Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 2668 – 2684
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A periodogram-based metric for time series classification

Jorge Caiado[a, b, *], Nuno Crato[a, c], Daniel Peña[d]

[a] *Centro de Matemática Aplicada à Previsão e Decisão Económica, Rua do Quelhas 6, 1200-781 Lisboa, Portugal*

[b] *Department of Economics and Management, Escola Superior de Ciências Empresariais, Instituto Politécnico de Setúbal, Campus do IPS, Estefanilha, 2914-503 Setúbal, Portugal*

[c] *Department of Mathematics, Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa, Rua do Quelhas 6, 1200-781 Lisboa, Portugal*

[d] *Department of Statistics, Universidad Carlos III de Madrid, Calle Madrid 126, 28903 Getafe, Spain*

Received 21 September 2004; received in revised form 4 April 2005; accepted 18 April 2005
Available online 17 May 2005

Abstract

The statistical discrimination and clustering literature has studied the problem of identifying similarities in time series data. Some studies use non-parametric approaches for splitting a set of time series into clusters by looking at their Euclidean distances in the space of points. A new measure of distance between time series based on the normalized periodogram is proposed. Simulation results comparing this measure with others parametric and non-parametric metrics are provided. In particular, the classification of time series as stationary or as non-stationary is discussed. The use of both hierarchical and non-hierarchical clustering algorithms is considered. An illustrative example with economic time series data is also presented.

Keywords: Autocorrelation function; Classification; Clustering; Euclidean distance; Periodogram; Stationary and non-stationary time series

[*] Corresponding author. Department of Economics and Management, Escola Superior de Ciências Empresariais, Instituto Politécnico de Setúbal, Campus do IPS, Estefanilha, 2914-503 Setúbal, Portugal. Tel.: +351 265 709 438; fax: +351 265 709 301.

E-mail address: jcaiado@esce.ips.pt (J. Caiado).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.04.012

1. Introduction

Classification and clustering time series is becoming an important area of research in several fields, such as economics, marketing, business, finance, medicine, biology, physics, psychology, zoology, and many others. For example, in Economics we may be interested in classifying the economic situation of a country by looking at some time series indicators, such as Gross National Product, investment expenditure, disposable income, unemployment rate or inflation rate. In Medicine, a patient may be classified in different classes using the information from the electrocardiogram time series.

The problem of identifying similarities or dissimilarities in time series data has been studied in the discrimination and clustering literature (see for instance Jonhson and Wichern, 1992). Some studies use non-parametric approaches for splitting a set of time series into clusters by looking at their Euclidean distances in the space of points. As pointed out by Galeano and Peña (2000), this metric has the important limitation of being invariant to transformations that modify the order of observations over time, and, therefore, it does not take into account the correlation structure of the time series. Piccolo (1990) introduced a metric for ARIMA models based on the autoregressive representation and applied this measure to the identification of similarities between industrial production series. Tong and Dabas (1990) investigated the affinity among some linear and non-linear fitted models by applying classical clustering techniques to the estimated residuals. Diggle and Fisher (1991) introduced a non-parametric approach to compare the spectrum of two time series based on the underlying cumulative periodograms. Diggle and al Wasel (1997) developed inference methods in spectral analysis based in the likelihood ratio to compare replicated time series data. Kakizawa et al. (1998) proposed parametric models for discriminating and clustering multivariate time series, with applications to environmental data (for discriminant analysis for time series, see also Shumway and Unger, 1974; Shumway, 1982; Dargahi-Noubary and Laycock, 1981; Dargahi-Noubary, 1992; Zhang and Taniguchi, 1994). Maharaj (2000) used a test of hypothesis in the comparison of two stationary time series based on the autoregressive parameters and proposed a classification method using the *p*-value of this test as a measure of similarity. Maharaj (2002) compared two non-stationary time series using the evolutionary spectra approach in order to take into account the structural changes over time. Other related works on clustering of time series are by Bohte et al. (1980), Kosmelj and Batagelj (1990), Shaw and King (1992), Maharaj (1999) and Xiong and Yeung (2004).

In this paper, we propose a metric based on the normalized periodogram and we use it for time series classification. We provide simulation results comparing this metric to the one by Piccolo (1990) and the ones based on autocorrelation, partial autocorrelation and inverse autocorrelation coefficients. In particular, we discuss the classification of time series as stationary or as non-stationary.

The remainder of the paper is organized as follows. In Section 2 we discuss briefly previous related methods on clustering time series and present our periodogram-based metrics. In Section 3 we discuss the methodology used for empirical classification of ARMA and ARIMA models and in Section 4 we present results from various approaches. In Section 5 we present an illustrative example with economic time series data to identify similarities among industrial production index series in United States, and in Section 6, we summarize the paper and discuss possible future research.

2. Time series metrics

A fundamental problem in classification analysis of time series is the choice of a relevant metric. Let $X_t = \left(x_{1,t}, \ldots, x_{k,t}\right)'$ be a vector time series with components represented by autoregressive integrated moving average or ARIMA(p, d, q) models,

$$\phi_i(B)(1-B)^d x_{i,t} = \theta_i(B)\varepsilon_{i,t}, \quad i = 1, \ldots, k, \tag{1}$$

where $\phi_i(B)$ is the autoregressive operator of order p and $\theta_i(B)$ is the moving average operator of order q; B is the back-shift operator and $(1-B)^d$ is the differencing operator of order d. The autoregressive and moving average polynomials in model (1) are assumed to have all roots outside the unit circle, so that each process $X_{i,t} = (1-B)^d x_{i,t}$ is causal and invertible.

Piccolo (1990) defined a metric for the class of invertible ARIMA models as the Euclidean distance between their autoregressive expansions. Let x_t be a zero mean stochastic process following an invertible ARIMA$(p, 0, q)$ model: $\phi(B)x_t = \theta(B)\varepsilon_t$. Then it can be represented by the AR(∞) operator $\pi(B) = \theta^{-1}(B)\phi(B) = 1 - \pi_1 B - \pi_2 B^2 - \cdots$, and the π coefficients contain all the information about the stochastic dependence structure of a time series. Piccolo (1990) introduced a metric by comparing the respective π sequences, defined by the distance

$$d_{\mathrm{PIC}}(x, y) = \sqrt{\sum_{j=1}^{\infty} \left(\pi_{j,x} - \pi_{j,y}\right)^2}. \tag{2}$$

An alternative metric suggested by Galeano and Peña (2000) is based on the estimated autocorrelation function (ACF). Suppose that we have a set of time series $X = \left(x_{1,t}, \ldots, x_{k,t}\right)'$ and $\hat{\rho}_i = \left(\hat{\rho}_{i,1}, \ldots, \hat{\rho}_{i,m}\right)$ is the vector of the estimated autocorrelation coefficients of the time series i for some m such that $\hat{\rho}_j \cong 0$ for $j > m$. This distance between the time series x e y is defined by

$$d_{\mathrm{ACF}}(x, y) = \sqrt{\left(\hat{\rho}_x - \hat{\rho}_y\right)' \Omega \left(\hat{\rho}_x - \hat{\rho}_y\right)}, \tag{3}$$

where Ω is some matrix of weights. When $\Omega = I$ (identity matrix), one obtains the Euclidean distance between the autocorrelation coefficients of time series x and y. When $\Omega = \left[\operatorname{cov}\left(\hat{\rho}\right)\right]^{-1}$ is the inverse covariance matrix of the autocorrelations, one obtains the Mahalanobis distance between the autocorrelations. It is also common to use weights that decrease with the autocorrelation lag.

Others possible distances that do not seem to have yet been applied in the clustering literature are the ones based on the partial autocorrelation function (PACF) and on the inverse autocorrelation function (IACF) introduced by Cleveland (1972) and developed by Chatfield (1979). For instance, a distance between the inverse autocorrelations can be given by

$$d_{\mathrm{IACF}}(x, y) = \sqrt{\left(\hat{\rho}_x^{(I)} - \hat{\rho}_y^{(I)}\right)' \Omega \left(\hat{\rho}_x^{(I)} - \hat{\rho}_y^{(I)}\right)}, \tag{4}$$

where the sample inverse autocorrelation functions $\hat{\rho}_x^{(I)}$ and $\hat{\rho}_y^{(I)}$ can be estimated by methods presented by Bhansali (1980, 1983), Battaglia (1983, 1986, 1988), Kanto (1987), and Subba Rao and Gabr (1989).

We now introduce a new distance based on the normalized periodogram. Let $P_x\left(w_j\right)=(1/n)|\sum_{t=1}^{n} x_t \mathrm{e}^{-\mathrm{i}tw_j}|^2$ and $P_y\left(w_j\right)=(1/n)|\sum_{t=1}^{n} y_t \mathrm{e}^{-\mathrm{i}tw_j}|^2$ be the periodograms of time series x and y, respectively, at frequencies $w_j=2\pi j/n,\ \ j=1,\dots,[n/2]$ in the range 0 to π (where $[n/2]$ is the largest integer less or equal to $n/2$). A distance between x and y can be defined by

$$d_{\mathrm{P}}(x,y)=\sqrt{\sum_{j=1}^{[n/2]}\left[P_x\left(w_j\right)-P_y\left(w_j\right)\right]^2}. \tag{5}$$

If we are not interested in the process scale, but only on its correlation structure, it is better to use the normalized periodogram (or rescaled periodogram) by replacing $P\left(w_j\right)$ in (5) by $NP\left(w_j\right)=P\left(w_j\right)/\hat{\gamma}_0$, where $\hat{\gamma}_0$ is the sample variance of the time series, that is

$$d_{\mathrm{NP}}(x,y)=\sqrt{\sum_{j=1}^{[n/2]}\left[NP_x\left(w_j\right)-NP_y\left(w_j\right)\right]^2}. \tag{6}$$

Since the variance of periodogram ordinates is proportional to the spectrum value at the corresponding frequencies, it makes sense to use the logarithm of the normalized periodogram,

$$d_{\mathrm{LNP}}(x,y)=\sqrt{\sum_{j=1}^{[n/2]}\left[\log NP_x\left(w_j\right)-\log NP_y\left(w_j\right)\right]^2}. \tag{7}$$

It is straightforward to show that distances (6) and (7) satisfy the usual properties of a metric: $d(x,y)=d(y,x)$ (symmetry); $d(x,y)>0$, with $x\neq y$ (non-negativity); and $d(x,y)\leqslant d(x,z)+d(z,y)$ (triangle inequality).

As can be expected, measures based on the autocorrelations and measures based on the periodogram are related. It is well known that the periodogram has the equivalent representation $P\left(w_j\right)=2\left[\hat{\gamma}_0+2\sum_{k=1}^{n-1}\hat{\gamma}_k\cos\left(w_jk\right)\right]$, where $\hat{\gamma}_k$ is the sample autocovariance function (for details, see Brockwell and Davis, 1991; Wei, 1990), and dividing $P\left(w_j\right)$ by $\hat{\gamma}_0$, we obtain the normalized periodogram given by

$$NP\left(w_j\right)=2\left[1+2\sum_{k=1}^{n-1}\hat{\rho}_k\cos\left(w_jk\right)\right], \tag{8}$$

which is the transform of the sample autocorrelation function and vice versa. The relation between the normalized periodogram metric (6) and the ACF metric (3) can be expressed by

$$\begin{aligned} d_{\mathrm{NP}}(x, y) &= \sqrt{\sum_{j=1}^{[n/2]} \left[NP_x\left(w_j\right) - NP_y\left(w_j\right)\right]^2} \\ &= \sqrt{\sum_{j=1}^{[n/2]} \left[\left(2 + 4\sum_{k=1}^{n-1} \hat{\rho}_{k,x} \cos\left(w_j k\right)\right) - \left(2 + 4\sum_{k=1}^{n-1} \hat{\rho}_{k,y} \cos\left(w_j k\right)\right)\right]^2} \\ &= \sqrt{\sum_{j=1}^{[n/2]} \left[4\sum_{k=1}^{n-1} \cos\left(w_j k\right)\left(\hat{\rho}_{k,x} - \hat{\rho}_{k,y}\right)\right]^2}. \end{aligned} \tag{9}$$

By the orthogonality properties of the cosine functions, $\sum_{j=1}^{[n/2]} \cos^2\left(w_j k\right) = n/4$ (if n is even) and $\sum_{j=1}^{[n/2]} \cos\left(w_j k\right)\cos\left(w_j s\right) = 0$ (for $k \neq s$), we get

$$\begin{aligned} d_{\mathrm{NP}}(x, y) &= \sqrt{16\left[\frac{n}{4}\left(\hat{\rho}_{1,x} - \hat{\rho}_{1,y}\right)^2 + \cdots + \frac{n}{4}\left(\hat{\rho}_{n-1,x} - \hat{\rho}_{n-1,y}\right)^2\right]} \\ &= 2\sqrt{n}\sqrt{\sum_{k=1}^{n-1}\left(\hat{\rho}_{k,x} - \hat{\rho}_{k,y}\right)^2}, \end{aligned} \tag{10}$$

or, using matrix notation,

$$\begin{aligned} d_{\mathrm{NP}}(x, y) &= \left(2\sqrt{n}\right)\sqrt{\left(\hat{\rho}_x - \hat{\rho}_y\right)' I \left(\hat{\rho}_x - \hat{\rho}_y\right)} \\ &= \left(2\sqrt{n}\right) d_{\mathrm{ACF}}(x, y). \end{aligned} \tag{11}$$

These two measures are thus equivalent. However, their application with a truncated number of autocorrelations or a truncated number of periodogram ordinates could yield different results.

Another useful measure in the time domain, based in the Kullback–Leibler information distance (KLTD measure), is defined by

$$d_{\mathrm{KLTD}}(x, y) = \operatorname{tr}\left(R_x R_y^{-1}\right) - \log\left(|R_x|/|R_y|\right) - n, \tag{12}$$

where R_x and R_y are the $L \times L$ autocorrelation matrices of time series x and y, respectively, made at L successive times. The Kullback–Leibler information distance in the frequency domain (KLFD measure) is asymptotically equivalent to (12), and much easier to compute.

It is given by

$$d_{\mathrm{KLFD}}(x, y) = \sum_{j=1}^{[n/2]} \left[\frac{NP_x\left(w_j\right)}{NP_y\left(w_j\right)} - \log \frac{NP_x\left(w_j\right)}{NP_y\left(w_j\right)} - 1 \right]. \tag{13}$$

This measure is greater or equal to zero, with equality if and only if $NP_x\left(w_j\right) = NP_y\left(w_j\right)$ almost everywhere. Its potentially success should by related with metric (7). The likelihood ratio statistic for testing equality of spectra is given by,

$$\log\ L = \sum_{j=1}^{[n/2]} \log\ \frac{\left[NP_x\left(w_j\right) NP_y\left(w_j\right)\right]^{1/2}}{\frac{1}{2}\left[NP_x\left(w_j\right) + NP_x\left(w_j\right)\right]}, \tag{14}$$

and it is distributed proportionally to a chi-square random variable and it could also be used for measuring distances between spectra. This statistic can also be expressed equivalently as the sum of differences between the average log spectra and the log of the spectra average.

3. Methodology of time series classification

In this section we will use the following previously discussed distances for time series classification:

Step 1: Find similarities or dissimilarities between every pair of time series in the data set. For each data we compute a distance matrix with $k(k-1)/2$ different pairs using the following metrics:

(i) Classical Euclidean (EUCL) distance, $d_{\mathrm{EUCL}}(x, y) = \sqrt{\sum_{t=1}^{n} (x_t - y_t)^2}$.

(ii) Piccolo's distance, $d_{\mathrm{PIC}}(x, y) = \sqrt{\sum_{j=1}^{\infty} \left(\pi_{j,x} - \pi_{j,y}\right)^2}$. The application of this distance requires the fitting of an ARIMA model to the time series. This has been done by fitting ARMA(p, d, q) models to the series with $d = 0, 1$, $p = 0, 1, 2, 3$ and $q = 0, 1, 2, 3$, and selecting the order by three possible model selection criteria. The first is the AIC, the second the AICC and the third the BIC criterion. These criteria have been applied as recommended by Beran et al. (1998). Then the π_i weights are obtained and used to compute the distance matrices.

(iii) ACF distance. We implemented three possible ways of computing a distance by using the autocorrelation coefficients. The first uses a uniform weighting and is equivalent to the Euclidean distance between autocorrelations coefficient vectors, $d_{\mathrm{ACF}}(x, y) = \sqrt{\sum_{i=1}^{L} \left(\hat{\rho}_{i,x} - \hat{\rho}_{i,y}\right)^2}$. The second uses a geometric decay, $d_{\mathrm{ACFG}}(x, y) = \sqrt{\sum_{i=1}^{L} m_i \left(\hat{\rho}_{i,x} - \hat{\rho}_{i,y}\right)^2}$,where L is the number of autocorrelations, $m_i = pq^i$ for $i = 1, \ldots, L$, $p = 1 - q$ and $0 < p < 1$. The third uses the Mahalanobis distance, $d_{\mathrm{ACFM}}(x, y) = \sqrt{\left(\hat{\rho}_x - \hat{\rho}_y\right)' \Omega^{-1} \left(\hat{\rho}_x - \hat{\rho}_y\right)}$, where Ω is the sample covariance matrix of the autocorrelation coefficients with elements given by the truncated Bartlett's formula (Brockwell and Davis, 1991, p. 222). That is, assuming that time series x and y are independent, and denoting $w_{ij} = \sum_{k=1}^{L} \left[\left(\hat{\rho}_{(k+i)} \hat{\rho}_{(k-i)} - 2\hat{\rho}_{(i)} \hat{\rho}_{(k)} \right) \left(\hat{\rho}_{(k+j)} \ \hat{\rho}_{(k-j)} \ -2\hat{\rho}_{(j)} \hat{\rho}_{(k)} \right) \right]$,then the variances and covariances in Ω are given by

$$\operatorname{var}\left(\hat{\rho}_x - \hat{\rho}_y\right) = \operatorname{var}\left(\hat{\rho}_x\right) + \operatorname{var}\left(\hat{\rho}_y\right) = w_{ii,x} + w_{ii,y},$$

$$\operatorname{cov}\left(\hat{\rho}_{i,x} - \hat{\rho}_{i,y}, \hat{\rho}_{j,x} - \hat{\rho}_{j,y}\right) = \operatorname{cov}\left(\hat{\rho}_{i,x}, \hat{\rho}_{j,x}\right) + \operatorname{cov}\left(\hat{\rho}_{i,y}, \hat{\rho}_{j,y}\right) = w_{ij,x} + w_{ij,y}.$$

(iv) PACF Euclidean distance, $d_{\text{PACF}}(x, y) = \sqrt{\sum_{i=1}^{L} \left(\hat{\phi}_{ii,x} - \hat{\phi}_{ii,y} \right)^2}$, where $\hat{\phi}_{ii}$ are the sample partial autocorrelations. We also explored weighting these coefficients, but as the results were similar to the uniform weighting we do not report them here.

(v) IACF Euclidean distance, $d_{\text{IACF}}(x, y) = \sqrt{\sum_{i=1}^{L} \left(\hat{\rho}_{i,x}^{(I)} - \hat{\rho}_{i,y}^{(I)} \right)^2}$, where $\hat{\rho}^{(I)}$ are inverse autocorrelation estimates calculated from the autocorrelation function using approximation Kanto's formula (Kanto, 1987) for ARMA processes, that is

$$\hat{\rho}^{(I)} = (A + B)^{-1} \hat{\rho},$$

where $\hat{\rho}^{(I)} = \left(\hat{\rho}_1^{(I)}, \ldots, \hat{\rho}_L^{(I)} \right)$, $\hat{\rho} = \left(\hat{\rho}_1, \ldots, \hat{\rho}_L \right)$, $A = \left\{ \hat{\rho}_{L-1} \right\}_{L \times L}$, and $B = \left\{ \hat{\rho}_{L+1} \right\}_{L \times L}$. As in the PACF metric, we only give the results for uniform weighting of these coefficients.

(vi) Distance based on the log-normalized periodogram (LNP metric),

$$d_{\text{LNP}}(x, y) = \sqrt{\sum_{j=1}^{[n/2]} \left[\log NP_x\left(w_j\right) - \log NP_y\left(w_j\right) \right]^2}.$$

(vii) KL information distance, $d_{\text{KL}}(x, y) = \sum_{j=1}^{[n/2]} \left[\frac{NP_x(w_j)}{NP_y(w_j)} - \log \frac{NP_x(w_j)}{NP_y(w_j)} - 1 \right]$.

We also investigated the performance of the likelihood ratio statistic (14) to compare the spectrum of the stationary and non-stationary time series, but the results are not included as it did not work well.

Step 2: Group the time series into two clusters (stationary and non-stationary) using an appropriate agglomerative hierarchical clustering algorithm as the single linkage (which maximizes the minimum distance between objects in the same group), the complete linkage (which minimizes the maximum distance between objects in the same group) or the average linkage (which averages the distances between objects in different groups). These linkages algorithms are concerned with the partition of a set of objects into groups or clusters, in such a way that objects in the same group are similar to one another and objects in different clusters are as distinct as possible. For more details, see for instance, Jonhson and Wichern (1992) and Gordon (1996). Alternatively, we may use a non-hierarchical clustering procedure such as the k-means algorithm, where one determines a preliminary set of k clusters, move each time series to the cluster whose centroid is closest in Euclidean distance, recalculate the cluster centroid and repeat the reassignment procedure until no time series is reassigned. The implementation of the k-means clustering algorithm in our approach is based on Euclidean distances among standardized observations, autoregressive weights, autocorrelation coefficients, partial autocorrelation coefficients, inverse autocorrelation coefficients, and normalized periodogram ordinates in the logarithm scale.

The autocorrelation coefficients and the spectrum are usually defined for stationary processes but their definition can be extended for integrated processes (see Peña and Poncela, 2005). With the usual definition for stationary time series (S), $\hat{\rho}_k^S \to 0$ as $k \to \infty$ and $n \to \infty$, and for non-stationary time series (NS), for any fixed k, $\hat{\rho}_k^{\mathrm{NS}} \to 1$ as $n \to \infty$. Hence, $\sum_{k=1}^{m} \left(\hat{\rho}_k^{\mathrm{NS}} - \hat{\rho}_k^S\right)^2 \to \infty$ as $m \to \infty$ and $n \to \infty$ with $m < n$. Consequently, by (10) it follows that $\sum \left(NP^{\mathrm{NS}}(w_j) - NP^S(w_j)\right)^2$ also diverges. Thus, the estimated ACF and the normalized periodogram should be able to discriminate between stationary and non-stationary time series.

4. Simulation results

We simulated one thousand time series replications of each of the following six stationary [(a)–(f)] and six non-stationary [(g)–(l)] models. All the series have zero mean and unit variance white noise. The samples sizes were taken equal to 50, 100, 200, 500, 1000 and 10000 observations:

Model (a): AR(1), with $\phi_1 = 0.9$;
Model (b): AR(2), with $\phi_1 = 0.95$ and $\phi_2 = -0.1$;
Model (c): ARMA(1,1), with $\phi_1 = 0.95$ and $\theta_1 = 0.1$;
Model (d): ARMA(1,1),with $\phi_1 = -0.1$ and $\theta_1 = -0.95$;
Model (e): MA(1), with $\theta_1 = -0.9$;
Model (f): MA(2), with $\theta_1 = -0.95$ and $\theta_2 = -0.1$.
Model (g): ARIMA(1,1,0), with $\phi_1 = -0.1$;
Model (h): ARIMA(0,1,0);
Model (i): ARIMA(0,1,1), with $\theta_1 = 0.1$;
Model (j): ARIMA(0,1,1), with $\theta_1 = -0.1$;
Model (k): ARIMA(1,1,1), with $\phi_1 = 0.1$ and $\theta_1 = -0.1$;
Model (l): ARIMA(1,1,1), with $\phi_1 = 0.05$ and $\theta_1 = -0.05$.

We have chosen all the models with parameter values close to the random walk in order to make it not easy to classify time series into stationary and non-stationary. Fig. 1 shows the typical shape of each series. We group the series into two clusters (stationary and non-stationary) and we obtain the percentage of successes in the classification. Table 1 gives the simulated results by using the complete linkage hierarchical clustering algorithm. In Table 2, we present the k-means clustering method. All simulations were obtained by using MATLAB 6.

From Table 1, it can be seen that both the Euclidean distance and Piccolo's metric are unable to distinguish successfully between stationary and non-stationary time series, since the percentages of successes obtained with those measures did not exceed 68%, even with large samples. In fact, the metric of Piccolo cannot discriminate between the ARMA models with large autoregressive coefficients [models (a), (b) and (c)] and the ARIMA models (g)–(l). This can be easily understood by noting that the autoregressive weights π_i of both

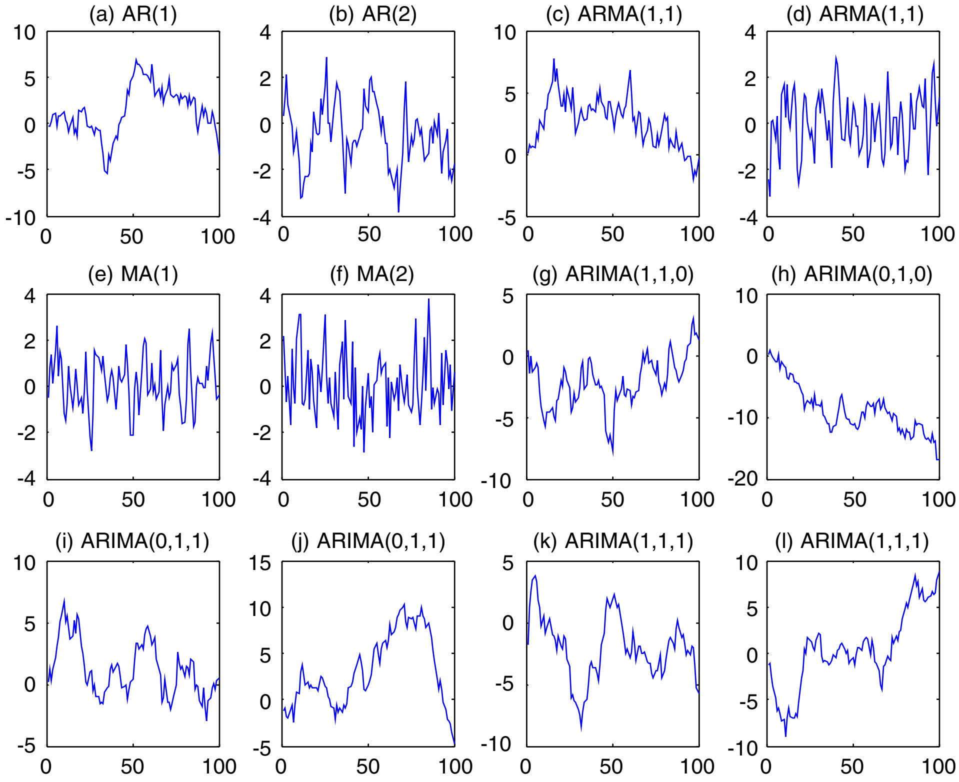

Fig. 1. Stationary and non-stationary simulated time series.

these two classes of models are similar for the first lags.

The ACFU metric, the ACFG metric, the KL metric and the LNP metric produced high percentages of success that increase with the sample size. Both ACF metrics seem to work better when using $L = n/10$ autocorrelations, when n is the sample size. The LNP metric works quite well in all cases (low, high and all frequencies). For time series with 10 000 observations, it can be seen that the ACF metrics (uniform and geometric decay), the KL metric and the LNP metric can distinguish perfectly (with percentages of success of 100%) between stationary and non-stationary time series. The ACF with uniform weighting works worse for $L = 1000$, 2500 and 5000. This may be due to the noise pattern of the correlogram of stationary time series for large lags. For all the samples, the LNP and KL metrics provide slightly better results than the ACF metrics, and the KL metric works very well for high frequency components. The ACF Mahalanobis, the PACF and the IACF distance-based methods seem to be poor metrics in this comparison.

From Table 2, it can be seen that the k-means clustering algorithm gives similar results to the ones obtained by the hierarchical clustering procedure. This shows the robustness of the proposed approaches.

Table 1
Percentages of success on time series classification by hierarchical method

n	EUCL	L	π weights			Autocorrelations, partial and inverse autoc.					Periodogram ordinates		
			AIC	AICC	BIC	ACFU	ACFG	ACFM	PACF	IACF	Freq	LNP	KL
50	**67.44**	5	66.01	66.12	66.95	76.43	76.41	73.51	68.92	**66.19**			
		10	66.17	66.67	66.92	**76.73**	**78.08**	73.92	**70.33**	65.26	Low	76.61	70.58
		15	**66.75**	**67.03**	66.97	76.58	76.12	73.67	67.83	64.61	High	77.75	**80.17**
		25	66.68	66.92	**67.02**	73.57	75.38	**75.00**	58.92	NA	All	**78.88**	74.42
100	**66.67**	5	65.67	65.87	66.87	77.04	76.68	73.33	72.50	66.20			
		10	65.73	66.00	67.01	**81.49**	80.88	73.83	72.08	**66.76**	Low	82.57	76.82
		25	**66.07**	**66.13**	**67.10**	80.67	80.62	74.72	**72.58**	65.17	High	83.04	**91.58**
		50	65.87	66.07	67.08	79.87	**81.31**	**75.08**	58.33	NA	All	**83.33**	84.75
200	**67.24**	5	65.60	65.73	66.94	75.81	75.51	74.50	73.01	**66.57**			
		10	**66.33**	65.80	67.11	82.83	81.30	75.08	71.83	66.48	Low	88.32	83.84
		20	66.27	**66.53**	**67.07**	**88.42**	**87.50**	75.15	72.83	66.39	High	89.21	**96.42**
		50	65.87	65.93	66.99	84.75	87.43	**75.21**	**73.33**	66.51	All	**90.08**	88.96
		100	66.19	66.39	66.98	83.33	87.17	NA	57.92	NA			
500	**67.38**	5	65.87	65.60	66.89	75.18	75.05	74.88	73.48	**66.67**			
		10	66.00	65.67	66.92	81.42	79.18	75.00	74.02	**66.67**	Low	92.83	92.25
		25	66.13	65.87	66.75	95.15	94.14	75.08	73.51	66.57	High	**98.01**	**98.87**
		50	65.33	**66.47**	66.67	**95.37**	**95.33**	**75.22**	**75.01**	65.92	All	97.73	96.74
		125	65.87	66.13	**67.03**	88.66	95.29	NA	73.33	64.61			
		250	**66.92**	66.07	67.01	87.42	94.67	NA	57.85	NA			

1000	**67.05**	5	65.27	65.67	66.93	75.00	75.01	74.72	74.67	**66.67**			
		10	65.87	66.13	66.99	80.34	78.08	74.92	74.33	**66.67**	Low	95.08	96.49
		25	66.00	**67.00**	66.98	97.28	96.00	**75.08**	74.03	**66.67**	High	99.58	**99.58**
		50	66.33	66.66	**67.02**	**98.85**	98.43	**75.08**	75.00	66.10	All	**99.83**	99.23
		100	**66.40**	66.67	67.01	97.98	**98.67**	NA	**75.01**	64.98			
		250	66.07	66.00	66.98	90.35	98.57	NA	74.38	64.68			
		500	66.02	66.13	66.97	89.58	98.17	NA	58.91	NA			
10 000	**67.65**	50	66.67	65.83	66.71	**100.0**	**100.0**	**75.00**	74.97	**66.67**			
		100	66.27	66.07	66.89	**100.0**	**100.0**	NA	75.49	**66.67**	Low	97.83	99.17
		250	65.83	65.53	66.97	**100.0**	**100.0**	NA	**76.30**	**66.67**	High	**100.0**	**100.0**
		500	65.97	65.89	66.95	**100.0**	**100.0**	NA	75.78	65.83	All	**100.0**	**100.0**
		1000	**66.33**	66.13	66.99	98.08	**100.0**	NA	75.43	64.92			
		2500	66.07	66.57	**67.01**	91.02	**100.0**	NA	73.85	NA			
		5000	66.01	**66.72**	66.98	88.01	**100.0**	NA	59.45	NA			

Notes: n is the sample size; EUCL is the Euclidean metric; L is the number of autoregressive weights (Piccolo's metric with model selection criteria AIC, AICC and BIC), autocorrelations (ACF uniform metric, ACF geometric decay metric with $p = 0.05$, ACF Mahalanobis metric), partial autocorrelations (PACF metric) and inverse autocorrelations (IACF metric); "low" frequencies of the LNP and KL metrics correspond to ordinates 1 to $[\sqrt{n}]$ and "high" frequencies to ordinates $[\sqrt{n} + 1]$ to $n/2$. The higher percentages are indicated in bold. The results were based on 1000 simulations of each time series, except for Piccolo's metric, ACFM, IACF and for $n = 10\,000$, which were based on 100 simulations.

Table 2
Percentages of success on time series classification by *k*-means method

n	EUCL	*L*	AIC	AICC	BIC	ACF	PACF	IACF	Freq	LNP
50	**67.67**	5	67.84	68.00	68.08	76.77	69.33	63.58		
		10	67.92	68.51	68.08	**78.54**	**66.75**	**65.42**	Low	76.93
		15	68.33	**68.74**	68.17	76.67	66.13	64.42	High	76.42
		25	**68.42**	68.14	**68.33**	73.67	59.34	NA	All	**77.61**
100	**69.75**	5	67.92	67.92	68.08	78.40	**72.21**	**68.82**		
		10	67.92	68.00	68.17	81.88	71.55	65.33	Low	84.43
		25	68.42	68.17	68.67	**82.50**	66.67	64.75	High	84.01
		50	**68.58**	**68.50**	**68.83**	79.21	59.53	NA	All	**84.39**
200	**71.33**	5	67.84	68.08	68.00	77.03	71.46	65.31		
		10	67.92	68.25	68.25	84.01	**72.73**	**66.01**	Low	88.83
		20	68.00	68.33	**68.67**	**89.29**	70.37	65.75	High	88.75
		50	**68.25**	68.33	**68.67**	87.79	68.37	64.33	All	**92.25**
		100	68.17	**68.50**	68.33	85.79	60.11	NA		
500	**71.50**	5	68.00	68.00	67.92	76.30	72.58	65.65		
		10	68.08	68.17	68.08	84.79	**72.64**	65.58	Low	94.17
		25	68.17	68.33	**68.42**	94.02	71.47	**65.83**	High	97.63
		50	68.33	68.42	**68.42**	**96.88**	70.09	66.58	All	**97.67**
		125	**68.42**	68.67	68.33	94.25	66.89	64.42		
		250	**68.42**	**68.75**	68.17	92.17	59.74	NA		

1000	**70.42**	5	67.84	68.08	68.08	75.67	72.13	65.82		
		10	68.00	68.08	68.17	85.54	72.05	**65.92**	Low	96.52
		25	68.33	68.17	68.33	96.29	71.97	**65.92**	High	99.62
		50	68.33	68.33	68.42	**99.17**	**72.65**	65.67	All	**99.88**
		100	**68.42**	**68.67**	**68.67**	98.75	66.83	64.21		
		250	68.33	**68.67**	68.33	96.08	60.18	64.05		
		500	68.33	**68.33**	68.17	94.42	58.34	NA		
10 000	**70.14**	50	67.92	**68.08**	68.17	**100.00**	72.92	66.67		
		100	68.08	68.00	68.33	**100.00**	**73.08**	**65.83**	Low	98.61
		250	68.17	68.17	68.42	**100.00**	72.92	65.00	High	**100.00**
		500	68.33	68.17	68.67	**100.00**	72.50	64.58	All	**100.00**
		1000	68.33	**68.67**	**68.83**	99.92	66.67	64.12		
		2500	**68.67**	68.33	68.67	99.67	65.83	63.39		
		5000	**68.67**	68.42	68.67	98.75	63.39	NA		

Notes: n is the sample size; EUCL is the Euclidean metric; L is the number of autoregressive weights (Piccolo's metric with model selection criteria AIC, AICC and BIC), autocorrelations (ACF metric), partial autocorrelations (PACF metric) and inverse autocorrelations (IACF metric); "low" frequencies of the LNP and KL metrics correspond to ordinates 1 to $[\sqrt{n}]$ and "high" frequencies to ordinates $[\sqrt{n}+1]$ to $n/2$. The higher percentages are indicated in bold. The results were based on 1000 simulations of each time series, except for Piccolo's metric, IACF and for $n = 10\,000$, which were based on 100 simulations.

Table 3
Industrial production indices series in United States (US)

No.	Series	No.	Series
1	Manufacturing	11	Non-ferrous metals
2	Primary processing	12	Fabricated metal products
3	Advanced processing	13	Industrial machinery and equipment
4	Durable manufacturing	14	Computer and office equipment
5	Lumber and products	15	Electrical machinery
6	Furniture and fixtures	16	Transportation equipment
7	Stone, clay, and glass products	17	Motor vehicles and parts
8	Primary metals	18	Autos and light trucks
9	Iron and steel	19	Aerospace and miscellaneous transp. equip.
10	Raw steel	20	Instruments

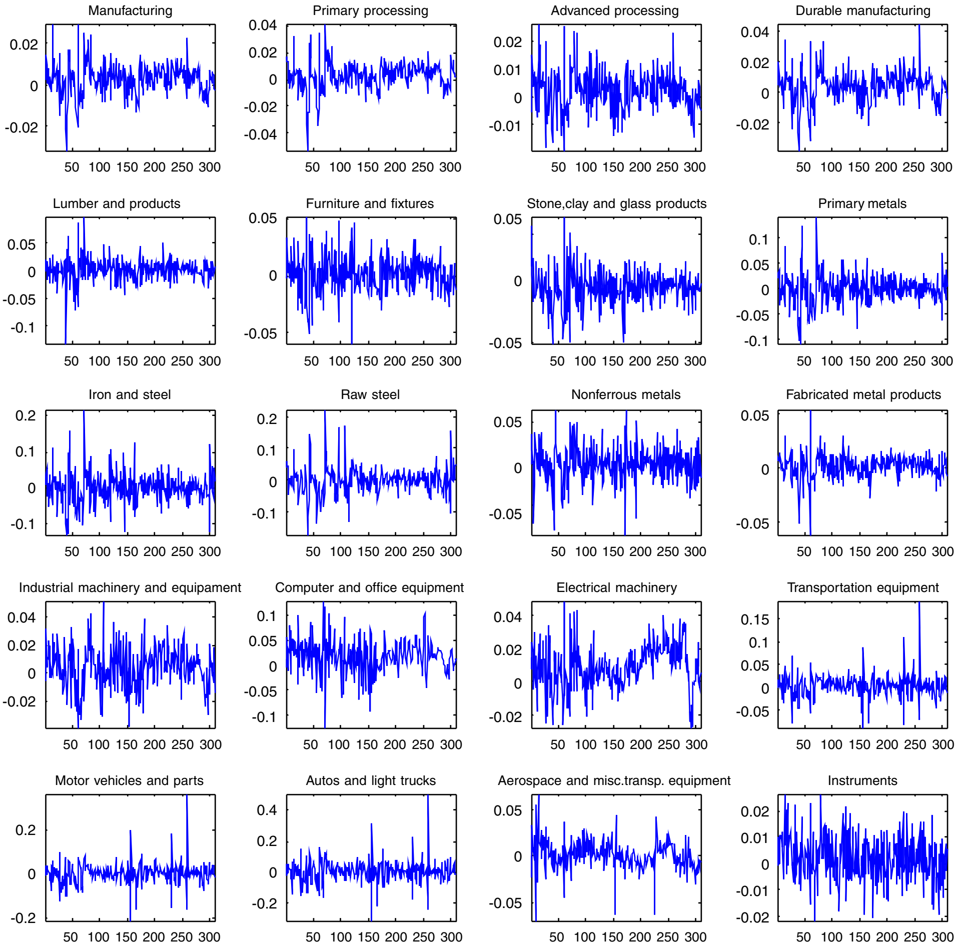

Fig. 2. Plots of industrial production differenced log series in US.

5. Application

As an illustrative example we use the Industrial Production (by Market Group) indices in United States (source: http://www.economagic.com). The 20 time series indices (seasonally adjusted) with sample sizes of $n = 309$, from January 1977 to September 2002, are reported in Table 3.

Before carrying out clustering analysis, the series were transformed in differences of the logarithm, $\log x_t - \log x_{t-1}$, as shown in Fig. 2, in order to get the percentages increases from period to period. This gets rid of the low frequency trends, forcing the metrics to work on the stationary parts of the series. In Fig. 3 we can see the dendrogram of the hierarchical cluster tree (complete linkage method) of the Industrial Production series by using the LNP metric with low frequencies.

The choice of the number of clusters in the data is sometimes subjective and depends on the researcher experience. However, we may find a natural partition in the data set by using the so-called inconsistency coefficient (see Statistics Toolbox User's Guide, 2001), which compares the length of each link in a cluster tree with the average length of all the other links. The LNP metric split the series data into three clusters: C1 = {1, 2, 3, 4, 6, 7, 12, 13, 14, 19, 20}, C2={8, 9, 10} and C3={5, 11, 16, 17, 18}, and separate from the others the series 15 (Computer and office equipment). Group C1 includes the fast growing sectors, group C2 includes sectors with some large negative growth in the period, and group D1 includes sectors of steady growth or with a small decay during the considered period.

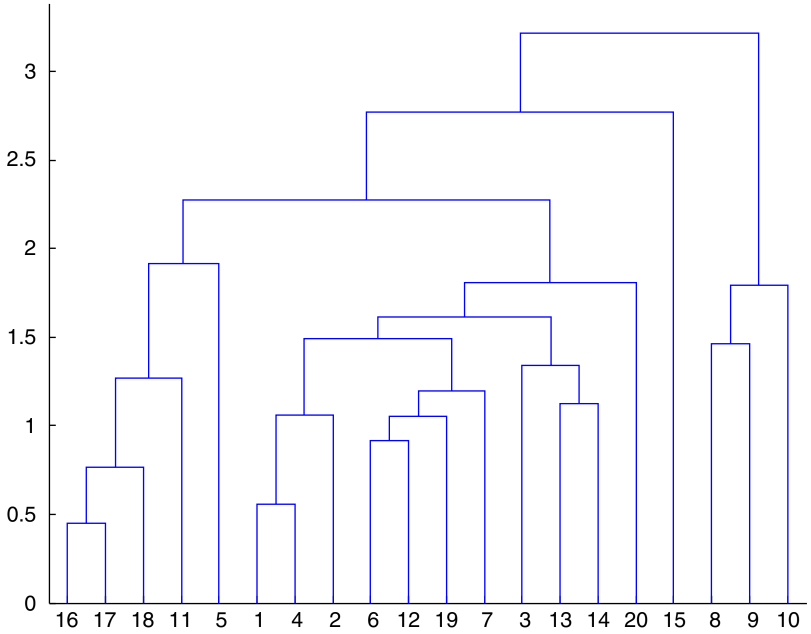

Fig. 3. Dendrogram of industrial production differenced log series in US by using the LNP metric.

6. Conclusions

In this paper, we have studied metrics based on different dependence measures to classify time series as stationary or as non-stationary. Simulation results show that the metrics based on the logarithm of the normalized periodogram and the metric based on the autocorrelation coefficients can all distinguish empirically with high success ARMA from ARIMA models, while this does not happen with the classic Euclidean distance nor with the metric based in the autoregressive weights proposed by Piccolo (1990). The first metrics above have also an important advantage over the distance-based methods proposed by Piccolo (1990) and by Tong and Dabas (1990): they do not need to fit previously an ARIMA model in order to compute the distances.

We have shown that the normalized periodogram metric is closely related to the autocorrelation function metric. However, the use of the normalized periodogram metric in the logarithm scale seems to provide better results than the normalized periodogram metric in levels or, equivalently, the ACF metric with uniform weights.

While we only studied the correlation between stationary and non-stationary processes, the methods we developed can be used to distinguish other type processes, namely different classes of stationary time series. One possible extension is the comparison between linear and non-linear time series processes.

Acknowledgements

This research was partially supported by a grant from the Fundação para a Ciência e Tecnologia (POCTI/FCT) and by a grant from the Fundação Calouste Gulbenkian. The third author acknowledges support from grant SEJ2004-03303 and from Fundación BBVA, Spain. Part of this work was completed during the visit of Jorge Caiado to the Department of Statistics, Universidad Carlos III de Madrid, Spain. The authors gratefully acknowledge the helpful comments and suggestions of associate editor and an anonymous referee on an earlier version of this paper.

References

Battaglia, F., 1983. Inverse autocovariances and a measure of linear determinism for a stationary process. J. Time Ser. Anal. 4, 79–87.

Battaglia, F., 1986. Recursive estimation of the inverse correlation function. Statistica 46, 75–82.

Battaglia, F., 1988. On the estimation of the inverse correlation function. J. Time Ser. Anal. 9, 1–10.

Beran, J., Bhansali, R.J., Ocker, D., 1998. On unified model selection for stationary and nonstationary short and long memory autoregressive processes. Biometrika 85, 921–934.

Bhansali, R.J., 1980. Autoregressive and window estimates of the inverse autocorrelation function. Biometrika 67, 551–566.

Bhansali, R.J., 1983. A simulation study of autoregressive and window estimators of the inverse correlation function. Appl. Statist. 32, 141–149.

Bohte, Z.D., Cepar, D., Kosmelu, K., 1980. Clustering of time series. Proc. COMPSTAT 80, 587–593.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods. second ed. Springer, New York.

Chatfield, C., 1979. Inverse autocorrelations. J. Roy. Statist. Soc. Ser. A 142, 363–377.

Cleveland, W.S., 1972. The inverse autocorrelations of a time series and their applications. Technometrics 14, 277–293.

Dargahi-Noubary, G.R., 1992. Discrimination between Gaussian time series based on their spectral differences. Commun. Statist. Theory Methods 21, 2439–2458.

Dargahi-Noubary, G.R., Laycock, P.J., 1981. Spectral ratio discriminant and information theory. J. Time Ser. Anal. 2, 70–86.

Diggle, P.J., Fisher, N.I., 1991. Nonparametric comparison of cumulative periodograms. Appl. Statist. 40, 423–434.

Diggle, P.J., al Wasel, I., 1997. Spectral analysis of replicated biomedical time series. Appl. Statist. 46, 31–71.

Galeano, P., Peña, D., 2000. Multivariate analysis in vector time series. Resenhas 4, 383–404.

Gordon, A.D., 1996. Hierarchical classification. In: Arabie, P., Hubert, L.J., De Soete, G. (Eds.), Clustering and Classification. Word Scientific Publishing, River Edge, NJ.

Jonhson, R.A., Wichern, D.W., 1992. Applied Multivariate Statistical Analysis. third ed. Prentice-Hall, Englewood Cliffs.

Kakizawa, Y., Shumway, R.H., Taniguchi, M., 1998. Discrimination and clustering for multivariate time series. J. Amer. Statist. Assoc. 93, 328–340.

Kanto, A.J., 1987. A formula for the inverse autocorrelation function of an autoregressive process. J. Time Ser. Anal. 8, 311–312.

Kosmelj, K., Batagelj, V., 1990. Cross-sectional approach for clustering time varying data. J. Classification 7, 99–109.

Maharaj, E.A., 1999. Comparison and classifying of stationary multivariate time series. Pattern Recognition 32, 1129–1138.

Maharaj, E.A., 2000. Clusters of time series. J. Classification 17, 297–314.

Maharaj, E.A., 2002. Comparison of non-stationary time series in the frequency domain. Comput. Statist. Data Anal. 40, 131–141.

Peña, D., Poncela, P., 2005. Nonstationary dynamic factor models. J. Statist. Planning and Inference, in press.

Piccolo, D., 1990. A distance measure for classifying ARIMA models. J. Time Ser. Anal. 11, 152–164.

Shaw, C.T., King, G.P., 1992. Using cluster analysis to classify time series. Physica D 58, 288–298.

Shumway, R.H., 1982. Discriminant analysis for time series. In: Krishnaiah, P.R., Kanals, L.N. (Eds.), Handbook of Statistics, vol. 1, pp. 1–46.

Shumway, R.H., Unger, A.N., 1974. Linear discriminant function for stationary time series. J. Amer. Statist. Assoc. 69, 948–956.

Statistics Toolbox User's Guide, 2001. Matlab, fifth ed. The MathWorks.

Subba Rao, T., Gabr, M.M., 1989. The estimation of spectrum, inverse spectrum and inverse autocovariances of a stationary time series. J. Time Ser. Anal. 10, 183–202.

Tong, H., Dabas, P., 1990. Cluster of time series models: an example. J. Appl. Statist. 17, 187–198.

Wei, W.W.S., 1990. Time Series Analysis: Univariate and Multivariate Methods. Addison-Wesley, Redwood City, CA.

Xiong, Y., Yeung, D., 2004. Time series clustering with ARMA mixtures. Pattern Recognition 37, 1675–1689.

Zhang, G., Taniguchi, M., 1994. Discriminant analysis for stationary vector time series. J. Time Ser. Anal. 15, 117–126.