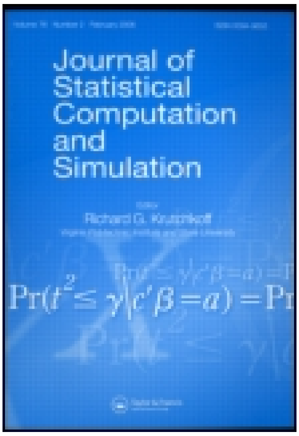

Journal of Statistical Computation and Simulation
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Tests for comparing time series of unequal lengths

Jorge Caiado [a], Nuno Crato [a] & Daniel Peña [b]
[a] CEMAPRE, ISEG, Technical University of Lisbon, Lisbon, Portugal
[b] Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain
Version of record first published: 06 Jul 2011.

Journal of Statistical Computation and Simulation
Vol. 82, No. 12, December 2012, 1715–1725

Tests for comparing time series of unequal lengths

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

[a]*CEMAPRE, ISEG, Technical University of Lisbon, Lisbon, Portugal;* [b]*Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain*

(*Received 25 January 2011; final version received 27 May 2011*)

This paper deals with hypothesis testing for independent time series with unequal length. It proposes a spectral test based on the distance between the periodogram ordinates and a parametric test based on the distance between the parameter estimates of fitted autoregressive moving average models. Both tests are compared with a likelihood ratio test based on the pooled spectra. In all cases, the null hypothesis is that the two series under consideration are generated by the same stochastic process. The performance of the three tests is investigated by a Monte Carlo simulation study.

Keywords: hypothesis testing; ARMA models; spectral analysis

1. Introduction

The classification and clustering of time series have been studied in literature using both time and frequency domain methods. The comparison of time series using spectral analysis approaches was considered by Coates and Diggle [1], Diggle and Fisher [2], Dargahi-Noubary [3], Diggle and al Wasel [4], Kakizawa et al. [5], Maharaj [6], Caiado et al. [7,8], among others. In the time domain, some relevant works are by Maharaj [9,10], Piccolo [11] and Corduas and Piccolo [12]. For a review of the basics of time series clustering and its applications, see the survey by Liao [13].

Caiado et al. [8] proposed a spectral domain method for clustering time series of unequal length. Their approach consists in calculating the periodogram ordinates of time series of any length at a common Fourier frequency and then using a discrepancy statistic to investigate similarities between the time series. We now extend this procedure for hypothesis testing purposes. In particular, we introduce a periodogram-based test of the hypothesis that two independent time series are realizations of the same stationary process. This nonparametric test is compared with a parametric approach based on the distance between the parameter estimates of autoregressive moving average (ARMA) models and a likelihood ratio test based on the pooled spectra. We assess the power and size of the three test statistics for different sample sizes by a Monte Carlo simulation study.

*Corresponding author. Email: jcaiado@iseg.utl.pt

ISSN 0094-9655 print/ISSN 1563-5163 online

http://dx.doi.org/10.1080/00949655.2011.592985
http://www.tandfonline.com

The remainder of the paper is organized as follows. In Section 2, we present a periodogram-distance-based test of the hypothesis that two time series of unequal lengths are realizations of the same generating process. In Section 3, we develop a parametric test in the time domain also to determine whether two time series are generated by the same stochastic mechanism. In Section 4, we present a simulation study to compare the performance of the proposed tests with a likelihood ratio approach. Section 5 summarizes the paper.

2. A periodogram-distance-based test

2.1. *Interpolated periodogram*

Let $\{x_t, t = 1, \ldots, n_x\}$ and $\{y_t, t = 1, \ldots, n_y\}$ be two stationary processes with different sample sizes. Without loss of generality assume that $n_x > n_y$. The periodogram of x_t is given by

$$P_x(\omega_j) = \frac{1}{n_x} \left| \sum_{t=1}^{n_x} x_t \mathrm{e}^{-\mathrm{i}t\omega_j} \right|^2, \tag{1}$$

where $\omega_j = 2\pi j/n_x$, for $j = 1, \ldots, m_x$, with $m_x = [n_x/2]$, the largest integer less or equal to $n_x/2$. Similar expression is defined for $P_y(\omega_p)$, with $\omega_p = 2\pi p/n_y$, for $p = 1, \ldots, m_y$, with $m_y = [n_y/2]$. When $m_x \neq m_y$, the Fourier frequencies at which the periodogram ordinates are computed do not coincide. Therefore, the Euclidean distance between the periodogram ordinates $P_x(\omega_j)$ and $P_y(\omega_p)$ is not adequate for comparing time series x_t and y_t. To deal with the problem of unequal length, Caiado et al. [8] suggest to construct an interpolated periodogram for the longer series at the frequencies defined by the shorter series. Without loss of generality, let $r = [p(m_x/m_y]$ be the largest integer less or equal to $p(m_x/m_y)$ for $p = 1, \ldots, m_y$ and $m_y < m_x$. The periodogram ordinates of x_t can be estimated as

$$\begin{aligned} P_x^I(\omega_p) &= P_x(\omega_r) + (P_x(\omega_{r+1}) - P_x(\omega_r)) \times \frac{\omega_{p,y} - \omega_{r,x}}{\omega_{r+1,x} - \omega_{r,x}} \\ &= P_x(\omega_r)\left(1 - \frac{\omega_{p,y} - \omega_{r,x}}{\omega_{r+1,x} - \omega_{r,x}}\right) + P_x(\omega_{r+1})\left(\frac{\omega_{p,y} - \omega_{r,x}}{\omega_{r+1,x} - \omega_{r,x}}\right). \end{aligned} \tag{2}$$

This procedure will yield an interpolated periodogram with the same Fourier frequencies of the shorter periodogram $P_y(\omega_p)$. If we are only interested in the dependence structure and not in the process scale, we can standardize the periodograms, dividing them by the sample variances, and then taking logarithms: $\log \mathrm{NP}_x^I(\omega_p) = \log(P_x^I(\omega_p)/\hat{\sigma}_x)$ and $\log \mathrm{NP}_y(\omega_p) = \log(P_y(\omega_p)/\hat{\sigma}_y)$. Figure 1 illustrates the interpolation procedure with two simulated processes. This method seems to work well for comparison purposes. An illustrative application of the spectral-discrepancy statistic to clustering time series is presented by Caiado et al. [8].

2.2. *Hypothesis testing*

We now suggest a test of hypotheses to determine whether two independent time series are realizations of stochastic processes. Given two independent stationary series x_t and y_t, with $n_x = n_y$, the null hypothesis to be tested is $H_0 : f_x(\omega_j) = f_y(\omega_j)$, that is, there is no difference between the underlying spectra of the series $\{x_t\}$ and $\{y_t\}$ at all Fourier frequencies ω_j.

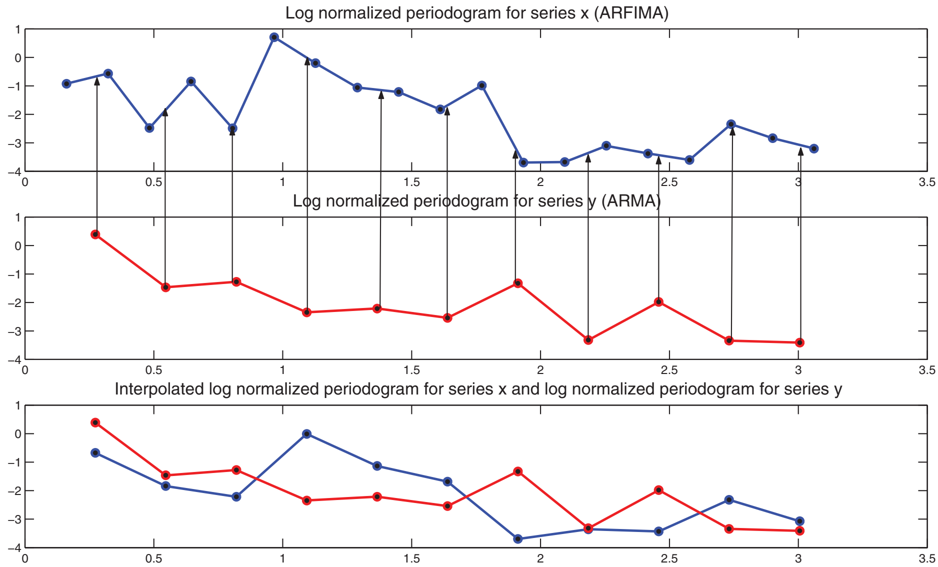

Figure 1. Interpolation of the log-normalized periodogram ordinates of an ARFIMA(0,0.45,0) with $n_x = 40$ from an ARMA(1,0), $\phi = 0.95$ with $n_y = 24$.

Since, asymptotically, $P_x(\omega_j) \sim f_x(\omega_j)\chi^2_{(2)}/2$, where $f_x(\omega)$ is the spectral density function, it follows that

$$E\left[\frac{P_x(\omega_j)}{\hat{\gamma}_{0,x}}\right] = \frac{f_x(\omega_j)}{\sigma_x^2} \tag{3}$$

and

$$\mathrm{Var}\left[\frac{P_x(\omega_j)}{\hat{\gamma}_{0,x}}\right] = \frac{f_x^2(\omega_j)}{\sigma_x^4}. \tag{4}$$

Similar expressions apply to the periodogram of y_t. A good approximation to $E[\log(\hat{f}_x(\omega_j)/\hat{\gamma}_{0,x})]$ and $\mathrm{Var}[\log(\hat{f}_x(\omega_j)/\hat{\gamma}_{0,x})]$ is the sample mean and sample variance of the log-normalized periodogram, that is,

$$E\left[\log\left(\frac{\hat{f}_x(\omega_j)}{\hat{\gamma}_{0,x}}\right)\right] \approx \frac{1}{m_x}\sum_{p=1}^{m_x} \log \mathrm{NP}_x(\omega_j) = \bar{x}_{\mathrm{LNP}} \tag{5}$$

and

$$\mathrm{Var}\left[\log\left(\frac{\hat{f}_x(\omega_j)}{\hat{\gamma}_{0,x}}\right)\right] \approx \frac{1}{m_x}\sum_{p=1}^{m_x} [\log \mathrm{NP}_x(\omega_j) - \bar{x}^2_{\mathrm{LNP}}]^2 = s^2_{\mathrm{LNP},x}. \tag{6}$$

Similar expressions are given for $\bar{y}_{\mathrm{LNP}}$ and $s^2_{\mathrm{LNP},y}$.

Under some suitable conditions, the logarithmic transformation of the sample spectrum is closer to the normal distribution than to a chi-square distribution [14]. The following statistic provides a test of significance for comparing the log-normalized periodograms of the two series,

$$D_{\mathrm{NP}} = \frac{\bar{x}_{\mathrm{LNP}} - \bar{y}_{\mathrm{LNP}}}{\sqrt{(s^2_{\mathrm{LNP},x} + s^2_{\mathrm{LNP},y})/m}}. \tag{7}$$

This statistic is approximately normally distributed with zero mean and unit variance. For different lengths, $m_x \neq m_y$, regular periodograms cannot be used as the Fourier frequencies are different.

To deal with this problem, we calculate the interpolated periodogram for the longer series x_t at the frequencies $\omega_p = 2\pi p/n_y$, for $p = 1, \dots, m_y$, as defined in Equation (2), and then use the test statistic

$$D^I_{\mathrm{NP}} = \frac{\bar{x}_{\mathrm{LNP}^I} - \bar{y}_{\mathrm{LNP}}}{\sqrt{(s^2_{\mathrm{LNP}^I,x} + s^2_{\mathrm{LNP},y})/m_y}}, \tag{8}$$

where $\bar{x}_{\mathrm{LNP}^I} = (1/m_y)\sum_{p=1}^{m_y} \log \mathrm{NP}^I_x(\omega_p)$, $s^2_{\mathrm{LNP}^I,x} = (1/m_y)\sum_{p=1}^{m_y}[\log \mathrm{NP}^I_x(\omega_p) - \bar{x}^2_{\mathrm{LNP}^I}]^2$ and $m_y < m_x$.

For reference, we consider the problem of testing equality of spectral densities using a likelihood ratio approach (see, e.g. [15]). For equal lengths, $m_x = m_y = m$, we compute the pooled normalized periodogram $\mathrm{NP}(\omega_j) = \frac{1}{2}[\mathrm{NP}_x(\omega_j) + \mathrm{NP}_y(\omega_j)]$ under the null hypothesis $H_0 : f_x(\omega_j) = f_y(\omega_j)$ and use the likelihood ratio test

$$\mathrm{LRT} = -\sum_{j=1}^{m} \log \mathrm{NP}_x(\omega_j) - \sum_{j=1}^{m} \log \mathrm{NP}_y(\omega_j) + 2\sum_{j=1}^{m} \log \mathrm{NP}(\omega_j), \tag{9}$$

which is distributed proportionally to a chi-square random variable. This test can be easily extended to the time series of unequal length using the reduced periodogram approach [8]. In practice, this consists in calculating both periodograms at a common Fourier frequency. In our simulation study, we suggest computing the periodogram of the longer series at the frequencies of the shorter series.

3. A parametric approach

The problem of comparison of series of unequal length can also be analysed by a parametric approach. Let $z_t, t = 1, 2, \dots, n$ be a time series, represented by a stationary and invertible ARMA or ARMA(p,q) process,

$$\phi_p(B)z_t = \theta_q(B)\varepsilon_t, \tag{10}$$

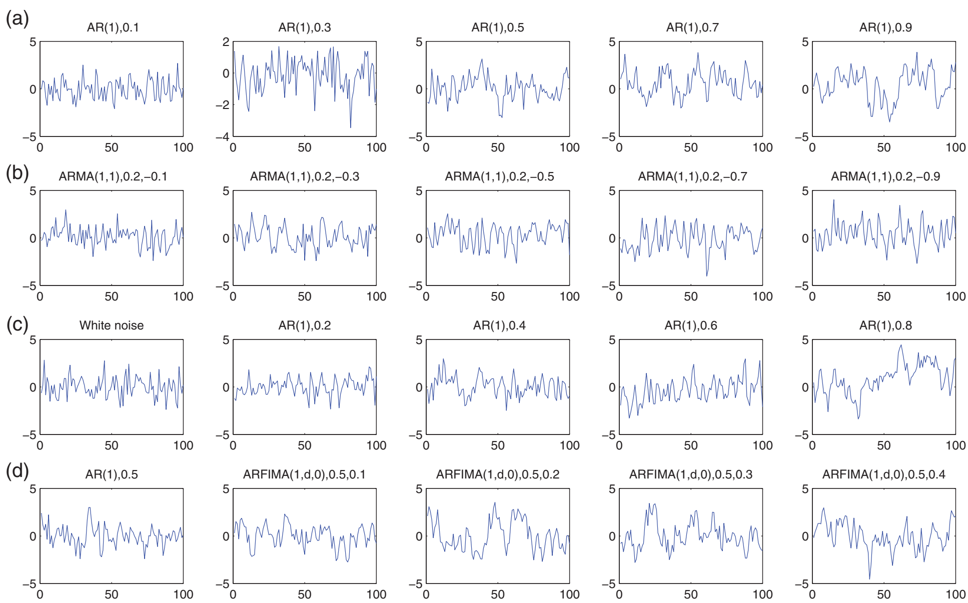

Figure 2. Times series generated from processes (a) AR(1), $\phi = 0.1$, 0.3, 0.5, 0.7, 0.9, (b) ARMA(1,1), $\phi = 0.2$, $\theta = -0.1, -0.3, -0.5, -0.7, -0.9$, (c) AR(1), $\phi = 0, 0.2, 0.4, 0.6, 0.8$ and (d) ARFIMA(1,d,0), $d = 0, 0.1, 0.2, 0.3, 0.4$.

where $\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ are polynomials of orders p and q, B is the back-shift operator $B^k z_t = z_{t-k}$, and ε_t is a Gaussian white noise process with zero mean and constant variance σ_ε^2.

Suppose we have two independent time series x_t and y_t generated by the same ARMA(p,q) process, but with different parameter values. Let the $k = p + q$ estimated parameters be grouped in the vectors $\hat{\beta}_x$ and $\hat{\beta}_y$ with estimated covariance matrices V_x and V_y, respectively. We want to check whether they are different realizations of the same stochastic process, so that $E[\hat{\beta}_x] = E[\hat{\beta}_y] = \beta$. Then, $\delta = \hat{\beta}_x - \hat{\beta}_y$ for large samples will be an approximately normally distributed vector with zero mean and covariance matrix

$$V_\delta = V_x + V_y, \tag{11}$$

and therefore, we can use the statistic

$$D_P = \delta' V_\delta^{-1} \delta, \tag{12}$$

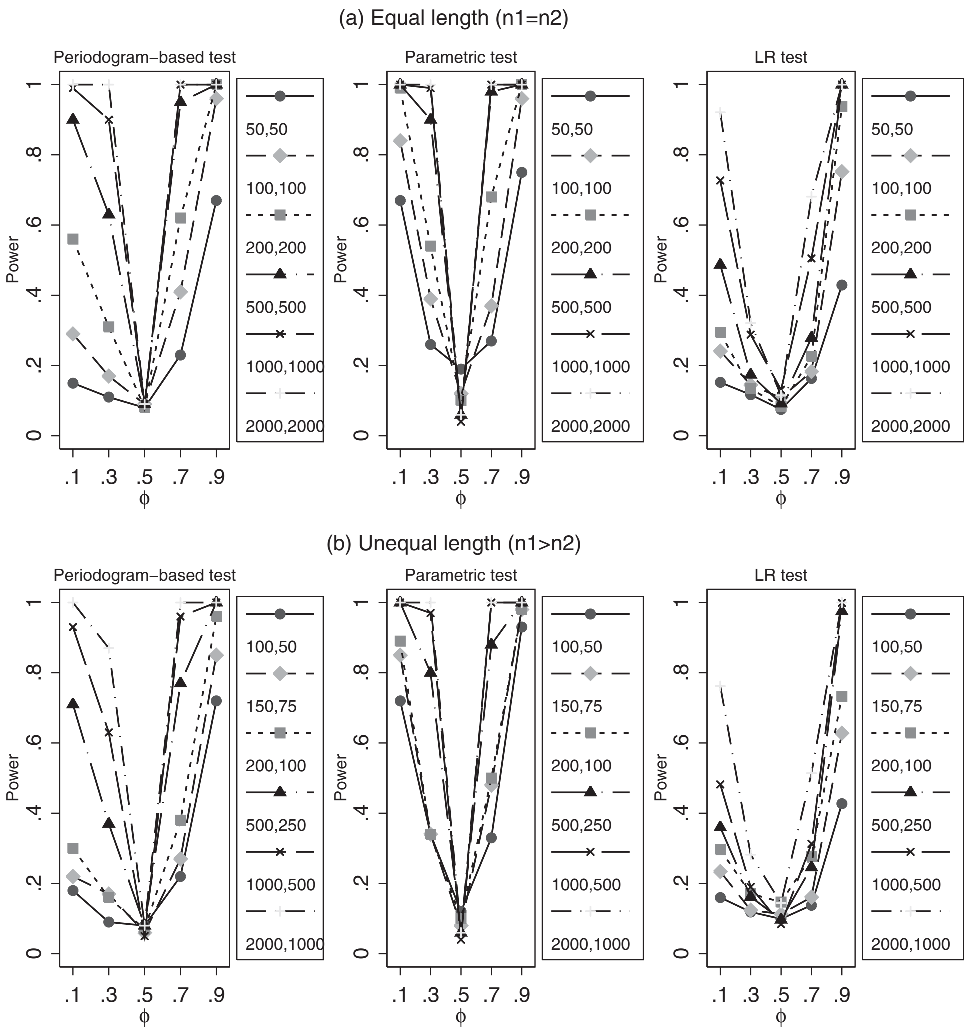

Figure 3. Estimates of power and size of 10% tests of significance for AR(1), $\phi = 0.5$ versus AR(1), $\phi = 0.1, 0.3, 0.5, 0.7, 0.9$.

which is asymptotically a chi-square distribution with k degrees of freedom under the null $\beta_x = \beta_y$. Hamilton [16, Section 14.3] suggested a similar statistic to test for structural stability of ARMA models over different subperiods.

In order to test if two generating ARMA processes are equal, the model for each time series can be selected by Akaike's information criterion (AIC) or Bayesian's information criterion (BIC). If the obtained model is the same for the two time series, then the statistic D_P is computed by using the estimated parameters in each time series. If the selected models are different, then the problem is more involved. Two methods have been proposed:

(i) We can start fitting a large ARMA model to both processes which encompass the two models to be compared, for instance, the larger of two AIC or BIC selected models. This method has two main problems: (1) the estimated parameters will be highly correlated for the overparametrized estimated model (or models) and the corresponding covariance matrix (or matrices) may be close to singular; (2) we have to be very careful to avoid possible near cancellation of the autoregressive (AR) and moving average (MA) roots on both sides.

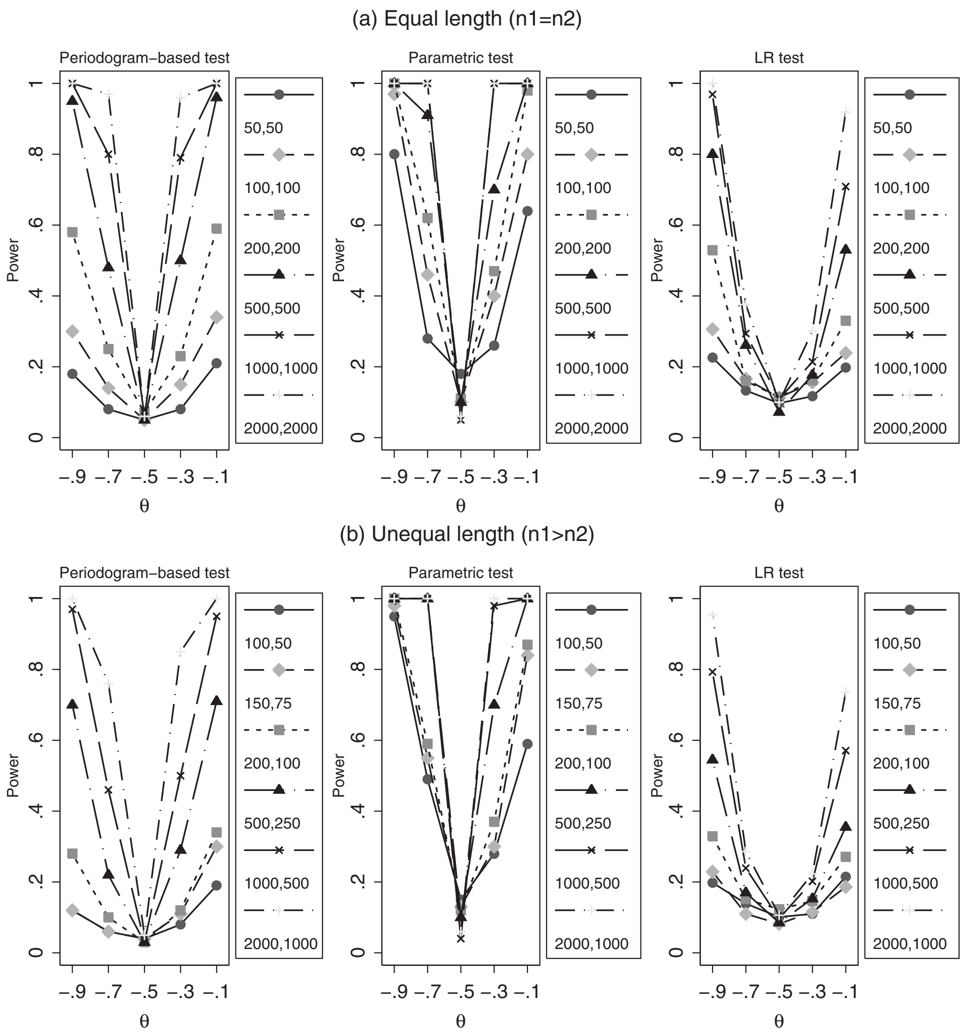

Figure 4. Estimates of power and size of 10% tests of significance for ARMA(1,1), $\phi = 0.2$ and $\theta = -0.5$ versus ARMA(1,1), $\phi = 0.2$ and $\theta = -0.1, -0.3, -0.5, -0.7, -0.9$.

(ii) Alternatively, in order to avoid the serious problem of near cancellation of roots, we can use AR approximations and thus fit to both processes the larger selected AR model and then compare the estimated parameters [9]. This method has two drawbacks. First, we may get very much correlated estimated parameters, specially when we have MA generating processes. Second, we may need very large AR models.
Given these problems, we propose an alternative approach to apply the parametric test when the selected models are different.

(iii) We fit both selected models, say M1 and M2, to both time series and compute the statistic D_P in these two situations, that is, $D_P(\text{M1})$ and $D_P(\text{M2})$. If $D_P(\text{M1}) \leq \chi^2_{(k_1)}$ and $D_P(\text{M2}) \leq \chi^2_{(k_2)}$, where k_1 and k_2 are the degrees of freedom associated with the models M1 and M2, the null hypothesis is not rejected and we conclude that the processes are generated by the same model. If the null hypothesis is rejected in one of the models, or in both, then we conclude that the generating processes are different. Since we have two comparison statements to be made, the Bonferroni inequality suggests each test with a significance level $\alpha/2$ to ensure that the overall significance level is at most α. In our simulation study, we will use this alternative approach (iii).

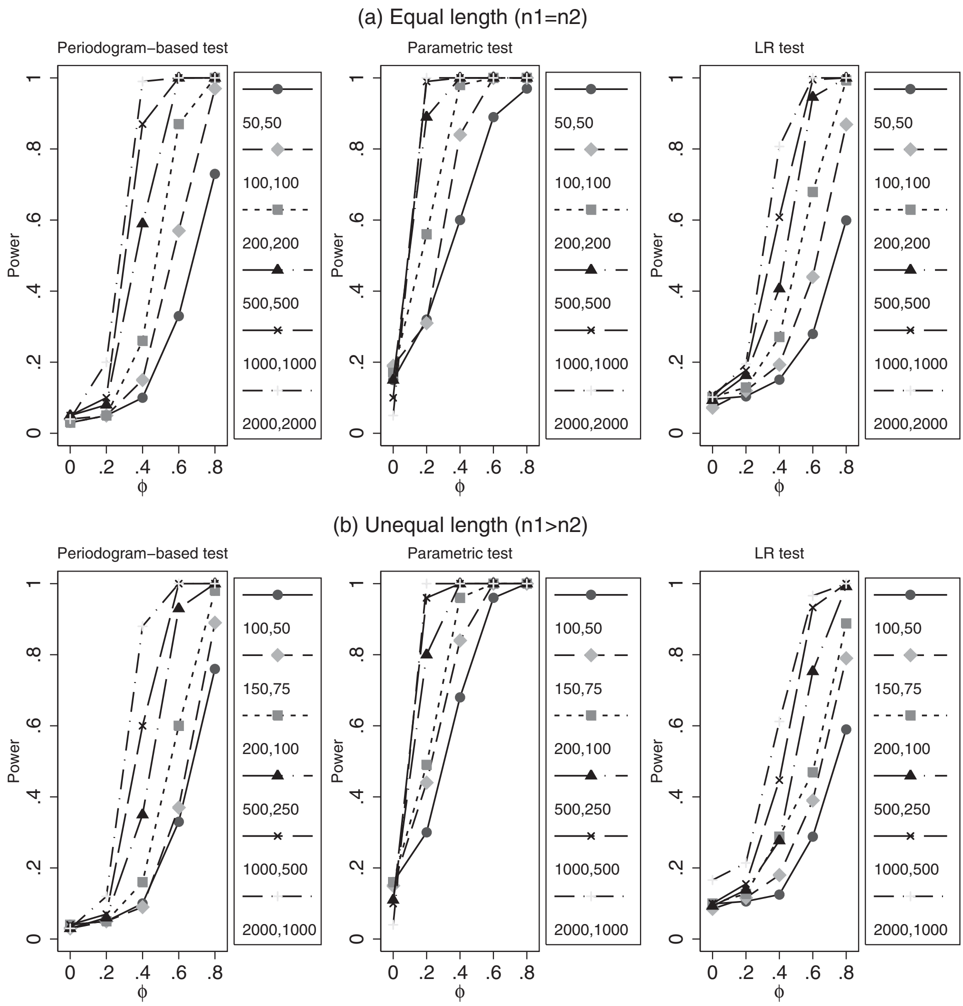

Figure 5. Estimates of power and size of 10% tests of significance for white noise versus AR(1), $\phi = 0, 0.2, 0.4, 0.6, 0.8$.

4. Monte Carlo simulations

We obtained the estimates of the power and size of the proposed tests for simulated series from the following processes:

(a) AR(1), $\phi = 0.5$ versus AR(1), $\phi = 0.1, 0.3, 0.5, 0.7, 0.9$;
(b) ARMA(1,1), $\phi = 0.2$, $\theta = -0.5$ versus ARMA(1,1), $\phi = 0.2$, $\theta = -0.1, -0.3, -0.5, -0.7, -0.9$;
(c) White noise versus AR(1), $\phi = 0, 0.2, 0.4, 0.6, 0.8$;
(d) ARMA(1,0), $\phi = 0.5$ versus ARFIMA(1,d,0), $\phi = 0.5$, $d = 0, 0.1, 0.2, 0.3, 0.4$;
(e) AR(1), $\phi = 0.7$ versus AR(1), $\phi = 0.7, 0.8, 0.9, 1.0$.

In case (a), we compare low-order models of similar type and similar autocorrelation functions. In case (b), we compare selected ARMA processes in order to deal with peak spectra. In case (c), we compare a white noise process with a low-order AR process. In case (d), we compare

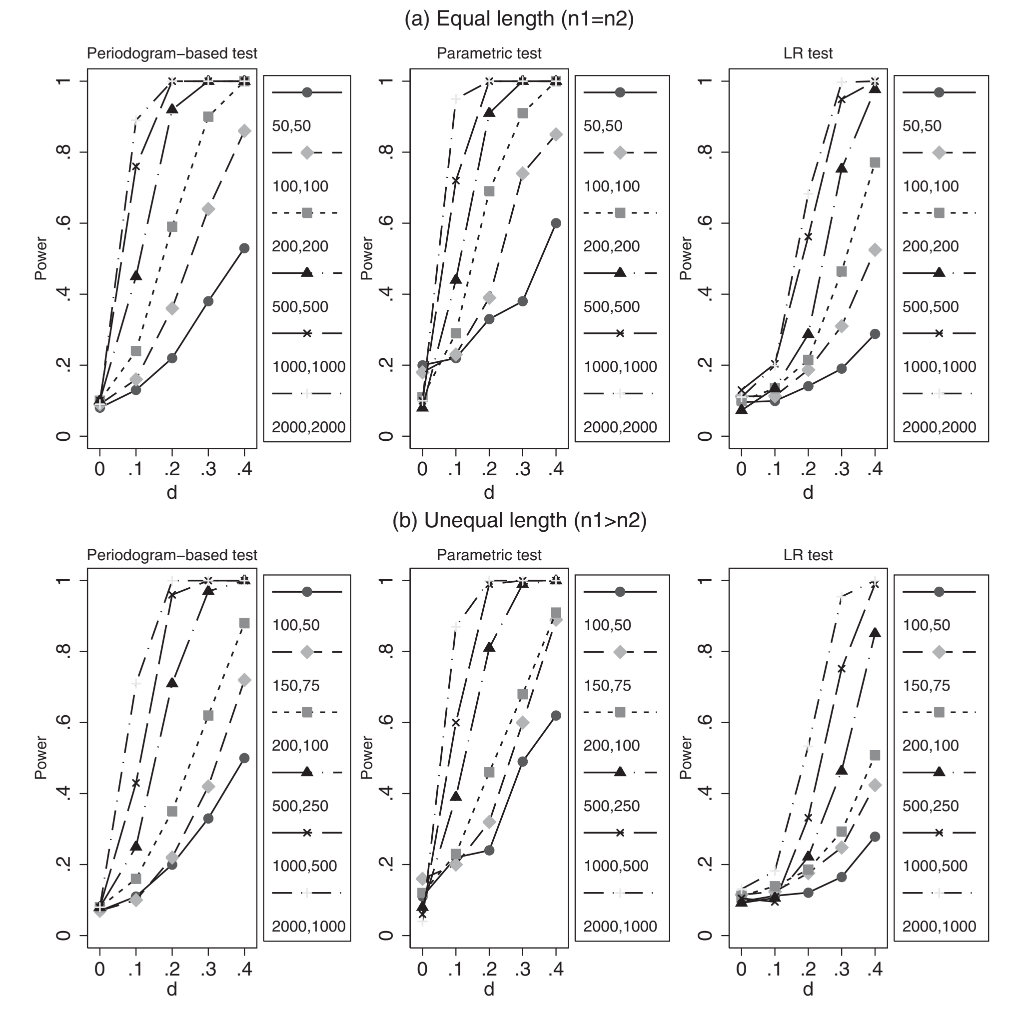

Figure 6. Estimates of power and size of 10% tests of significance for ARMA(1,0), $\phi = 0.5$ versus ARFIMA(1,d,0), $\phi = 0.5$ and $d = 0, 0.1, 0.2, 0.3, 0.4$.

stationary long-memory and short-memory processes. In case (e), we compare stationary and near-nonstationary AR processes. The ARMA models were generated with zero mean and unit variance white noise. The fractional noise in autoregressive fractionally integrated moving average (ARFIMA) model was simulated using the finite Fourier method of Davies and Harte [17]. Figure 2 shows the plots of time series from the generated processes (a)–(d).

From these comparisons, we are able to see how the tests work for distinguishing similar models with different parameters. From the considerable set of values for the parameters, we can verify whether an increasing difference leads to better test power. For each of the considered processes, we simulated pairs of series of equal lengths $(n_1, n_2) = \{(50, 50), (100, 100), (200, 200), (500, 500), (1000, 1000), (2000, 2000)\}$ and unequal lengths $(n_1, n_2) = \{(100, 50), (150, 75), (200, 100), (500, 250), (1000, 500), (2000, 1000)\}$. The results were based on 1000 replications of each pair of processes. For the parametric approach, we fitted ARMA(p,q) models to the series, with the orders $p = 0, 1, 2, 3$ and $q = 0, 1, 2, 3$ selected by BIC (the AIC does not work in selecting models for hypothesis testing, as noted by Peña and Rodriguez [18]).

Figures 3–6 provide estimates of the size and power for cases (a)–(d) (stationary versus stationary) for 10% level of significance using the three tests discussed above. Figure 7 gives the

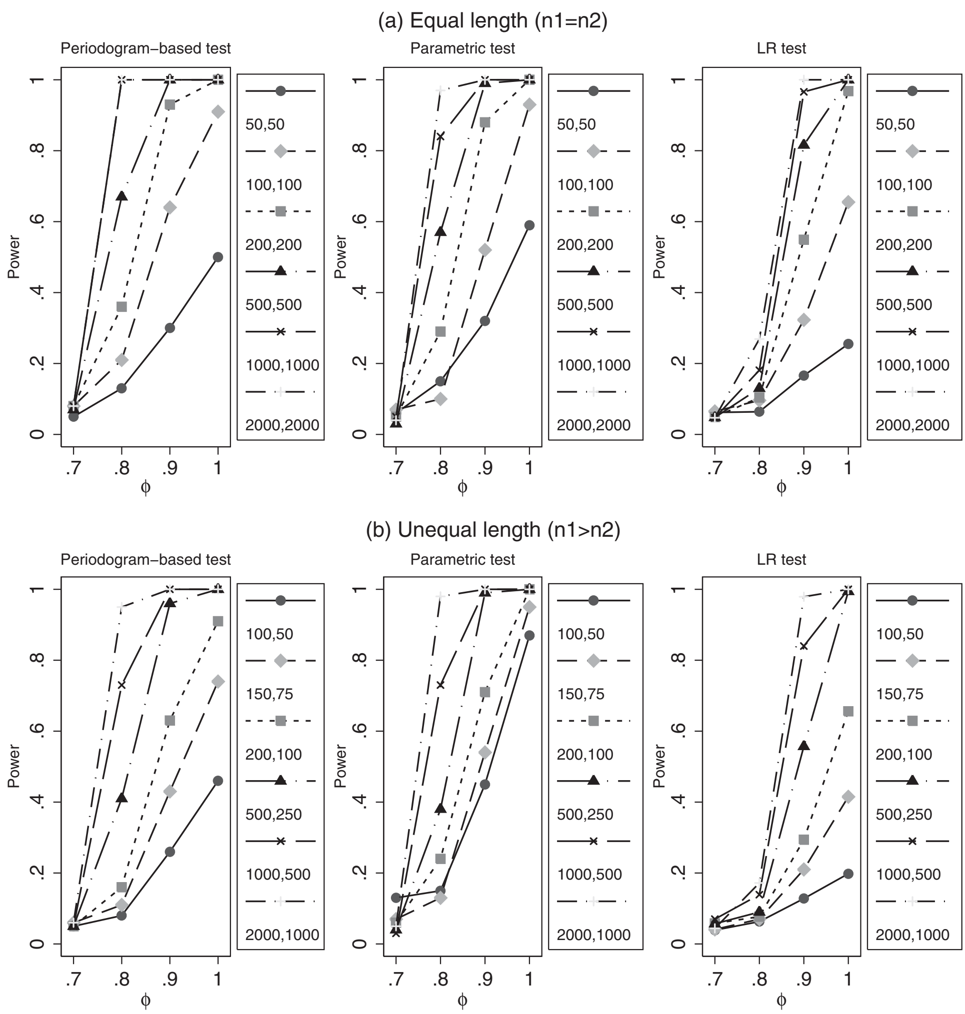

Figure 7. Estimates of power and size of 5% tests of significance for AR(1), $\phi = 0.7$ versus AR(1), $\phi = 0.7, 0.8, 0.9, 1.0$.

power and size estimates in the case (e) (stationary versus near nonstationary) for 5% level of significance.

As expected, the tests for large samples are more powerful than the tests for small samples.

When the series are generated from the pairs of processes: (a) AR(1), $\phi = 0.5$ versus AR(1), $\phi \neq 0.5$; (b) ARMA(1,1), $\phi = 0.2$, $\theta = -0.5$ versus ARMA(1,1), $\phi = 0.2$, $\theta \neq -0.5$; and (c) white noise versus AR(1), $\phi > 0$, the power of the parametric test for small samples is larger than the periodogram-based test.

When the series are generated from the pairs of processes: (d) AR(1), $\phi = 0.5$ versus ARFIMA(1,d,0), $\phi = 0.5, d > 0$; and (e) AR(1), $\phi = 0.7$ versus AR(1), $\phi > 0.7$, the performance of the two tests seems to be similar for all the sample sizes considered.

For the parametric test, overall estimates of the size for small samples exceed slightly the significance levels when the two series were simulated from the same process, whereas for the periodogram-based test the estimated sizes were very close to the significance levels and do not change significantly with the increasing sample.

A limitation of the parametric test is the need of previously fitting ARMA models of several time series in order to compute the distances. In contrast, the periodogram-based test is relatively simple and computationally efficient and does not require the choice of particular models.

From the simulation study, we also see that, in general, the performance of the periodogram-distance-based test is better than the likelihood ratio test. For all the cases, the contrast between the tests for equal and unequal lengths is similar.

5. Concluding remarks

This paper focuses on the development of a periodogram-based test for comparison of time series with unequal length. The proposed test uses the interpolated periodogram method of Caiado et al. [8] to compute the individual periodogram ordinates at different Fourier frequencies. It can perform very well for comparing stationary processes with similar sample properties, for comparing stationary and near-nonstationary processes, and for comparing short-memory and long-memory processes. The estimated power and sizes of the tests for both equal and unequal lengths give similar results, which shows the robustness of the proposed approach.

We have also developed a time domain parametric approach based on the distance between parameter estimates of the same model. We found that the parametric test had very high power to distinguish between two distinct processes. However, contrarily to the periodogram-based test, which is easy to implement and computational fast, the parametric approach needs *ad hoc* ARMA modelling of several time series.

For reference, the proposed tests are compared with a likelihood ratio test based on the pooled spectra. The simulation study shows that the periodogram-distance-based tests are generally more powerful than the likelihood ratio approach for comparison of time series generated from different processes.

References

[1] D.S. Coates and P.J. Diggle, *Tests for comparing two estimated spectral densities*, J. Time Ser. Anal. 7 (1986), pp. 7–20.
[2] P.J. Diggle and N.I. Fisher, *Nonparametric comparison of cumulative periodograms*, Appl. Stat. 40 (1991), pp. 423–434.
[3] G.R. Dargahi-Noubary, *Discrimination between Gaussian time series based on their spectral differences*, Comm. Statist. Theory Methods 21 (1992), pp. 2439–2458.
[4] P.J. Diggle and I. al Wasel, *Spectral analysis of replicated biomedical time series*, Appl. Stat. 46 (1997), pp. 31–71.

[5] Y. Kakizawa, R.H. Shumway, and M. Taniguchi, *Discrimination and clustering for multivariate time series*, J. Amer. Statist. Assoc. 93 (1998), pp. 328–340.
[6] E.A. Maharaj, *Comparison of non-stationary time series in the frequency domain*, Comput. Statist. Data Anal. 40 (2002), pp. 131–141.
[7] J. Caiado, N. Crato, and D. Peña, *A periodogram-based metric for time series classification*, Comput. Statist. Data Anal. 50 (2006), pp. 2668–2684.
[8] J. Caiado, N. Crato, and D. Peña, *Comparison of time series with unequal length in the frequency domain*, Comm. Statist. Simulation Comput. 38 (2009), pp. 527–540.
[9] E.A. Maharaj, *A significance test for classifying ARMA models*, J. Statist. Comput. Simul. 54 (1996), pp. 305–331.
[10] E.A. Maharaj, *Clusters of time series*, J. Classification 17 (2000), pp. 297–314.
[11] D. Piccolo, *A distance measure for classifying ARIMA models*, J. Time Ser. Anal. 11 (1990), pp. 152–164.
[12] M. Corduas and D. Piccolo, *Time series clustering and classification by the autoregressive metric*, Comput. Statist. Data Anal. 52 (2008), pp. 1860–1872.
[13] T.W. Liao, *Clustering of time series data – a survey*, Pattern Recognit. 38 (2005), pp. 1857–1874.
[14] G.M. Jenkins and M.B. Priestley, *The spectral analysis of time series*, J. R. Stat. Soc. Ser. B 19 (1957), pp. 1–12.
[15] P.J. Diggle, *Time Series*, Oxford University Press, New York, 1990.
[16] J.D. Hamilton, *Time Series Analysis*, Princeton University Press, Princeton, NJ, 1994.
[17] R.B. Davies and D.S. Harte, *Tests for Hurst effect*, Biometrika 74 (1987), pp. 95–102.
[18] D. Peña and J. Rodriguez, *Detecting nonlinearity in time series by model selection criteria*, Int. J. Forecast. 21 (2005), pp. 731–748.