Alonso, A. M., Casado, D., López-Pintado, S. & Romo, J. (2014). Robust Functional Supervised Classification for Time Series. *Journal of Classification, 31*(3), pp. 325–350.

DOI: 10.1007/s00357-014-9163-x

Robust Functional Supervised Classification for Time Series

Andrés M. Alonso

Universidad Carlos III de Madrid, Spain and INAECU, Spain

David Casado

Universidad Complutense de Madrid, Spain

Sara López-Pintado

Columbia University, USA

Juan Romo

Universidad Carlos III de Madrid, Spain

Abstract: We propose using the integrated periodogram to classify time series. The method assigns a new time series to the group that minimizes the distance between the series integrated periodogram and the group mean of integrated periodograms. Local computation of these periodograms allows the application of this approach to nonstationary time series. Since the integrated periodograms are curves, we apply functional data depth-based techniques to make the classification robust, which is a clear advantage over other competitive procedures. The method provides small error rates for both simulated and real data. It improves existing approaches and presents good computational behavior.

Keywords: Time series; Supervised classification; Integrated periodogram; Functional data depth.

All authors supported in part by CICYT (Spain) grants SEJ2007-64500, and MICINN (Spain) grant ECO2008-05080. Research partially supported by grant ECO2011-25706 of the Spanish Ministry of Science and Innovation. A.M. Alonso supported in part by MICINN (Spain) grant ECO2012-38442.

Corresponding Author's Address: David Casado, Universidad Complutense de Madrid, Av Sneca, 2, 28040 Madrid, Spain, e-mail: david@casado-d.org.

1. Introduction

Classification of time series is an important tool in several fields. Time series can be studied from both time and frequency domains. For short stationary series, a time domain approach based on usual multivariate techniques can be applied. Nevertheless, the frequency point of view is particularly important for nonstationary series (Huang, Ombao, and Stoffer 2004), which justifies why our proposal follows a frequency domain approach. There exist many papers on supervised classification methods for stationary processes in both domains (see e.g. references in Chapter 7 of Taniguchi and Kakizawa 2000). Several authors have already proposed methods for discriminating between nonstationary models. By using optimal scoring, Hastie, Buja, and Tibshirani (1995) cast the classification problem into the regression framework, where a penalized technique can be applied to the coefficients. As they wrote, "it is natural, efficient and sometimes essential to impose a spatial smoothness constraint on the coefficients, both for improved prediction performance and interpretability". Their proposal is designed for situations where the discriminant variables (predictors) are highly correlated, e.g. when a function is discretized. The following approaches are based on Dahlhaus's (1996; 1997) local stationarity framework. Shumway (2003) uses the Kullback-Leibler discrimination information measure (it is not a real distance), which is evaluated by using the smoothed time-varying spectral estimator. For clustering, they consider the symmetrized version of that measure. In a first step, Huang, Ombao, and Stoffer (2004) select from SLEX a basis explaining the difference between the classes of time series as well as possible; in a second step, they construct a discriminant criterion that is related to the SLEX spectra of the different classes: a time series is assigned to the class minimizing the Kullback-Leibler divergence between the estimated spectrum and the spectrum of the class. Sakiyama and Taniguchi (2004) use a consistent classification criterion that is an approximation of the Gaussian likelihood ratio. By introducing an influence function, they investigate the behavior of their measure with respect to infinitesimal perturbations of the spectra. In Hirukawa (2004) the approximation of the measure introduced by Sakiyama and Taniguchi for multivariate, non-Gaussian, locally stationary process is generalized to nonlinear time-varying spectral measures (including the Kullback-Leibler and Chernoff discrimination information measures). For non-Gaussian processes, Hirukawa proposes a specific asymptotically optimal criterion based on the concept of *quasi-log-likelihood ratio*, instead of the log-likelihood ratio. The discrimination of Chandler and Polonik (2006) is based on some features – shape measures or, better, measures of concentration of the variance function – that are measured for each time series. Since it is not distance-based, their approach

does not require aligning the series. Both time and frequency domains are connected in Maharaj and Alonso (2007), who combine the techniques of wavelet analysis with those of discriminant analysis. Other related line of research is unsupervised classification of time series – see Liao (2005) for a comprehensive survey.

In this paper, we propose using the integrated periodogram for classifying (locally stationary) time series. The integrated periodogram has the following properties that improve the classification procedure: i) it is a non-decreasing, smooth curve; ii) it presents good asymptotic properties: while the periodogram is an asymptotically unbiased but inconsistent estimator of the spectral density, the integrated periodogram is a consistent estimator of the spectral distribution (see Chapter 6 of Priestly 1981); iii) although for stationary processes the integrated spectrum is usually estimated through the spectrum, from a theoretical point of view, the spectral distribution always exists whereas the spectral density only exists under absolutely continuous distributions.

Since the integrated periodogram is a function, we shall use specific techniques for functional data. There is a vast body of literature on the statistical analysis of functional data and, particularly, on their classification. For example, a penalized discriminant analysis is proposed in Hastie, Buja, and Tibshirani (1995); it is adequate for situations with many highly correlated predictors, as those obtained by discretizing a function. Non-parametric tools to classify a set of curves have been introduced in Ferraty and Vieu (2003), where the authors calculate the posterior probability of belonging to a given class of functions by using a consistent kernel estimator. A new method for extending classical linear discriminant analysis to functional data has been analyzed in James and Hastie (2001): this technique is particularly useful when only fragments of the curves are observed. The problem of unsupervised classification or clustering of curves is addressed in James and Sugar (2003), who elaborate a flexible model-based approach for clustering functional data; it is effective when the observations are sparse, irregularly spaced or occur at different time points for each subject. In Abraham, Cornillon, Matzner-Løber, and Molinari (2003), unsupervised clustering of functions is considered; they fit the data using B-splines and the partition is done over the estimated model coefficients using a k-means algorithm. In a related problem, Hall, Poskitt, and Presnell (2001) explore a functional data-analytic approach to perform signal discrimination. Nonetheless, many of these procedures are highly sensitive to outliers. A natural, simple way to classify functions is to minimize the distance between the new curve and a reference function of the group. The technique presented in this paper follows this approach. We first consider the mean of the integrated periodograms as the group representative element and then,

as a second approach, we use the idea of "deepest" curves to robustify the method.

The notion of statistical depth has already been extended to functional data (see e.g. López-Pintado and Romo 2009). In López-Pintado and Romo (2006) the concept of depth is used to classify curves. A statistical depth expresses the "centrality" or "outlyingness" of an observation within a set of data and provides a criterion to order observations from the center outward. Since robustness is an interesting feature of statistical methods based on depth, we have applied the ideas in López-Pintado and Romo (2006) to add robustness to our time series classification procedure. Their method orders the curves within a sample, based on a notion of depth for functions, and works with the α-trimmed mean as a reference curve of each group.

The paper is organized as follows. In Section 2 we include some definitions and describe the classification algorithm based on the integrated periodogram. Section 3 explains how depth can be used to make the method robust. Next two sections, 4 and 5, show the behavior of the procedure with simulated and real data, respectively. A brief summary of conclusions is given in Section 6.

2. Classifying Time Series

We propose transforming the initial time series into functional data by considering the integrated periodogram of each time series (see e.g. Figure 1). This permits us to use functional data classification techniques. Let $\{X_t\}$ be a stationary process with autocovariance function $\sigma_h = cov(X_t, X_{t-h})$, such that $\sum_{h=-\infty}^{+\infty} |\sigma_h| < +\infty$, and autocorrelation function $\rho_h = \sigma_h/\sigma_0$. The spectral density is $f(\omega) = \sum_{h=-\infty}^{+\infty} \rho_h \exp(-2\pi i h\omega)$, and it holds that $\rho_h = \int_{-1/2}^{+1/2} \exp(2\pi i h\omega) dF(\omega)$, where F is the spectral distribution function.

The *periodogram* is the corresponding sample version of the spectral density and it expresses the contribution of the frequencies to the variance of the series. Let $X = (x_1, \ldots, x_T)$ be a time series. The periodogram is given by

$$I_T(\omega_k) = \sum_{h=-(T-1)}^{(T-1)} \hat{\rho}_h \exp(-2\pi i h\omega_k), \tag{1}$$

where $\hat{\rho}_h$ denotes the sample autocorrelation at lag h and ω_k takes values in $\{k/T \mid k = 0, \ldots, [T/2]\}$, the discrete *Fourier frequencies* set. The *integrated* or *cumulative periodogram* is defined as $F_T(\omega_k) = \sum_{i=1}^{k} I_T(\omega_i)$ or, in its normalized version

$$F_T(\omega_k) = \sum_{i=1}^{k} I_T(\omega_i) / \sum_{i=1}^{m} I_T(\omega_i), \tag{2}$$

where m is the number of Fourier frequencies. Notice that the denominator in (2) is proportional to the variance of the time series, since $2\sum_{i=1}^{m} I_T(\omega_i) = \sum_{t=1}^{T}(x_t - \bar{x})^2$. Therefore, the nonnormalized version of the cumulative periodogram considers not only the shape of the integrated spectrum but also the scale, whereas the normalized version of the cumulative periodogram emphasizes the shape of the curves instead of the scale. For instance, if two time series have spectral densities such that $f_X(\omega) = c f_Y(\omega)$ for some $c > 1$, then they will have different integrated periodograms but equal normalized integrated periodograms. See Diggle and Fisher (1991) for details on the comparison of cumulative periodograms. As a simple criterion we recommend using the normalized version of the cumulative periodogram when the graphs of the functions of the different groups tend to intersect inside their domain of definition. If this is not the case, we recommend using the nonnormalized version. Notice also that the integrated periodogram is a consistent estimator of the integrated spectrum (see e.g. Chapter 6 of Priestley 1981).

Definitions (1) and (2) correspond to some particular values of ω, but they can be extended to any value in the interval $(-1/2, +1/2)$. Since the periodogram is defined only for stationary series, to classify nonstationary time series we shall consider them as locally stationary; this allows us to split the series into blocks, compute the integrated periodogram of each block and merge these periodograms in a final curve: the idea is to approximate the locally stationary processes by piecewise stationary processes. Figure 2(b) provides a blockwise spectral distribution estimation of the locally stationary process spectrum. There are two opposite effects when we increase the numbers of blocks: first, we get closer to the locally stationarity assumption; second, the integrated periodogram becomes a worse estimator of the integrated spectrum. Notice that this blockwise approach is compatible with the locally stationary time series model of Dahlhaus (1997) where an increasing T implies that more and more data of local structures are available, which allows us to consider the number of blocks as an increasing function of T. In the appendix, we present the locally stationary model of Dahlhaus (1997) and we propose an integrated spectrum based on this model. This integrated spectrum can be thought of as a population version of our blockwise integrated spectrum.

A simple criterion to classify functions is to assign a new observation to the group to which, on the basis of some distance, the function is nearest. In our context, we propose classifying a new series in the group minimizing the distance between the integrated periodogram of the series and a reference curve from the group. We first consider the group mean as a reference curve.

If $\Psi_{gi}(\omega)$, $i = 1,\ldots,N$, are functions of group g, the mean is

$$\bar{\Psi}_g = \frac{1}{N}\sum_{i=1}^{N}\Psi_{gi}(\omega). \tag{3}$$

In our case, $\Psi_{gi}(\omega)$ is the concatenated integrated periodogram of each block of the ith series in group g. To measure proximity, we have chosen the L_1 distance,

$$\begin{aligned} d(\Psi_1,\Psi_2) &= \int_{-1/2}^{+1/2}|\Psi_1(\omega)-\Psi_2(\omega)|d\omega \\ &= \sum_{j=1}^{k}\int_{-1/2}^{+1/2}|F_1^{(j)}(\omega)-F_2^{(j)}(\omega)|d\omega, \end{aligned} \tag{4}$$

where k is the number of blocks in which the time series is divided and $F^{(j)}$ is the integrated periodogram of the jth block. The integrated periodograms belong to the $L_1[-1/2,+1/2]$ space. Some other distances could have also been considered; for example, the L_2 distance would highlight large differences between functions.

Based on these definitions we introduce the following classification algorithm:

Algorithm 1:

Let $\{X_1,\ldots,X_M\}$ be a sample containing M series from population P_X and let $\{Y_1,\ldots,Y_N\}$ be a sample containing N series from P_Y. The classification method includes the following steps:

1. Split each series into k blocks, calculate the integrated periodogram in each block, and merge these integrated periodograms: $\{\Psi_{X_1},\ldots,\Psi_{X_M}\}$ and $\{\Psi_{Y_1},\ldots,\Psi_{Y_N}\}$, where $\Psi_{X_i} = (F_{X_i}^{(1)}\ldots F_{X_i}^{(k)})$, $\Psi_{Y_i} = (F_{Y_i}^{(1)}\ldots F_{Y_i}^{(k)})$, and $F_{X_i}^{(j)}$ is the integrated periodogram of the jth block of the ith series of population X; and analogously for Y. Figures 2(b) and 4 illustrate the obtained Ψ_{X_i}.
2. Calculate the corresponding group means, $\bar{\Psi}_X$ and $\bar{\Psi}_Y$.
3. Let $\Psi_Z = (F_Z^{(1)}\ldots F_Z^{(k)})$ be the curve of a new series Z. Then Z is classified in the group P_X if $d(\Psi_Z,\bar{\Psi}_X) < d(\Psi_Z, \bar{\Psi}_Y)$; and in the group P_Y otherwise.

Remark 1: Set $k = 1$ to apply the algorithm to stationary series. For nonstationary series, in the computations with both simulated and real data we have used a dyadic splitting of the series into blocks: $k = 2^p$, $p = 0, 1, \ldots$ The implementation with blocks of different lengths, as suggested by visual inspection of the data, is also possible. To select the number of blocks, our code implements an optional nested/secondary cross-validation loop to select, in each run, the value of k that minimizes the global error (we register these values during the runs to form weights that can be thought of as relative frequencies). When the previous loop is not called, the minimum global error of each run is registered and the user is given an estimate of the error that would have arisen if the number of blocks had been optimized. For this loop to be applicable to small real data sets, the data of the primary cross-validation loop are used for both optimizing the number of blocks and estimating the final misclassification error rates.

Remark 2: Although we are considering $G = 2$, the classification method is obviously extended to the general case in which there are G different groups or populations P_g, $g = 1, \ldots, G$.

Remark 3: The same methodology could be implemented by using different classification criteria between curves, reference functions for each group (as we do in the following section) or distances between curves.

Remark 4: Notice that in this paper we only consider nonstationarities in the autocovariance structure, as we assume that the series are mean stationary. In the case of nonstationarities in the mean (trends, level shifts, piecewise trends, et cetera), we should divide the analysis in two cases: (*i*) The nonstationaries in the mean are different in the two populations so they will be useful to improve the classification procedure. In this case, an option would be the admissible linear procedure described in Section 7.2.3 of Taniguchi and Kakizawa (2000), though this is out of the scope of this paper. (*ii*) The nonstationaries in the mean are equal in the two populations so they will not be useful to improve the classification procedure. In this case, we should remove the nonstationaries in the mean by, for instance, the Hodrick-Prescott filter (see Hodrick and Prescott 1997) or the detrending procedure based on Loess (see Cleveland, Cleveland, McRae, and Terpenning 1990).

3. Robust Time Series Classification

Our classification method depends on the reference curve used to measure the distance to the group. The mean of a set of functions is not robust to the presence of outliers. Thus, robustness can be added to this tech-

nique by using a robust reference curve. Instead of considering the mean of the integrated periodograms in the group, we shall consider the α-trimmed mean, where only the deepest elements are averaged. This trimming adds robustness by making the reference function more resistant to outliers.

Statistical depth measures the "centrality" of each element inside the group. Different definitions of depth are already available. In this section we first describe the concept of depth extended to functional data by López-Pintado and Romo (2009) and then we propose a robust version of our classification algorithm.

Let $G(\Psi) = \{(t, \Psi(t)) \mid t \in [a,b]\}$ denote the graph in $\mathbb{R}^2$ of a function $\Psi \in C[a,b]$, the set of real continuous functions on the interval $[a,b]$. Let $\Psi_i(t)$, $i = 1,\ldots,N$, be functions in $C[a,b]$. The functions $\Psi_{i_j}(t)$, $j = 1,\ldots,h$, determine a band in $\mathbb{R}^2$,

$$\begin{aligned} &B(\Psi_{i_1},\ldots,\Psi_{i_h}) = \\ &\quad \{(t,y) \in [a,b] \times \mathbb{R} \mid \min_{r=1,\ldots,h} \Psi_{i_r}(t) \leq y \leq \max_{r=1,\ldots,h} \Psi_{i_r}(t)\}. \end{aligned} \tag{5}$$

Given a function Ψ,

$$\begin{aligned} BD_N^{(j)}(\Psi) = \tbinom{N}{j}^{-1} \cdot \textstyle\sum_{1 \leq i_1 < i_2 < \ldots < i_j \leq N} I\{G(\Psi) \\ \subset B(\Psi_{i_1},\ldots,\Psi_{i_j})\}, \;\; 2 \leq j \leq N, \end{aligned} \tag{6}$$

expresses the proportion of bands determined by different curves $\Psi_{i_1},\ldots,\Psi_{i_j}$ that contain the graph of Ψ (the indicator function takes the value $I\{A\} = 1$ if A occurs, and $I\{A\} = 0$ otherwise). For functions $\Psi_i(t)$, $i = 1,\ldots,N$, the *band depth* of any of these curves Ψ is

$$BD_{N,J}(\Psi) = \sum_{j=2}^{J} BD_N^{(j)}(\Psi), \;\; 2 \leq J \leq N. \tag{7}$$

If $\tilde{\Psi}$ is the stochastic process generating the observations $\tilde{\Psi}_i(t)$, $i = 1,\ldots,N$, the population versions of these indexes are:

$$BD^{(j)}(\Psi) = P\{G(\Psi) \subset B(\tilde{\Psi}_{i_1},\ldots,\tilde{\Psi}_{i_j})\}, 2 \leq j \leq N,$$

and

$$\begin{aligned} BD_J(\Psi) = \textstyle\sum_{j=2}^{J} BD^{(j} = \sum_{j=2}^{J} P\{G(\Psi) \subset B(\tilde{\Psi}_{i_1},\ldots,\tilde{\Psi}_{i_j})\}, \\ 2 \leq J \leq N, \end{aligned}$$

respectively. In order to illustrate the calculation of the *band depth*, consider the following example: Assume that we have two time series generated by AR(1) models:

$$X_t^{(i)} = \phi X_{t-1}^{(i)} + \varepsilon_t^{(i)},$$

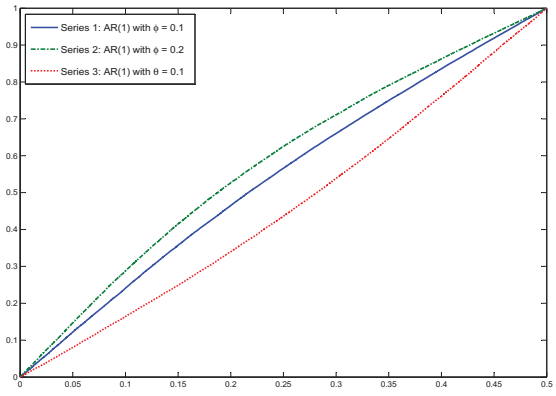

Figure 1: Example of three integrated periodograms

with $\phi^{(1)} = 0.1$, $\phi^{(2)} = 0.2$ and an additional time series generated by a MA(1) model:

$$X_t^{(3)} = \theta\varepsilon_{t-1}^{(3)} + \varepsilon_t^{(3)},$$

where $\theta = 0.1$ and the $\varepsilon_t^{(i)}$ are i.i.d. N(0,1). Figure 1 shows the three (theoretical) integrated periodograms. To calculate the depth of each function (integrated periodogram), we determine the $\binom{3}{2} = 3$ bands defined by these three functions, i.e. the bands defined by (1,2), (1,3) and (2,3). Notice that the integrated periodogram of the first series is included in the three bands and the integrated periodograms of the second and third series are include in only two bands; therefore, their band depths are 1, 2/3 and 2/3, respectively. For instance, as the graph shows, the integrated periodogram of the first series is the deepest element.

The *modified band depth* is a more flexible notion of depth also defined in López-Pintado and Romo (2009). The indicator function in (6) is replaced by the length of the set where the function is inside the corresponding band. For any function Ψ of $\Psi_i(t)$, $i = 1,\ldots,N$, and $2 \leq j \leq N$, let

$$\begin{aligned} A_j(\Psi) \equiv A(\Psi;\Psi_{i_1},\ldots,\Psi_{i_j}) \equiv \{t \in [a,b] \mid \min_{r=i_1,\ldots,i_j} \Psi_r(t) \\ \leq \Psi(t) \leq \max_{r=i_1,\ldots,i_j} \Psi_r(t)\} \end{aligned} \tag{8}$$

be the set of points in the interval $[a,b]$ where the function Ψ is inside the band. If λ is the Lebesgue measure on the interval $[a,b]$, $\lambda(A_j(\Psi))$ is the "proportion of time" that Ψ is inside the band. Thus,

$$MBD_N^{(j)}(\Psi) = \binom{N}{j}^{-1}(\lambda[a,b])^{-1}\cdot \sum_{1\le i_1<i_2<\ldots<i_j\le N}\lambda(A(\Psi;\Psi_{i_1},\ldots,\Psi_{i_j})),\ \ 2\le j\le N, \tag{9}$$

is the generalized version of $BD_N^{(j)}$. If Ψ is always inside the band, the measure $\lambda(A_j(\Psi))$ is 1 and this definition generalizes the definition of depth given in (7). Finally, the modified band depth of any of the curves Ψ in $\Psi_i(t)$, $i=1,\ldots,N$, is

$$MBD_{N,J}(\Psi) = \sum_{j=2}^{J} MBD_N^{(j)}(\Psi),\ \ 2\le J\le N. \tag{10}$$

If $\tilde{\Psi}_i(t)$, $i=1,\ldots,N$, are independent copies of the stochastic process $\tilde{\Psi}$, the population version of these indexes are

$$MBD^{(j)}(\Psi) = \mathbb{E}[\lambda(A(\Psi;\tilde{\Psi}_{i_1},\ldots,\tilde{\Psi}_{i_j}))],\ 2\le j\le N,$$
and
$$MBD_J(\Psi) = \textstyle\sum_{j=2}^{J} MBD^{(j)}(\Psi) = \sum_{j=2}^{J}\mathbb{E}[\lambda(A(\Psi;\tilde{\Psi}_{i_1},\ldots,\tilde{\Psi}_{i_j}))],$$
$$2\le J\le N,$$

respectively.

Given a sample of functions, $(\Psi_{g_1},\Psi_{g_2},\ldots,\Psi_{g_N})$, we can order the curves by calculating the sample modified band depth, $MBD_{N,J}(\Psi_{g_i})$, of each function Ψ_{g_i} for $i=1,2,\ldots,N$. The ordered sample is denoted by $(\Psi_{g_{(1)}},\Psi_{g_{(2)}},\ldots,\Psi_{g_{(N)}})$, where $\Psi_{g_{(1)}}$ is the deepest function, $\Psi_{g_{(2)}}$ is the second deepest function, and so on.

To robustify Algorithm 1, we apply the α-trimmed mean of the elements as the group reference function. If $\Psi_{g_{(i)}}(t)$, $i=1,\ldots,N$, are functions of the class g ordered by decreasing depth, the α-trimmed mean is

$$\bar{\Psi}_g^{\alpha} = \frac{1}{N-[N\alpha]}\sum_{i=1}^{N-[N\alpha]}\Psi_{g_{(i)}}(t), \tag{11}$$

where $[\cdot]$ is the integer part function. Notice that the median (in the sense of "the deepest") function is also included in the previous expression. We shall use $\alpha=0.2$ in our analyses with simulated and real data. This means that for each group the 20% least deep data are left out when the average is computed.

In step 2 of the new algorithm, the α-trimmed mean replaces the mean as the reference curve for each class, which will make the classification more robust.

Algorithm 2:

Let $\{X_1,\ldots,X_M\}$ be a sample containing time series from the population P_X, and let $\{Y_1,\ldots,Y_N\}$ be a sample from P_Y. The classification method includes the following steps:

1. Split each series into k blocks, calculate the integrated periodogram in each block and merge these integrated periodograms: $\{\Psi_{X_1},\ldots,\Psi_{X_M}\}$ and $\{\Psi_{Y_1},\ldots,\Psi_{Y_N}\}$, where $\Psi_{X_i}=(F_{X_i}^{(1)}\ldots F_{X_i}^{(k)})$, $\Psi_{Y_i}=(F_{Y_i}^{(1)}\ldots F_{Y_i}^{(k)})$, and $F_{X_i}^{(j)}$ is the integrated periodogram of the jth block of the ith series of the population X; and analogously for Y.
2. Obtain the corresponding group α-trimmed means, $\bar{\Psi}_X^{\alpha}$ and $\bar{\Psi}_Y^{\alpha}$.
3. Let $\Psi_Z=(F_Z^{(1)}\ldots F_Z^{(k)})$ be the curve of a new series Z. Then Z is classified in the group P_X if $d(\Psi_Z,\bar{\Psi}_X^{\alpha})<d(\Psi_Z,\bar{\Psi}_Y^{\alpha})$, and in the group P_Y otherwise.

Remark 5: We have used the sample modified band depth with $J=2$ because this depth is very stable in J – similar center-outward orderings are obtained in a collection of functions for different values of J (López-Pintado and Romo 2006; 2009).

Remark 6: The same algorithm could be implemented by using a different functional depth.

Remark 7: Computing the depth of functional data is the most time-consuming task in our proposed robust classification algorithm. We implement a preprocessing step to help scale the algorithm to large real data sets as follows. The deepest elements are identified at the beginning so as to maintain only them in the training samples during the runs (although all data are classified). On the one hand, the depth is calculated only once; on the other hand, because of the use of fewer but better elements in the training samples, the computational time may be reduced in some cases for which the time spent in calculating the depth is compensated. With this preprocessing step the sizes of the training samples are slightly reduced in most runs, although this has little effect when sample sizes are large. This technique can be applied outside the framework of this work.

MATLAB code is available at http://www.Casado-D.org. Methods *DbC* and *DbC*-α, as well as other characteristics (loop to select the number of blocks, robustifying approach, access to the computational times, et

cetera) are implemented in several scripts. The code is fast and easy to execute and extend. The reader can easily reproduce, apply or extend our results and plots. A help file is also included with the code.

4. Simulation Study

In this section, we evaluate our two algorithms and compare them with the method proposed in Huang et al. (2004), who use the SLEX (smooth localized complex exponentials) model for a nonstationary random process introduced by Ombao, Rax, von Sachs, and Malow (2001). SLEX is a set of Fourier-type bases that are at the same time orthogonal and localized in both time and frequency domains. In a first step, they select from SLEX a basis explaining the difference between the classes of time series as well as possible. After this, they construct a discriminant criterion that is related to the SLEX spectra of the different classes: a time series is assigned to the class minimizing the Kullback-Leibler divergence between the estimated spectrum and the spectrum of the class. For the SLEXbC method we have used an implementation provided by the authors (see http://hombao.ics.uci.edu/). To select the parameters, we have performed a small optimization for each simulation, and the results were similar to the values recommended to us by the authors.

We have evaluated the methods with some models considered by Huang et al. (2004). For each comparison of two classes, we run 1000 times the following steps. We generate training and test sets for each model/class. The training sets have the same sizes (sample size and series length) as those used by Huang et al. (2004), and all the test sets contain 10 series of the length involved in each particular simulation. The methods are tested with the same data sets so that, in all models, exactly the same simulated time series are used by the three methods, including our algorithms for different values of k.

Simulation 1. We compare two processes composed half by white noise and half by an autoregressive process of order one. The value of the AR(1) parameter is -0.1 in the first class and $+0.1$ in the second class:

$$X_t^{(i)} = \begin{cases} \varepsilon_t^{(i)} & \text{if } t = 1,\ldots,T/2 \\ X_t^{(i)} = -0.1 \cdot X_{t-1}^{(i)} + \varepsilon_t^{(i)} & \text{if } t = T/2+1,\ldots,T \end{cases}$$

$$Y_t^{(j)} = \begin{cases} \varepsilon_t^{(j)} & \text{if } t = 1,\ldots,T/2 \\ Y_t^{(j)} = +0.1 \cdot Y_{t-1}^{(j)} + \varepsilon_t^{(j)} & \text{if } t = T/2+1,\ldots,T \end{cases} \tag{12}$$

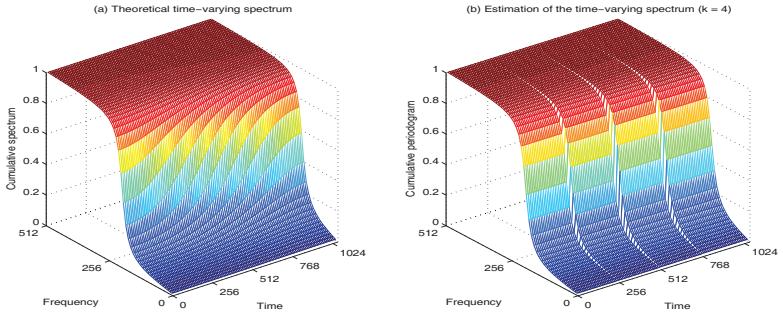

Figure 2: Time-varying autoregressive model with $\tau = 0.4$.

where ε_t are i.i.d. $N(0,1)$, with $i = 1,\ldots,M$ and $j = 1,\ldots,N$. Different combinations of training sample sizes—$M = N = 8$ and 16—and series lengths—$T = 512$, 1024 and 2048—are considered. In this case, the series are made up of stationary parts, but the whole series are not stationary.

Simulation 2. For this study, the stochastic models in both classes are slowly time-varying second order autoregressive processes:

$$\begin{aligned} X_t^{(i)} &= a_{t;0.5} \cdot X_{t-1}^{(i)} - 0.81 \cdot X_{t-2}^{(i)} + \varepsilon_t^{(i)} \qquad \text{if } t = 1,\ldots,T \\ Y_t^{(j)} &= a_{t;\tau} \cdot Y_{t-1}^{(j)} - 0.81 \cdot Y_{t-2}^{(j)} + \varepsilon_t^{(j)} \qquad \text{if } t = 1,\ldots,T \end{aligned} \tag{13}$$

where ε_t are i.i.d. $N(0,1)$, with $i = 1,\ldots,M$, $j = 1,\ldots,N$ and $a_{t;\tau} = 0.8 \cdot [1 - \tau\cos(\pi t/1024)]$, where τ is a parameter. Each training data set has $M = N = 10$ series of length $T = 1024$. Three comparisons have been done, the first class having always the parameter $\tau = 0.5$ and the second class having, respectively, the values $\tau = 0.4$, 0.3 and 0.2. Note that a coefficient of the autoregressive structure is not fixed but changes with time, making the processes nowhere stationary. See Figure 2(a) for an example of the integrated spectrum corresponding to these processes. As a precaution, we have checked that values between $\tau = -0.9$ and $\tau = +0.9$ do not generate, for any value of t, roots inside the unit circle for the characteristic polynomial of the autoregressive process.

To compare the three methods in terms of robustness, we have performed additional simulations where the training set is contaminated with an outlier time series. In all cases we contaminate the P_X population by replacing a series by another one following a different model. We consider three levels of contamination: one weak contamination (A) and two

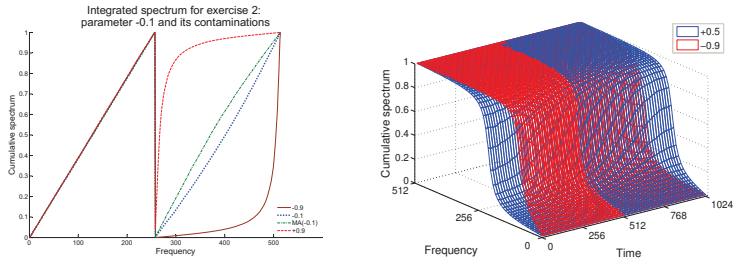

Figure 3: Examples of contamination for the two exercises, respectively.

strong contaminations (B and C). Since similar results are obtained for the two strong contaminations, only the former is reported in this paper. Other results on misclassification rates and computational times can be found in Alonso, Casado, López-Pintado, and Romo (2008) and Casado (2013).

> **Contamination A**. For simulation 1, we replace the autoregressive structure—half of the series—for a moving average process; that is, we generate a MA(1) model—with the MA parameter equal to the AR parameter—instead of a AR(1) model. For simulation 2, we contaminate the set of slowly time-varying autoregressives of parameter +0.5 with a series of the same model but with parameter value +0.2.

> **Contamination B**. This type of contamination corresponds to a parameter value of $\phi = -0.9$ in simulation 1 and $\tau = -0.9$ in simulation 2, instead of the correct values. Therefore, we are always applying the correct model except in one case, where we modify the parameter value. (Contamination C consists in using the value +0.9 instead of -0.9.)

In Figure 3, some cases of contamination are shown for the two simulation exercises. The error rates estimated for the first simulation are presented in Tables 1, 2 and 3; for the second simulation, in Tables 4, 5 and 6. Each cell includes the mean and the standard error (in parenthesis) of the 1000 runs.

For all tables we use the following notation: DbC (from *depth-based classification*) for algorithm 1, DbC-α for algorithm 2 (using $\alpha = 0.2$ for the α-trimmed mean and $J = 2$ for the modified band depth) and SLEXbC for the method of Huang et al. (2004). When a number follows DbC or DbC-α, it indicates the number k of blocks into which the series are split. Given a length T, SLEXbC considers several levels or number of partitions of the series $(1, 2, 2^2, ..., 2^p)$ and usually selects and combines blocks from

different levels, that is, blocks of different length, to calculate the SLEX spectrum. For example, for $T = 1024$, partitions into 1, 2, 4 and 8 blocks are managed by SLEXbC, and that is why the same values have been considered for our methods. Finally, the digits in bold correspond to the minima (when they are different to zero).

Comments on Error Rates

Tables 1, 2 and 3 provide the results of the first simulation experiment. When contamination is not present, DbC and DbC-α provide similar error rates, about half the size of those obtained by SLEXbC. As we could expect, for DbC and SLEXbC, error rates increase slightly with contamination A (weak) and notably with contamination B (strong), while changes are negligible for DbC-α because the trim keeps the contamination out. The DbC error rate is about half of SLEXbC error rate for contamination A, but their error rates are similar with contamination B. DbC-α is the only method maintaining the same pattern (with and without contamination) and having a considerable amount of error values close to zero. All three methods missclassify no elements for values of $\phi \in \{-0.5, -0.3, 0.3, 0.5\}$ (these results have not been included in the paper). As we might expect, error rates decrease when either N or T increases. Our methods reach the minima when series are divided into two blocks. While our error rates are larger than the rates of SLEXbC, when we consider the whole series (without splitting them into blocks), they fall with the first division. As we mentioned before, the length of the blocks decreases with k, and this negatively affects the performance of the periodogram as an estimator. We can observe this effect of splitting in all the tables of simulation 1, and it is also evident that the increase with k is higher for short series than for longer ones. Nevertheless, we observe that even with $k = 8$ the misclassification rates are smaller than the ones obtained by the SLEXbC procedure or those obtained by our procedures with $k = 1$. Recall that, like our procedure, the SLEXbC method implicitly splits the series into blocks. Regarding the contaminations, for DbC and SLEXbC, error rates increase slightly with contamination A and greatly for contamination B, while DbC-α maintains its error rates and outperforms all the other methods, mainly with strong contamination and when two blocks are considered. As could be expected, contaminating a series has major effects when samples sizes are $N_x = N_y = 8$ than when $N_x = N_y = 16$.

For simulation 2, conclusions similar to the previous ones can be derived from Tables 4, 5 and 6. They also show that, in our proposal, penalization for splitting too much is not serious when series are long enough. Generally, the best results with both methods are obtained with $k = 4$, but even with $k = 8$ the misclassification rates are smaller than those obtained by the SLEXbC procedure or the ones obtained by our procedures with $k = 1$.

Table 1: Misclassification rates estimates for simulation 1 without contamination.

	NxT = **7x512**	**16x512**	**7x1024**	**16x1024**	**7x2048**	**16x2048**
DbC 1	0.141 (0.0024)	0.131 (0.0024)	0.062 (0.0017)	0.060 (0.0017)	0.014 (0.0008)	0.014 (0.0008)
2	**0.066** (0.0017)	**0.061** (0.0017)	**0.015** (0.0009)	**0.014** (0.0008)	**0.001** (0.0003)	**0.001** (0.0003)
4	0.078 (0.0019)	0.069 (0.0018)	**0.015** (0.0009)	0.014 (0.0009)	**0.001** (0.0003)	**0.001** (0.0003)
7	0.090 (0.0020)	0.080 (0.0019)	0.020 (0.0010)	0.018 (0.0009)	0.002 (0.0003)	**0.001** (0.0003)
DbC-α 1	0.143 (0.0024)	0.132 (0.0024)	0.063 (0.0017)	0.061 (0.0017)	0.015 (0.0009)	0.014 (0.0008)
2	0.069 (0.0018)	0.064 (0.0017)	0.016 (0.0009)	0.015 (0.0009)	**0.001** (0.0003)	**0.001** (0.0003)
4	0.083 (0.0020)	0.073 (0.0018)	0.017 (0.0010)	0.016 (0.0009)	0.002 (0.0003)	**0.001** (0.0003)
7	0.105 (0.0023)	0.088 (0.0020)	0.024 (0.0011)	0.019 (0.0010)	0.002 (0.0004)	0.002 (0.0003)
SLEXbC	0.114 (0.0023)	0.086 (0.0020)	0.038 (0.0014)	0.025 (0.0011)	0.007 (0.0006)	0.003 (0.0004)

Table 2: Misclassification rates estimates for simulation 1 with contamination A.

	NxT = **8x512**	**16x512**	**8x1024**	**16x1024**	**8x2048**	**16x2048**
DbC 1	0.143 (0.0025)	0.132 (0.0024)	0.063 (0.0017)	0.062 (0.0017)	0.018 (0.0010)	0.015 (0.0008)
2	**0.070** (0.0018)	**0.062** (0.0017)	0.018 (0.0010)	**0.014** (0.0008)	0.002 (0.0003)	**0.001** (0.0003)
4	0.083 (0.0020)	0.071 (0.0019)	0.019 (0.0010)	0.015 (0.0009)	0.002 (0.0003)	**0.001** (0.0003)
8	0.102 (0.0022)	0.083 (0.0020)	0.026 (0.0012)	0.019 (0.0010)	0.003 (0.0004)	0.002 (0.0003)
DbC-α 1	0.145 (0.0025)	0.132 (0.0023)	0.063 (0.0017)	0.061 (0.0017)	0.015 (0.0009)	0.014 (0.0008)
2	0.072 (0.0018)	0.064 (0.0017)	**0.015** (0.0009)	0.015 (0.0009)	**0.001** (0.0002)	**0.001** (0.0003)
4	0.086 (0.0021)	0.073 (0.0018)	0.018 (0.0010)	0.016 (0.0009)	0.002 (0.0003)	**0.001** (0.0003)
8	0.114 (0.0024)	0.089 (0.0021)	0.025 (0.0011)	0.019 (0.0010)	0.003 (0.0004)	0.002 (0.0003)
SLEXbC	0.128 (0.0025)	0.092 (0.0021)	0.050 (0.0016)	0.027 (0.0012)	0.012 (0.0008)	0.004 (0.0004)

Table 3: Misclassification rates estimates for simulation 1 with contamination B.

	NxT = **8x512**	**16x512**	**8x1024**	**16x1024**	**8x2048**	**16x2048**
DbC 1	0.258 (0.0029)	0.168 (0.0026)	0.252 (0.0029)	0.117 (0.0022)	0.250 (0.0029)	0.065 (0.0018)
2	0.135 (0.0024)	0.082 (0.0020)	0.088 (0.0021)	0.030 (0.0012)	0.049 (0.0016)	0.007 (0.0006)
4	0.137 (0.0025)	0.085 (0.0020)	0.089 (0.0021)	0.031 (0.0012)	0.049 (0.0016)	0.007 (0.0006)
8	0.143 (0.0025)	0.092 (0.0021)	0.093 (0.0022)	0.034 (0.0014)	0.050 (0.0016)	0.007 (0.0006)
DbC-α 1	0.145 (0.0024)	0.134 (0.0024)	0.064 (0.0017)	0.061 (0.0017)	0.015 (0.0008)	0.014 (0.0008)
2	**0.070** (0.0018)	**0.065** (0.0017)	**0.017** (0.0010)	**0.015** (0.0009)	0.003 (0.0006)	**0.001** (0.0003)
4	0.081 (0.0020)	0.071 (0.0019)	**0.017** (0.0010)	0.017 (0.0009)	**0.002** (0.0003)	0.002 (0.0003)
8	0.104 (0.0023)	0.087 (0.0020)	0.023 (0.0011)	0.019 (0.0010)	0.002 (0.0004)	0.002 (0.0003)
SLEXbC	0.239 (0.0031)	0.134 (0.0024)	0.228 (0.0030)	0.081 (0.0020)	0.220 (0.0030)	0.037 (0.0013)

Table 4: Misclassification rates estimates for simulation 2 without contamination.

	$\tau = 0.4$	$\tau = 0.3$	$\tau = 0.2$
DbC 1	0.218 (0.0031)	0.063 (0.0017)	0.019 (0.0010)
2	0.119 (0.0023)	0.006 (0.0006)	**0.000** (0.0000)
4	**0.101** (0.0022)	**0.002** (0.0003)	**0.000** (0.0000)
8	0.123 (0.0024)	0.003 (0.0004)	**0.000** (0.0000)
DbC-α 1	0.226 (0.0032)	0.065 (0.0018)	0.021 (0.0010)
2	0.128 (0.0023)	0.006 (0.0006)	**0.000** (0.0000)
4	0.112 (0.0023)	**0.002** (0.0003)	**0.000** (0.0000)
8	0.139 (0.0026)	0.004 (0.0004)	**0.000** (0.0000)
SLEXbC	0.181 (0.0031)	0.011 (0.0009)	**0.000** (0.0000)

Table 5: Misclassification rates estimates for simulation 2 with contamination A.

	$\tau = 0.4$	$\tau = 0.3$	$\tau = 0.2$
DbC 1	0.232 (0.0032)	0.062 (0.0017)	0.019 (0.0009)
2	0.143 (0.0026)	0.006 (0.0006)	**0.000** (0.0000)
4	0.144 (0.0026)	0.004 (0.0004)	**0.000** (0.0000)
8	0.177 (0.0028)	0.005 (0.0005)	**0.000** (0.0000)
DbC-α 1	0.241 (0.0035)	0.065 (0.0018)	0.020 (0.0010)
2	0.131 (0.0025)	0.007 (0.0006)	**0.000** (0.0000)
4	**0.121** (0.0026)	**0.003** (0.0004)	**0.000** (0.0000)
8	0.150 (0.0029)	0.005 (0.0005)	**0.000** (0.0000)
SLEXbC	0.234 (0.0033)	0.016 (0.0011)	**0.000** (0.0000)

Table 6: Misclassification rates estimates for simulation 2 with contamination B.

	$\tau = 0.4$	$\tau = 0.3$	$\tau = 0.2$
DbC 1	0.254 (0.0029)	0.106 (0.0022)	0.043 (0.0015)
2	0.500 (0.0015)	0.067 (0.0021)	0.001 (0.0002)
4	0.500 (0.0012)	0.062 (0.0020)	0.001 (0.0002)
8	0.499 (0.0013)	0.082 (0.0024)	0.000 (0.0001)
DbC-α 1	0.231 (0.0031)	0.074 (0.0020)	0.026 (0.0012)
2	0.128 (0.0024)	0.007 (0.0006)	**0.000** (0.0000)
4	**0.113** (0.0023)	**0.002** (0.0004)	**0.000** (0.0000)
8	0.141 (0.0026)	0.003 (0.0004)	0 **0.000** (0.0000)
SLEXbC	0.492 (0.0019)	0.174 (0.0051)	0.015 (0.0009)

Notice that in this case there does not exist a theoretical optimum k. In the contaminated models, the best error rates are obtained with DbC-α for $k = 4$. As we can see, contamination A has a small effect.

Finally, in the two experiments only a subtle difference can be seen between DbC and DbC-α. When there is no contamination, it is natural that

the former provides slightly better error rates, since the latter, because of its trimming, uses only $100(1-\alpha)\%$ of the suitable training data available. Similar results were obtained when the L_2 distance is used instead of L_1.

5. Real Data Examples

In this section, we illustrate the performance of our proposal in two benchmark data sets: (*i*) Geological data consisting of 17 time series corresponding to earthquakes and explosions; and (*ii*) Speech recognition data consisting of three sets of 100 labeled time series corresponding to digitized speech frames.

5.1 Geological Data

In this section, we have evaluated our proposal in a data set containing eight explosions, eight earthquakes and one extra series—known as NZ event—not classified (but being an earthquake or an explosion). This data set was constructed by Blandford (1993). Each series consists of 2048 points, and its plot clearly shows two different parts—the first half is part P and the second half is S. This division is an assumption made by most authors, and is based on geological reasons. Both parts are also commonly considered stationary. Kakizawa, Shumway, and Taniguchi (1998) give a list of these measurements. Shumway and Stoffer (2000) include a detailed study of this data set and provide access to it at http://www.stat.pitt.edu/stoffer/tsa.html. Figure 4 presents examples of an earthquake, an explosion, the NZ event and their respective integrated periodograms.

Following the simple criterion given in Section 2 to choose between the normalized or the nonnormalized version of the cumulative periodogram, and after visual observation of these data, for each series we have built a curve by merging the nonnormalized integrated periodograms of parts P and S independently computed; that is, we take $k = 2$, as used by most authors. Let us define the eight earthquakes as group 1 and the eight explosions as group 2. We use leave-one-out cross-validation to classify the elements of these two groups: that is, removing a series at a time, using the rest of the data set to train the method and finally classifying the removed series. By doing this, both of our algorithms misclassify the first series of the group 2 (explosions). Regarding the NZ event, if we use the previous groups as training sets, both algorithms agree on assigning it to the explosions group, which agrees with the results obtained by, e.g., Kakizawa et al. (1998) or Huang et al. (2004).

Now we propose an additional scenario. We consider an artificial data set constructed by the eight earthquakes plus the NZ event as group

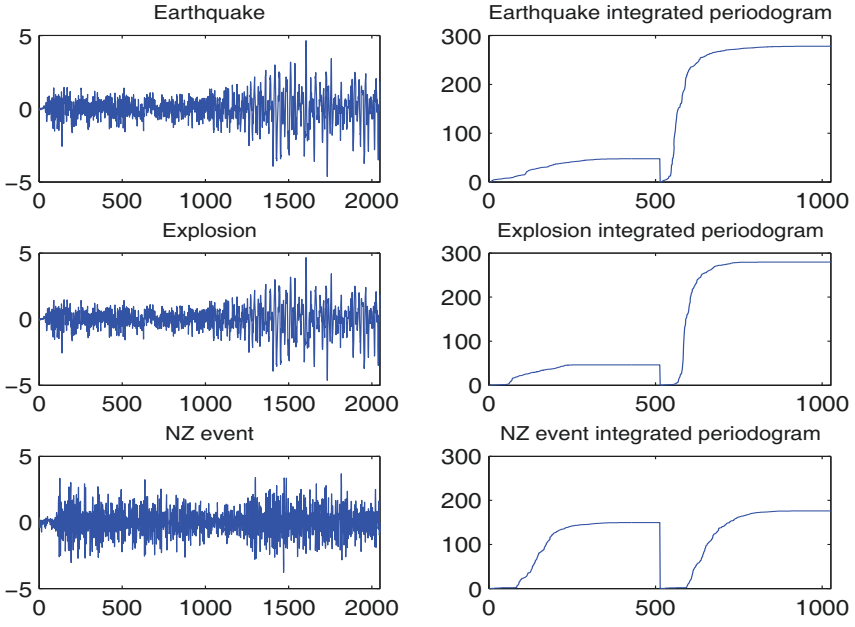

Figure 4: Geological data

1, and the eight explosions as group 2. (Note that our method and most of the published papers classify NZ as an explosion.) Then we could consider this artificial setting as a case where one atypical observation is presented in group 1. In this situation, algorithm 1 misclassifies the first and the third elements of group 2 (explosions), not only the first, whereas algorithm 2 still misclassifies only the first series of group 2. This seems to show the robustness of our second algorithm. Obviously, since we are applying leave-one-out cross-validation, both algorithms classify the NZ event in the explosions group, as we mentioned in the previous paragraph.

5.2 Speech Recognition Data

In this section, we have evaluated our proposal in a benchmark data set containing three subsets of 100 recordings of two short words or phonemes. These three data sets were used by Biau, Bunea, and Wegkamp (2003) to illustrate the performance of several classification procedures on functional data. Their procedures consider the time series as functional data. The first set corresponds to the words YES and NO with 52 and 48 speech frames, respectively; the second set corresponds to the words BOAT and GOAT with 55 and 45 speech frames, respectively; and the third set to the phonemes SH (as in SHE) and AO (as in WATER) with 42 and 58 speech frames,

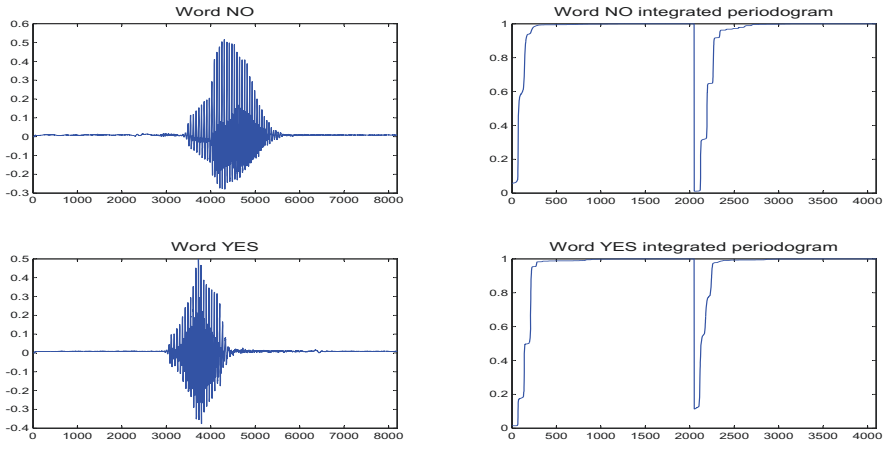

Figure 5: Words YES/NO data

respectively. Each speech frame consists of a time series of length 8192 observations. Figures 5, 6 and 7 present examples of different words or phonemes and their respective integrated periodograms. As is clear from those figures, the time series are nonstationary so, as a consequence, there must be $k > 1$ blocks in our procedures. For illustrative purposes, we use $k = 2$ in the figures, although the "best" k could be selected by a cross-validation procedure.

Biau et al. (2003) report their misclassification error rates based on a cross-validation procedure with 50 time series as training sample and the remaining 50 time series as testing sample. The results with their nonparametric functional classification procedure and two alternative procedures (nearest neighbour procedure and quadratic discriminant analysis) are 0.10–0.36–0.07, 0.21–0.42–0.35 and 0.16–0.42–0.19 for YES/NO, BOAT/GOAT and SH/AO, respectively. Table 7 shows our classification results after applying the same cross-validation scheme with different values of k. The misclassification error rates reported in Table 7 are based on 1000 replications.

Our results are similar or better than those obtained by Biau et al. (2003). The robust algorithm, DbC-α, provides the best results for the YES/NO and BOAT/GOAT sets, having misclassification rates around 0.05 (with k=4,8 or 16) and 0.150 (with k=16 or 32), respectively. Both methods, DbC and DbC-α, lead to almost perfect classification in the SH/AO set, which is a big improvement with respect to the three methods used in Biau et al. (2003). For this third set, the impact of k is not relevant.

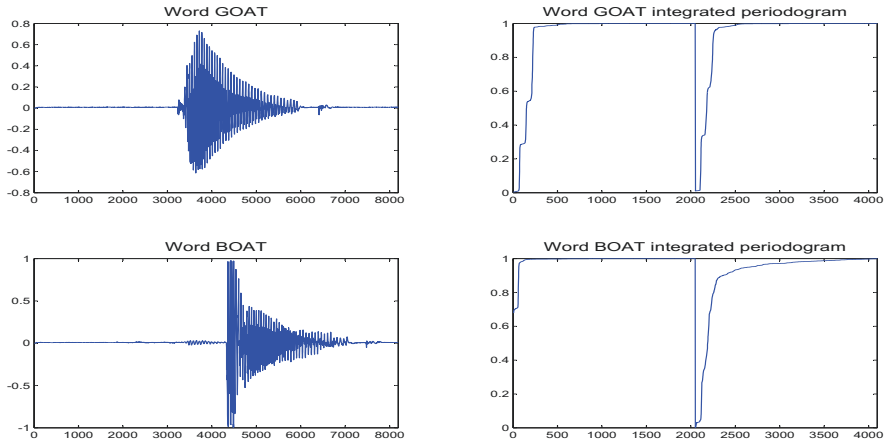

Figure 6: Words BOAT/GOAT data

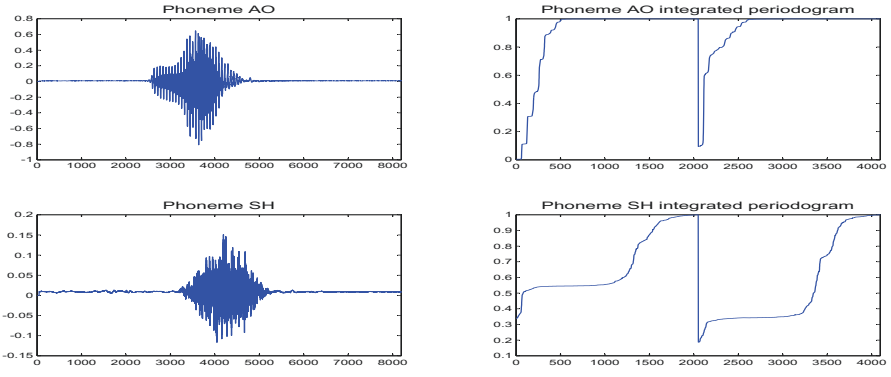

Figure 7: Phonemes SH/AO data

Additionally, in Figure 8 we show the overall error rate (based on 100 runs) for the YES/NO data set when from one to thirty two blocks are considered. The computational time spent generating Figure 8 was around 843.8 seconds, which confirms the practicability of the block selection procedure. Notice that the selection of blocks is performed only once. The best results for DbC and DbC-α are obtained with $k = 5$ and $k = 16$, respectively. Moreover, Figure 8 illustrates that, in this data set, once we select a $k > 4$, the misclassification rates are fairly stable.

Table 7: Misclassification rates estimates for speech recognition data.

	YES/NO	BOAT/GOAT	SH/AO
DbC 1	0.404 (0.0018)	0.387 (0.0030)	**0.000** (0.0000)
2	0.407 (0.0021)	0.345 (0.0026)	**0.000** (0.0000)
4	0.102 (0.0021)	0.285 (0.0023)	0.003 (0.0002)
8	**0.091** (0.0014)	0.265 (0.0017)	0.003 (0.0002)
16	0.100 (0.0015)	0.253 (0.0028)	0.003 (0.0002)
32	0.117 (0.0016)	**0.250** (0.0028)	0.008 (0.0005)
DbC-α 1	0.281 (0.0038)	0.360 (0.0041)	**0.000** (0.0000)
2	0.170 (0.0032)	0.293 (0.0041)	**0.000** (0.0001)
4	0.066 (0.0012)	0.217 (0.0033)	0.005 (0.0003)
8	**0.049** (0.0009)	0.223 (0.0026)	0.007 (0.0003)
16	0.058 (0.0010)	**0.143** (0.0032)	0.012 (0.0005)
32	0.085 (0.0012)	0.164 (0.0035)	0.023 (0.0006)

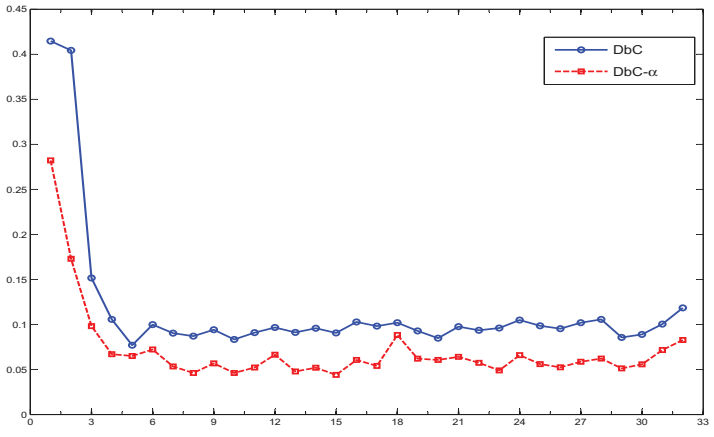

Figure 8: Overall error rate estimated by cross-validation in YES/NO data set.

6. Conclusions

We introduced a new time series classification method based on the series integrated periodogram. Notice that the calculation of the (integrated) periodogram does not involve a bandwidth selection, in contrast to other spectral (distribution) density estimators. This is a clear advantage with respect to methods that require smooth and consistent spectral density estimators. When the series are nonstationary, they are split into blocks and the integrated periodograms of the blocks are merged to construct a curve;

this idea relies on the assumption of local stationarity of the series. Since the integrated periodogram is a function, statistical methods recently developed for functional data can be applied. New series are assigned to the class minimizing the distance between its group mean curve and the new data function. Since the group mean can be affected by the presence of outliers, we propose robustifying the classification by replacing the mean curve by the depth-based α-trimmed mean, where for each group only the deepest elements are averaged. We have evaluated our proposal in different scenarios. We have run two simulations containing several models and parameter values, one with piecewise stationary series and the other with nowhere stationary series. After running the simulations without contamination, we have repeated all the comparisons twice more with exactly the same series but replacing one by a contaminated series. Two levels of contamination are considered: weak and strong. Our second algorithm exhibits robustness against outliers, while the performance of the SLEXbC procedure deteriorates noticeably. We also illustrate the performance of our procedure in two benchmark data sets. Our proposal provides small error rates, robustness and good computational behavior, which makes the method suitable for classifying long time series. Finally, this paper suggests that the integrated periodogram contains useful information for classifying time series, and that the concept of depth for functional data can be used to make classification robust, which is a clear advantage over other competitive procedures that are strongly affected by the presence of outliers.

Appendix

In this section we follow the papers of Dahlhaus (1996; 1997) to present a locally stationary time series model that allows us to define a time dependent integrated spectrum. In this nonstationary framework it is not possible to separate the time and the frequency domains. The strategy of Dahlhaus started with a spectral representation.

Definition 1 (Dahlhaus, 1996 and 1997). A sequence of stochastic processes $(X_{t,T}\ 1 \leq t \leq T,\ T \geq 1)$ is called *locally stationary with transfer function* A^0 *and trend* μ if such a representation exists

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{+\pi} e^{i\lambda t} A^0_{t,T}(\lambda) d\xi(\lambda), \tag{14}$$

where

(i) $\xi(\lambda)$ is a stochastic process on $[-\pi, +\pi]$ with $\overline{\xi(\lambda)} = \xi(-\lambda)$ and

$$\text{cum}\{d\xi(\lambda_1), \cdots, d\xi(\lambda_k)\} = \eta(\sum_{j=1}^{k} \lambda_j) g_k(\lambda_1, \cdots, \lambda_{k-1}) d\lambda_1 \cdots d\lambda_k,$$

where $g_1 = 0$, $g_2(\lambda) = 1$, $|g_k(\lambda_1, \cdots, \lambda_{k-1})| \leq \text{const}_k$ for all k, $\text{cum}\{\cdots\}$ denotes the cumulant of k-th order and $\eta(\lambda) = \sum_{j=-\infty}^{+\infty}$ $\delta(\lambda + 2\pi j)$ is the period 2π extension of the Dirac delta function.

(ii) There is a constant C and a 2π-periodic function $A : [0,1] \times \mathbb{R} \to \mathbb{C}$ with $A(u,-\lambda) = \overline{A(u,\lambda)}$ and

$$\sup_{t,\lambda} |A^0_{t,T}(\lambda) - A(t/T,\lambda)| \leq CT^{-1},$$

for all T; $A(u,\lambda)$ and $\mu(u)$ are assumed to be continuous in u.

Definition 2 (Dahlhaus, 1996 and 1997). The *(time-varying) spectral density* of the process (sequence of processes) is defined as:

$$f(u,\lambda) = A(u,\lambda)\overline{A(u,\lambda)} = |A(u,\lambda)|^2. \tag{15}$$

For these processes, Dahlhaus (1996) also defines the local covariance of lag k at time u, and gives kernel estimates of it, as well as of the spectral density. From the above definition, we propose the following spectral distribution function that could be estimated by our blockwise integrated periodogram.

Definition 3. The *(time-varying) spectral distribution* of the process (sequence of processes) is defined as:

$$F(u,\lambda) = \int_{-\pi}^{\lambda} f(u,l)dl. \tag{16}$$

Notice that if the underlying process (sequence of processes) is piecewise stationary, there exit some $u_1, u_2, \ldots, u_s$ such that $f(u,\lambda)$ is constant at intervals (u_i, u_{i+1}) as a function of u. Then the above definition leads to a piecewise constant spectral distribution. Of course, assuming the asymptotic framework proposed by Dahlhaus, the number of observations in each interval increases as T grows. The integrated periodogram, F_T, calculated using the observations inside a particular interval will provide a consistent estimator of the piecewise spectral distribution. Moreover, if we consider an increasing number of blocks, then most of these k blocks will be inside one of the intervals (u_i, u_{i+1}), $i = 1, 2, \ldots, s-1$, and therefore the length of the intervals that does not satisfy this inclusion property will be asymptotically negligible.

References

ABRAHAM, C., CORNILLON, P.A., MATZNER-LØBER, E. and MOLINARI, N. (2003), "Unsupervised Curve Clustering Using B-Splines", *Scandinavian Journal of Statistics,30(3)*, 581–595.

ALONSO, A.M., CASADO, D., LÓPEZ-PINTADO, S. and ROMO, J. (2008), "A Functional Data Based Method for Time Series Classification", Working Paper, Departamento de Estadística. Universidad Carlos III de Madrid, available at http://hdl.handle.net/10016/3381.

BIAU, G., BUNEA, F., and WEGKAMP, M.H. (2003), "Functional Classification in Hilbert Spaces", *IEEE Transactions on Information Theory, 1(11)*, 1–8.

BLANDFORD, R.R. (1993), "Discrimination of Earthquakes and Explosions at Regional Distances Using Complexity", Report AFTAC-TR-93-044 HQ, Air Force Technical Applications Center, Patrick Air Force Base, FL.

CASADO, D. (2013), *StatisCLAS: Methods for Statistical Classification*, Package of code, available at http://www.casado-d.org/edu/publications.html#Code.

CHANDLER, G., and POLONIK, W. (2006), "Discrimination of Locally Stationary Time Series Based on the Excess Mass Functional", *Journal of the American Statistical Association, 101(473)*, 240–253.

CLEVELAND, R.B., CLEVELAND, W.S., MCRAE, J.E., and TERPENNING, I. (1990), "STL: A Seasonal-Trend Decomposition Procedure Based on Loess", *Journal of Official Statistics, 6*, 2–73.

DAHLHAUS, R. (1996), "Asymptotic Statistical Inference for Nonstationary Processes with Evolutionary Spectra", in *Athens Conference on Applied Probability and Time Series Analysis*, eds. P.M. Robinson and M. Rosenblatt, New York: Springer.

DAHLHAUS, R. (1997), "Fitting Time Series Models to Nonstationary Processes", *The Annals of Statistics, 25(1)*, 1–37.

DIGGLE, P.J., and FISHER, N.I. (1991), "Nonparametric Comparison of Cumulative Periodograms", *Journal of the Royal Statistical Society, Series C (Applied Statistics), 40(3)*, 423–434.

FERRATY, F., and VIEU, P. (2003), "Curves Discrimination: A Nonparametric Functional Approach", *Computational Statistics and Data Analysis, 44*, 161–173.

HALL, P., POSKITT, D.S., and PRESNELL, B. (2001), "A Functional Data-Analytic Approach to Signal Discrimination", *Technometrics, 43(1)*, 1–9.

HASTIE, T., BUJA, A., and TIBSHIRANI, R.J. (1995), "Penalized Discriminant Analysis", *The Annals of Statistics, 23(1)*, 73–102.

HIRUKAWA, J. (2004), "Discriminant Analysis for Multivariate Non-Gaussian Locally Stationary Processes", *Scientiae Mathematicae Japonicae, 60(2)*, 357–380.

HODRICK, R., and PRESCOTT, E.C. (1997), "Postwar U.S. Business Cycles: An Empirical Investigation", *Journal of Money, Credit, and Banking, 29(1)*, 1–16.

HUANG, H., OMBAO, H. and STOFFER, D.S. (2004), "Discrimination and Classification of Nonstationary Time Series Using the SLEX Model", *Journal of the American Statistical Association, 99(467)*, 763–774.

JAMES, G.M., and HASTIE, T. (2001), "Functional Linear Discriminant Analysis for Irregularly Sampled Curves", *Journal of the Royal Statistical Society, Series B, 63*, 533–550.

JAMES, G.M., and SUGAR, C.A. (2003), "Clustering for Sparsely Sampled Functional Data", *Journal of the American Statistical Association, 98(462)*, 397–408.

KAKIZAWA, Y., SHUMWAY, R.H., and TANIGUCHI, M. (1998), "Discrimination and Clustering for Multivariate Time Series", *Journal of the American Statistical Association, 93(441)*, 328–340.

LIAO, T.W. (2005), "Clustering of Time Series Data Survey", *Pattern Recognition, 38*, 1857–1874.

LÓPEZ-PINTADO, S., and ROMO, J. (2006), "Depth-Based Classification for Functional Data", *DIMACS Series in Discrete Mathematics and Theoretical Computer Science (Vol. 72)*, Providence RI: American Mathematical Society.

LÓPEZ-PINTADO, S., and ROMO, J. (2009), "On the Concept of Depth for Functional Data", *Journal of the American Statistical Association, 104(486)*, 704–717.

MAHARAJ, E.A., and ALONSO, A.M. (2007), "Discrimination of Locally Stationary Time Series Using Wavelets", *Computational Statistics and Data Analysis, 52*, 879–895.

OMBAO, H.C., RAZ, J.A., VON SACHS, R., and B.A. MALOW, B.A. (2001), "Automatic Statistical Analysis of Bivariate Nonstationary Time Series", *Journal of the American Statistical Association, 96(454)*, 543–560.

PRIESTLEY, M. (1981), *Spectral Analysis and Time Series. Volume 1: Univariate Series*, London: Academic Press, Inc.

SAKIYAMA, K., and TANIGUCHI, M. (2004), "Discriminant Analysis for Locally Stationary Processes", *Journal of Multivariate Analysis, 90*, 282–300.

SHUMWAY, R.S. (2003), "Time-Frequency Clustering and Discriminant Analysis", *Statistics & Probability Letters, 63*, 307–314.

SHUMWAY, R.H., and STOFFER, D.S. (2000), *Time Series Analysis and Its Applications*, New York: Springer.

TANIGUCHI, M., and KAKIZAWA, Y. (2000), *Asymptotic Theory of Statistical Inference for Time Series*, New York: Springer.