An interpolated periodogram-based metric for comparison of time series with unequal lengths

Jorge Caiado and Nuno Crato and Daniel Peña

2006

An interpolated periodogram-based metric for comparison of time series with unequal lengths

Jorge Caiado[1,2], Nuno Crato[2,3] and Daniel Peña[4]

[1]College of Business and Administration, Polytechnic Institute of Setubal, Portugal
[2]Centro de Matemática Aplicada à Previsão e Decisão Económica, CEMAPRE/ISEG
[3]School of Economics and Business, Technical University of Lisbon, Portugal
[4]University Carlos III of Madrid, Spain

KEY WORDS Classification, Cluster analysis, Interpolation, Periodogram, Time series

Abstract

We propose a periodogram-based metric for classification and clustering of time series with different sample sizes. For such cases, we know that the Euclidean distance between the periodogram ordinates cannot be used. One possible way to deal with this problem is to interpolate lineary one of the periodograms in order to estimate ordinates of the same frequencies.

1. Introduction

The problem of comparison of time series has been studied in statistical literature using both time and frequency domain methods. Some related works are by Coates and Diggle (1986), Diggle and Fisher (1991), Diggle and al Wasel (1997), Kakizawa, Shumway and Taniguchi (1998), Maharaj (2002), Caiado, Crato and Peña (2006), among others. However, existing spectral methods for discrimination and clustering analysis of time series cannot be applied directly to series with different sample sizes. Caiado, Crato and Peña (2006) proposed a new measure of distance between time series based on the log normalized periodogram. In particular, they discuss the classification of time series as stationary or as nonstationary. We now extend this method for classifying times series with unequal different lengths. For such cases, we know that the Euclidean distance between the periodogram ordinates cannot be used. One possible way to deal with this problem is to interpolate the periodogram ordinates of the series with longer (shorter) length from the series with the shorter (longer) length.

2. Interpolated periodogram

Let $\{x_t, t = 1, ..., n_x\}$ and $\{y_t, t = 1, ..., n_y\}$ be two stationary processes with different sample sizes $n_x \neq n_y$. The periodogram ordinates of x_t are given by

$$P_x(\omega_j) = (2\pi n_x)^{-1} \left| \sum_{t=1}^{n_x} x_t e^{-it\omega_j} \right|^2 . \quad (1)$$

where $\omega_j = 2\pi j/n_x$, for $j = 1, ..., m_x$, with $m_x = [n_x/2]$, the largest integer less or equal to $n_x/2$, and the frequency ω is in the range $[-\pi, \pi]$. Similar expression is defined for $P_y(\omega_p)$, with $\omega_p = 2\pi p/n_y$, for $p = 1, ..., m_y$, with $m_y = [n_y/2]$. The Euclidean distance between the periodogram ordinates $P_x(\omega_j)$ and $P_y(\omega_p)$ is not adequate for comparison of series x_t and y_t since $m_x \neq m_y$. Without loss of generatility, let $r = [p\frac{m_x}{m_y}]$ be the largest integer less or equal to $p\frac{m_x}{m_y}$ for $p = 1, ..., m_y$, and $m_y < m_x$. We estimate the periodogram ordinates of x_t as

$$\begin{aligned} P'_x(\omega_p) &= P_x(\omega_r) + (P_x(\omega_{r+1}) - P_x(\omega_r)) \\ &\quad \times \frac{\omega_{p,y} - \omega_{r,x}}{\omega_{r+1,x} - \omega_{r,x}} \\ &= P_x(\omega_r) \left(1 - \frac{\omega_{p,y} - \omega_{r,x}}{\omega_{r+1,x} - \omega_{r,x}}\right) \\ &\quad + P_x(\omega_{r+1}) \left(\frac{\omega_{p,y} - \omega_{r,x}}{\omega_{r+1,x} - \omega_{r,x}}\right) \quad (2) \end{aligned}$$

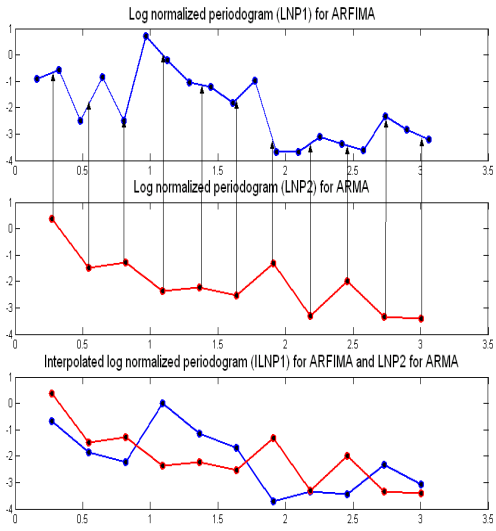

Figure 1: Interpolation of the log normalized periodogram ordinates of an ARFIMA(0,0.45,0) with $n1 = 40$ from an ARMA(1,0), $\phi = 0.95$ with $n2 = 24$

Since now the periodograms $P'_x(\omega_j)$ and $P_y(\omega_p)$ have the same number of frequencies, we can use the following distance between the periodogram ordinates of the two series,

$$d = \sqrt{\frac{1}{m_y}\sum_{p=1}^{m_y}(LNP'_x(\omega_p) - LNP_y(\omega_p))^2}, \tag{3}$$

where $LNP'_x(\omega_p) = \log(P'_x(\omega_p)/Var(x))$ and $LNP_y(\omega_p) = \log(P_y(\omega_p)/Var(y))$ are the logarithms of the normalized periodograms of time series x and y, as recommended by Caiado, Crato and Peña (2006). Figure 1 illustrates the interpolation procedure with two simulated processes.

3. Simulation results

To illustrate the performance of the interpolated periodogram based metric, two series of different sample sizes, $(n_1, n_2) = \{(100, 100), (200, 100), (500, 250), (1000, 500)\}$, were simulated from each of the following processes:

(a) AR(1), $\phi = 0.9$ versus ARIMA(0,1,0);

(b) IMA(1,1), $\theta = 0.8$ versus ARMA(1,1), $\phi = 0.95$, $\theta = 0.74$;

(c) ARFIMA(0,0.45,0) versus AR(1), $\phi = 0.95$.

The four generated series with zero mean and unit variance white noise were grouped into two clusters by hierarchical method of complete linkage using the Euclidean mean distance between the log normalized periodogram ordinates defined in (3). This was repeated 1000 times. The mean percentages of sucess on the comparison in cases (a), (b) and (c) are provided in Tables 1, 2 and 3, respectively. For instance, in Table 1, the value 59.8 means that 59.8% of the times the two AR(1), $\phi = 0.9$, $n_1 = 50$ and $n_2 = 100$ processes were grouped into one cluster and the two AR(1), $\phi = 0.5$, $n_1 = 50$ and $n_2 = 100$ processes were grouped into another cluster.

In the comparisons between ARMA and ARFIMA processes, the interpolated periodogram based metric shows a remarkable good performance. The simulations results on the comparison between ARMA versus ARIMA processes show a performance that increases significantly with the sample size. For unequal lengths, the discrimination between the two models works well.

4. Conclusion

In this paper, we introduced an interpolated periodogram based metric for comparison and clustering of time serie with unqual lengths. This metric is easy to implement and is computationally fast. It can perform very well for comparing between stationary and near stationary processes, and for comparing between short-memory and long-memory processes.

Table 1: Percentages of sucesses AR(1) vs ARIMA(0,1,0)

	ARIMA(0,1,0)			
AR(1): $\phi = 0.9$	(100,100)	(200,100)	(500,250)	(1000,500)
(100,100)	22.7	30.8	78.9	98.3
(200,100)	19.4	36.0	76.6	96.4
(500,250)	59.8	58.2	74.8	92.0
(1000,500)	100.0	96.4	79.4	89.0

Table 2: Percentages of sucesses IMA(1) vs ARMA(1,1)

	ARMA(1,1), $\phi = 0.95$, $\theta = 0.74$			
IMA(1,1): $\theta = 0.8$	(100,100)	(200,100)	(500,250)	(1000,500)
(100,100)	11.1	8.2	60.7	100.0
(200,100)	10.2	20.6	46.2	92.7
(500,250)	60.1	48.7	41.1	54.4
(1000,500)	97.3	90.4	62.6	60.4

Table 3: Percentages of sucesses ARFIMA(0,0.45,0) vs ARMA(1,0)

	ARMA(1,0): $\phi = 0.95$			
ARFIMA(0,0.45,0)	(100,100)	(200,100)	(500,250)	(1000,500)
(100,100)	83.1	86.0	96.0	99.8
(200,100)	82.5	85.2	95.2	99.7
(500,250)	95.1	93.4	93.9	97.6
(1000,500)	100.0	99.9	96.5	97.8

Acknowledgement 1 *This research was supported by a grant from the Fundação para a Ciência e Tecnologia (POCTI/FCT) and by MEC project SEJ2004-03303, Spain.*

REFERENCES

Caiado, J., Crato, N., Peña, D. (2006). "A periodogram-based metric for time series classification", *Comput. Stat. Data Anal.* 50, 2668-2684.

Coates, D. S., Diggle, P. J. (1986). "Tests for comparing two estimated spectral densities", *J. Time Ser. Anal.* 7, 7-20.

Diggle, P. J., Fisher, N. I. (1991). "Nonparametric comparison of cumulative periodograms", *Appl. Stat.* 40, 423-434

Diggle, P. J., al Wasel, I. (1997). "Spectral analysis of replicated biomedical time series", *Appl. Stat.* 46, 31-71.

Kakizawa, Y., Shumway, R. H., Taniguchi, M. (1998). "Discrimination and clustering for multivariate time series", *J. Amer. Stat. Assoc.* 93, 328-340.

Maharaj, E. A. (2002). "Comparison of non-stationary time series in the frequency domain", *Comput. Stat. Data Anal.* 40, 131-141.