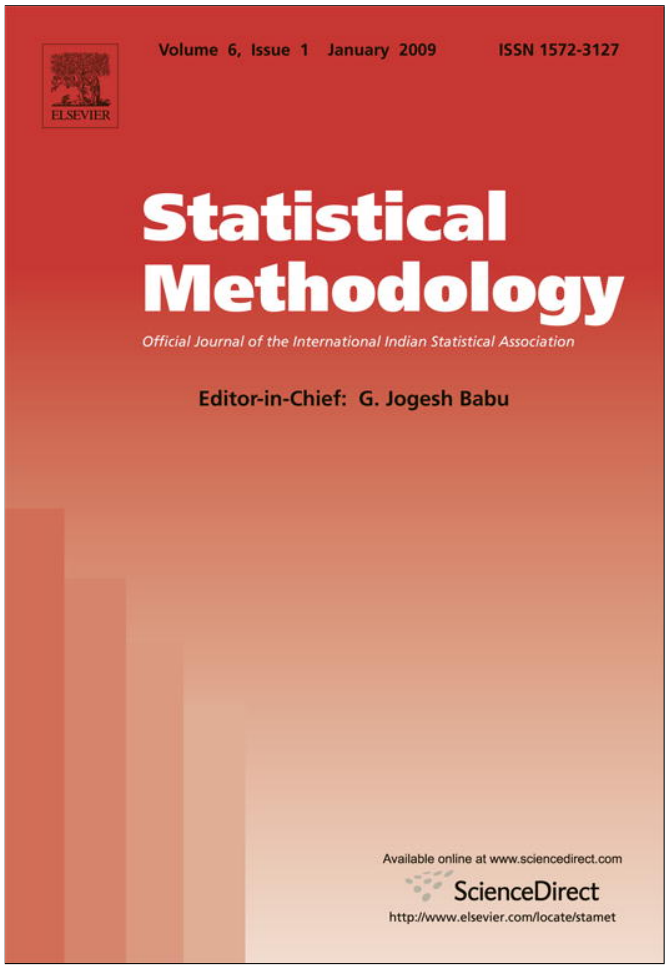

Available online at www.sciencedirect.com
ScienceDirect
Statistical Methodology 6 (2009) 1–20
Statistical Methodology
Official Journal of the International Indian Statistical Association
www.elsevier.com/locate/stamet

Spectral domain diagnostics for testing model proximity and disparity in time series data

Tucker McElroy[a,*], Scott Holan[b]

[a] *Statistical Research Division, US Census Bureau, 4700 Silver Hill Road, Washington, DC 20233-9100, United States*
[b] *Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO, 65211-6100, United States*

Received 1 November 2007; received in revised form 5 February 2008; accepted 13 February 2008

Abstract

Diagnostics for testing the proximity and disparity of a model spectral density to the truth for time series data are formulated by considering a convenient measure of a spectral density's departure from constancy. The method is illustrated through numerical experiments and several case studies.
Published by Elsevier B.V.

Keywords: ARMA; EXP models; Frequency domain; Nonstationary time series

1. Introduction

This paper presents diagnostics for testing whether a given model spectral density is close to the truth, by considering a convenient measure of a spectral density's departure from constancy. In the model-based approach to time series analysis, estimated residuals are computed once a fitted model has been obtained from the data, and these are then tested for "whiteness" i.e., it is determined whether they behave like white noise ([3], pp. 306–314). Tests for residual whiteness include Portmanteau tests, such as [10,9,13], and frequency domain tests — [1,12,4,7]; these model goodness-of-fit (gof) procedures generally postulate whiteness of the model errors as the Null Hypothesis, so that significant rejections indicate model *inadequacy*. Since the classical statistical paradigm dictates that the practitioner seeks to reject Null Hypotheses by obtaining low p-values, we have the paradoxical situation that gof tests are actually designed to identify

* Corresponding author. Tel.: +1 301 763 3227; fax: +1 301 763 8399.
E-mail addresses: tucker.s.mcelroy@census.gov (T. McElroy), holans@missouri.edu (S. Holan).

1572-3127/$ - see front matter. Published by Elsevier B.V.
doi:10.1016/j.stamet.2008.02.007

bad models. Nevertheless, what is needed is an indication that the fitted model is good, or at least adequate. In order to obtain a statistically significant indication of model *adequacy*, one must reject a Null Hypothesis of nonwhiteness; in other words, a badness-of-fit (bof) diagnostic is needed in order to find good models.

The above-mentioned gof procedures indicate whether or not a model is correct, i.e., whether the model class has been correctly specified. We consider the model class together with its parameter values, which can be summarized (for Gaussian processes) through the spectral density. Our tests assess the proximity and disparity between the model spectrum and the true spectrum. The proximity/disparity paradigm is similar to that of gof/bof: if our Null Hypothesis dictates that model and true spectrum are proximate, then significant p-values indicate problems with the model — either the parameter values or the model class (or both). Therefore, in order to obtain indications of adequacy of the model spectrum, we should formulate a Null Hypothesis of disparity between model and true spectrum.

The ratio of the model to the true spectrum will be a constant function when there is complete proximity, and it will be useful to map this ratio into a single summary number, or measure. To that end, we present a convenient characterization of whiteness of a time series, which is quite similar to the approach of Drouiche [7] and was developed independently. We demonstrate that our measure intuitively captures the concept of proximity/disparity for time series, and can be used to test for the inadequacy/adequacy of models and their parameter values in a flexible and rigorous fashion. The work of Drouiche [7], while similar in spirit, does not consider the bof applications. Section 2 defines our spectral measure – our characterization of whiteness – and motivates this choice by making connections to the work of Peña and Rodríguez [14,15]. We also derive some basic properties of our measure, which include those of [7], but with an additional facet that is arguably advantageous in the proximity/disparity context. Some examples are provided on familiar time series models in Section 3. Section 4 considers the statistical estimate of the spectral measure and its distributional properties. The asymptotics involve known techniques, but the computation of asymptotic variances is delicate and is included in the Appendix, along with all proofs. Section 5 provides an explicit discussion of proximity/disparity testing using our diagnostics, and Section 6 demonstrates their empirical properties through some simulation experiments and case studies. We compare the proximity test to the Ljung–Box [10] diagnostic and also provide some power results for the disparity procedure.

2. Measuring whiteness

We make use of some basic notations in this paper. Suppose that, after suitable transformations and differencing if necessary, we have a mean zero stationary time series $X_1, X_2, \ldots, X_n$, which will sometimes be denoted by the vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$. When the autocovariance function $\gamma_f(h)$ is absolutely summable, the spectral density f can be defined by

$$f(\lambda) = \sum_{h=-\infty}^{\infty} \gamma_f(h) \mathrm{e}^{-\mathrm{i}h\lambda} \tag{1}$$

with $\mathrm{i} = \sqrt{-1}$ and $\lambda \in [-\pi, \pi]$. For a general function g that is integrable on $[-\pi, \pi]$, we define its inverse Fourier transform via

$$\gamma_g(h) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda) \mathrm{e}^{\mathrm{i}h\lambda} \, \mathrm{d}\lambda,$$

a relation that we will use repeatedly in what follows. That is, γ_g and g are Fourier transform pairs. Furthermore, denote the $n \times n$ Toeplitz matrix associated with g by $\Sigma(g)$, which is defined by

$$\Sigma_{jk}(g) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda) \mathrm{e}^{\mathrm{i}(j-k)\lambda} \mathrm{d}\lambda.$$

So if f is the spectral density of a stationary process, $\Sigma(f)$ is the associated $n \times n$ autocovariance matrix. Finally, let $\widehat{f}(\lambda)$ denote the periodogram defined on a continuum of frequencies:

$$\widehat{f}(\lambda) = \frac{1}{n} \left| \sum_{t=1}^{n} X_t \mathrm{e}^{-\mathrm{i}t\lambda} \right|^2 = \sum_{h=1-n}^{n-1} R(h) \mathrm{e}^{-\mathrm{i}h\lambda} \quad \lambda \in [-\pi, \pi],$$

with $R(h)$ equal to the sample (uncentered) autocovariance function. We adopt the notation

$$\theta(g) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\lambda)\, \mathrm{d}\lambda. \tag{2}$$

The main proximity/disparity measure that we consider in this paper is a spectral variance of the logged spectral density, given by

$$\psi(f) = \theta(\log^2 f) - \theta^2(\log f), \tag{3}$$

where f is a positive spectral density. The empirical version of this measure is obtained by replacing f by the periodogram $\widehat{f}$, and replacing the integrals by a Riemann sum over grid points located at the Fourier frequencies. This is further discussed in Section 3; here we discuss some of the properties of (3) and its motivation.

In [15] a gof measure for time series data is introduced and developed, which is based on the logarithm of the determinant of the (empirical) autocorrelation matrix. The use of this quantity for gof tests is justified in several ways in [15]; for one, this quantity appears in the logarithm of the Gaussian likelihood function. Note that the above authors apply this statistic to the residual autocorrelations obtained from fitting a time series model to the data, and thus their method is similar in spirit to the use of Ljung–Box statistics [10]. Essentially, the statistic of Peña and Rodríguez [15] can be written as

$$\widehat{D} = \frac{1}{n} \log \det \Sigma(\widehat{f}).$$

The above authors consider the case that Σ is m-dimensional, where m is a fixed integer (i.e., it does not expand with sample size in their asymptotics). Also they consider the autocovariance matrix associated with estimated residuals normalized to have unit sample variance, and thus $\widehat{f}$ would be the periodogram of such residuals. Now following the treatment in ([16], Section 7.2.2), under some conditions there exist approximations for the Toeplitz matrices $\Sigma(g)$ of the form

$$\Sigma(g) \doteq Q\, D(g)\, Q^* \tag{4}$$

with $Q_{jk} = n^{-1/2} \exp\{\mathrm{i}2\pi jk/n\}$, and $*$ denoting the conjugate transpose. Here $D(g)$ is a diagonal matrix with entries given by $g(2\pi k/n)$ for $k = 1, \ldots, n$. Substituting $\widehat{f}$ for g in (4), we obtain $\frac{1}{n} \sum_{k=1}^{n} \log \widehat{f}(2\pi k/n)$, which is the Riemann-sum approximation of

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \log \widehat{f}(\lambda)\, \mathrm{d}\lambda. \tag{5}$$

Although such a measure looks quite a bit different from the $\widehat{D}$ of [15], the above heuristics show that it is similar in spirit to their statistic.

Now, the theoretical measure associated with (5) is $\theta(\log f)$ as in (2), a log moment of the spectrum. This is similar to the spectral moment approach of Miller and Rochwarger [11], although that work considers integrals of polynomials multiplying the spectrum. The idea is that moments of the spectrum (or log spectrum in our case) can reveal important properties of the frequency domain representation of a time series, and thus a statistic based on this measure will be useful for assessing the gof of a particular time series model.

In [11] some of the interest focuses on a spectral variance, which can be used to assess the spread (or entropy, loosely defined) in a spectrum. In a similar fashion, we will focus on the spectral variation measure $\psi(f)$ rather than $\theta(\log f)$. It is easy to see that

$$\psi(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} (\log f(\lambda) - \theta(\log f))^2 \, \mathrm{d}\lambda. \tag{6}$$

In contrast, the spectral measure of Drouiche [7] is $D(f) = \log\theta(f) - \theta(\log f)$. Now for any positive bounded f, we have from (6) the following properties of $\psi(f)$:

1. ψ is nonnegative.
2. $\psi(c\, f) = \psi(f)$ for any $c > 0$.
3. $\psi(f) = 0$ iff $f(\lambda) \propto 1$.
4. $\psi(f) = \psi(1/f)$.

Now the D functional of Drouiche [7] possesses (1)–(3), but not (4) in general. This latter property has the following benefit. If f is the spectrum of theoretical model errors, we may consider that peaks and troughs in f are equally persuasive in indicating departures from whiteness; in particular, ψ gives equal measure to both f and its reciprocal, so that peaks and troughs contribute equally to the assessment of nonwhiteness.

3. Examples: $\psi(f)$ for some familiar models

Recall that the basic EXP model [2] can be defined by a Fourier expansion of the log spectrum:

$$\log f(\lambda) = \sum_j \xi_j \mathrm{e}^{-\mathrm{i}\lambda j}.$$

By the evenness of the spectrum, $\xi_j = \xi_{-j}$. An EXP(m) model has $\xi_j = 0$ for $j > m$. Now for such models the spectral variation measure is given by

$$\psi(f) = \sum_{j \neq 0} \xi_j^2.$$

Since the ξ_0 parameter essentially corresponds to the scale of the spectral density, it makes sense (given the comments above) that ξ_0 is omitted from the summation. Below we give some simple examples of this formula.

Example 1 (*EXP(1)*). Let $\log f(\lambda) = \xi_0 + \xi_1(\mathrm{e}^{-\mathrm{i}\lambda} + \mathrm{e}^{\mathrm{i}\lambda})$. Then we obtain

$$\psi(f) = 2\xi_1^2 \left(1 + \gamma(2) - 2\gamma^2(1)\right).$$

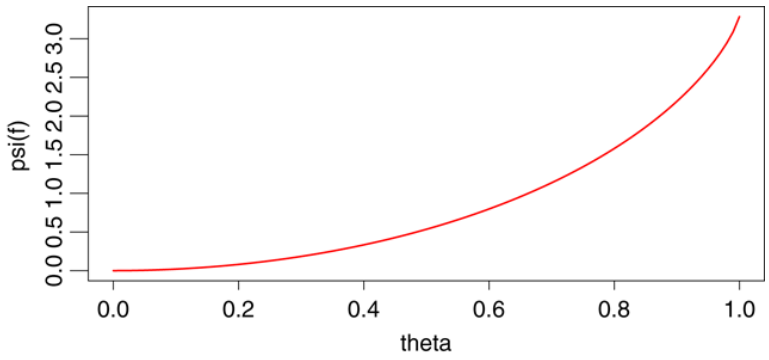

Fig. 1. This is a graph of θ vs. $\psi(f)$. Using this graph one can find equivalent MA(1) processes associated with different values of $\psi(f)$.

Example 2 (*MA(1)*). Let $f(\lambda) = |1 - \theta e^{-i\lambda}|^2 \sigma^2$, so that

$$\log f(\lambda) = \log \sigma^2 - \sum_{j \geq 1} \frac{\theta^j}{j} (e^{i\lambda j} + e^{-i\lambda j}).$$

Hence $\xi_j = -\theta^{|j|}/|j|$ and

$$\psi(f) = \sum_{j \neq 0} \frac{\theta^{2|j|}}{j^2} = 2 \int_0^{\theta^2} -\frac{\log(1-u)}{u}\, du.$$

It is clear that the overall spread of f increases with $|\theta|$, and $\psi(f)$ picks this up in a fairly direct fashion, giving a measure bounded between zero and $\pi^2/3$; Fig. 1 plots θ against $\psi(f)$. Let $h(x) = 2\int_0^x -\frac{\log(1-u)}{u}\, du$ (when $x < 0$, we have $h(x) = 2\int_x^0 \frac{\log(1-u)}{u}\, du$). Now by property 4 of Section 2, the preceding analysis also holds for $AR(1)$ models.

Example 3 (*MA(2)*). First considering the case that the MA polynomial has two real roots, we can write

$$f(\lambda) = |1 - \theta_1 e^{-i\lambda}|^2 |1 - \theta_2 e^{-i\lambda}|^2 \sigma^2$$
$$\log f(\lambda) = \log \sigma^2 - \sum_{j \geq 1} \frac{\theta_1^j + \theta_2^j}{j} (e^{i\lambda j} + e^{-i\lambda j}).$$

This defines ξ_j, and we obtain

$$\psi(f) = \sum_{j \neq 0} \frac{(\theta_1^{|j|} + \theta_2^{|j|})^2}{j^2} = h(\theta_1^2) + 2h(\theta_1\theta_2) + h(\theta_2^2).$$

If there are complex conjugate roots, we can write

$$f(\lambda) = |1 - \rho e^{-i(\lambda - \omega)}|^2 |1 - \rho e^{-i(\lambda + \omega)}|^2 \sigma^2.$$

By the same Taylor series techniques, we find that

$$\xi_j = -\frac{\rho^{|j|}}{|j|} 2 \cos \omega j.$$

Hence the spectral variance measure is

$$\psi(f) = \sum_{j \neq 0} \frac{\rho^{2|j|}}{j^2} 4\cos^2 \omega j.$$

It is interesting that the spectral peak/trough location parameter ω affects the variation. There is more variability in f when the peak/trough is in the center ($\omega = 0$) or at the end ($\omega = \pm\pi$). Also, the variation increases with ρ, which parametrizes the strength of the peak/trough.

4. Statistical properties

Although $\psi(\widehat{f})$ is our statistic of interest, there is some question of how to compute it, given that it involves an integral of the periodogram. The most straightforward approach – following [5] and [16] – is to use a Riemann-sum approximation with mesh points given by the Fourier frequencies. A delicate (and non-obvious) issue is that the asymptotic variance of any integral approximation to $\psi(\widehat{f})$ depends on the mesh size – see [6] for a related discussion. For coherency of treatment with the literature, we use the approximation described in Section 6.4 of [16]:

$$\theta(\widehat{f}) \approx \frac{1}{n} \sum_{j=-n/2}^{n/2} \widehat{f}(\lambda_j).$$

Here we suppose that n is even (else replace $n/2$ by its greatest integer), and $\lambda_j = 2\pi j/n$ (the Fourier frequencies). The generalizations to $\theta(\log \widehat{f})$, etc. are obvious. We denote these Riemann-sum approximation to θ and ψ via $\tilde{\theta}$ and $\tilde{\psi}$ (although these approximations depend on n, this will be suppressed in the notation).

We now present the asymptotic theory for the measure $\tilde{\psi}(\widehat{f})$, which is an estimate of $\psi(\tilde{f})$. Here $\tilde{f}$ is the true spectral density of the data, and may differ from a specified model f. Theorem 1 gives the joint asymptotic normality result for the first and second log moments, i.e., $\tilde{\theta}(\log \widehat{f})$ and $\tilde{\theta}(\log^2 \widehat{f})$. The basic assumption that we use is that the data are Gaussian. Let Γ denote the gamma function, and let $\dot{\Gamma}(x)$ denote the first derivative of the gamma function at x (and so on for higher derivatives).

Theorem 1. *Suppose that $\{X_t\}$ is a stationary zero mean Gaussian time series with $\sum_k |k\gamma_X(k)| < \infty$ and spectral density bounded away from zero. Then*

$$\begin{bmatrix} \sqrt{n}\left(\tilde{\theta}(\log \widehat{f}) - \theta(\log \tilde{f}) - \dot{\Gamma}(1)\right) \\ \sqrt{n}\left(\tilde{\theta}(\log^2 \widehat{f}) - \theta(\log^2 \tilde{f}) - 2\dot{\Gamma}(1)\theta(\log \tilde{f}) - \ddot{\Gamma}(1)\right) \end{bmatrix}$$

is asymptotically bivariate normal with zero mean vector and variance matrix V. The entries of V are given by:

$$\begin{aligned} V_{11} &= 2\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right) \\ V_{12} &= 4\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right)\theta(\log \tilde{f}) + 2\left(\dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1)\right) \\ V_{22} &= 8\left(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1)\right)\theta(\log^2 \tilde{f}) + 8\left(\dddot{\Gamma}(1) - \dot{\Gamma}(1)\ddot{\Gamma}(1)\right)\theta(\log \tilde{f}) \\ &\quad + 2\left(\ddddot{\Gamma}(1) - \ddot{\Gamma}^2(1)\right). \end{aligned}$$

Corollary 1. *Under the same assumptions and notation of Theorem 1,*

$$\sqrt{n}\left(\tilde{\psi}(\widehat{f}) - \psi(\tilde{f}) - \ddot{\Gamma}(1) + \dot{\Gamma}^2(1)\right) \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0, W)$$
$$W = 8(\ddot{\Gamma}(1) - \dot{\Gamma}^2(1))(\psi_1(\tilde{f}) + \dot{\Gamma}^2(1)) + 2(\dddot{\Gamma}(1) - \ddot{\Gamma}^2(1)) - 8(\dot{\Gamma}(1)\dddot{\Gamma}(1) - \ddot{\Gamma}(1)\dot{\Gamma}^2(1)).$$

5. Goodness-of-fit and badness-of-fit testing

We now focus on testing model errors for whiteness. Suppose that a fitted model is used to compute model residuals, which are denoted by $\widehat{\varepsilon}_t$. We then compute $\tilde{\psi}(\widehat{f})$ based on these model residuals, which are assumed to have true spectral density $\tilde{f}$. Assuming that $\widehat{\varepsilon}_t$ are Gaussian and $\tilde{f}$ is positive, we know by Corollary 1 that $\sqrt{n}\tilde{\psi}(\widehat{f})$ is asymptotically normal with mean and variance computable from a knowledge of $\psi(\tilde{f})$. Note that this formulation differs from the approach of gof testing, which makes assumptions about the spectrum of the *model errors* rather than the model residuals.

We illustrate this distinction with a short example. Consider an AR(1) process $X_t = \phi X_{t-1} + \varepsilon_t$, where ϕ is estimated by some estimator $\widehat{\phi}$ computed from the observed data. The model errors are $\varepsilon_t = X_t - \phi X_{t-1}$, whereas the model residuals are $\widehat{\varepsilon}_t = X_t - \widehat{\phi} X_{t-1}$ (which are available for times $t = 2, 3, \ldots, n$). A gof test (such as Ljung–Box) uses the Null Hypothesis that the AR(1) model is the correct specification, or equivalently that the model errors are *iid*. Our procedure, in contrast, uses the Null Hypothesis that the model residuals are white noise, or equivalently that the AR(1) model is the correct specification *and* our model parameter $\widehat{\phi}$ is correct, i.e., equal to the truth.

Our proximity test seeks to reject the whiteness of model residuals (which gives evidence of model misspecification and/or incorrect parameter values), which is equivalent to $\psi(\tilde{f}) = 0$ by Property 3 in Section 2. So for the proximity procedure, we have

$$H_0 : \psi(\tilde{f}) = 0$$
$$H_a : \psi(\tilde{f}) > 0.$$

Note that by Corollary 1, we can determine asymptotic power with only a knowledge of the value of $\psi(\tilde{f})$; it is not necessary to know the full spectrum $\tilde{f}$, only its characterization through ψ. For disparity testing, we instead specify a given level of nonwhiteness $\mu_0 > 0$, and test the hypotheses

$$H_0 : \psi(\tilde{f}) = \mu_0$$
$$H_a : \psi(\tilde{f}) < \mu_0.$$

Here we note that the alternative is lower one-sided, in contrast to the upper one-sided proximity procedure. Since by Corollary 1 the asymptotic distribution of $\tilde{\psi}(\widehat{f})$ only depends on $\tilde{f}$ through the measure $\psi(\tilde{f})$, we can compute the mean and variance under H_0. It is not clear how to adapt Ljung–Box statistics or $\widehat{D}$ of [15] to disparity testing, since the asymptotic distribution of these statistics under nonwhiteness of the underlying process is either unknown, or is a complex function of the entire spectral density.

What is the advantage of disparity testing over proximity testing? As mentioned in Section 1, gof tests are used to find bad models, whereas bof tests can be used to find good models. For a proximity test, we seek to reject the Null Hypothesis (that the model fit is perfect) with significant

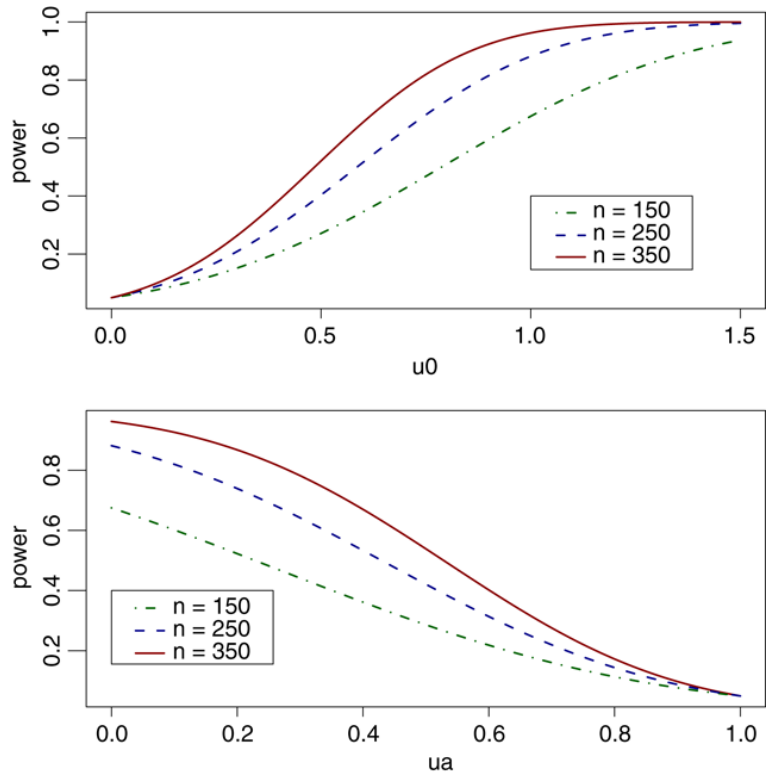

Fig. 2. The top panel of this figure contains a graph of the maximum theoretical power for $\mu_0 \in [0, 1.5]$ (i.e., $\mu_a = 0$). The bottom panel displays the theoretical power for different values of μ_a when μ_0 is held fixed and equal to 1.

p-values. Failing to reject this Null Hypothesis corresponds to having a good model, but how high should the p-values be? Rather than trying to fail to reject by getting large p-values, the disparity diagnostic can be used as an alternative, with a small p-value indicating rejection of the badness of a model (together with its parameter values) — in the direction of a better model. The drawback is that one must specify in H_0 a level of disparity μ_0. Our perspective is that disparity diagnostics are useful as a complement to the existing gof tests.

Of course such a procedure requires that the asymptotic distribution of the test statistic only depends on $\tilde{f}$ through the chosen characterization of whiteness. Moreover, the power of the procedure will understandably depend on the choice of μ_0; in fact, the power will be an increasing function of μ_0. Fig. 2 gives an indication of these relationships, using the formula for the asymptotic power of the disparity test given below. Let $c_1 = \ddot{I}(1) - \dot{I}^2(1)$ and

$$c_2 = 8\left(c_1\dot{I}^2(1) - \dot{I}(1)\dddot{I}(1) + \ddot{I}(1)\dot{I}^2(1)\right) + 2(\ddddot{I}(1) - \ddot{I}^2(1)).$$

Note that $c_1 = \pi^2/6$ and $c_2 \approx 23.811$. The disparity asymptotic power for a δ-level test is given by

$$\Phi\left(\frac{\sqrt{n}(\mu_0 - \mu_a) + z_\delta\sqrt{8c_1\mu_0 + c_2}}{\sqrt{8c_1\mu_a + c_2}}\right), \tag{7}$$

where Φ is the standard normal cdf and $z_\delta = \Phi^{-1}(\delta)$. Here μ_0 and μ_a are Null and Alternative specifications of $\psi(\tilde{f})$.

There is an exact relationship between the proximity and disparity test statistics. For a specified μ_0 (which equals zero in the proximity context), we define $\widehat{\psi}(\mu_0)$ to be the

corresponding test statistic, which is given by

$$\widehat{\psi}(\mu_0) = \sqrt{n}\,\frac{\tilde{\psi}(\widehat{f}) - \mu_0 - c_1}{\sqrt{8c_1\mu_0 + c_2}}.$$

Now the disparity test statistic is given by the above formula when $\mu_0 > 0$, but the proximity test statistic corresponds to $\widehat{\psi}(0)$. These test statistics are related by the formula

$$\widehat{\psi}(\mu_0) = \frac{\sqrt{c_2}\widehat{\psi}(0) - \mu_0\sqrt{n}}{\sqrt{8c_1\mu_0 + c_2}}. \tag{8}$$

Thus, there is an equivalence between proximity and disparity testing, and the link is μ_0. We see from (8) that small values of $\widehat{\psi}(0)$ – associated with failure to reject proximity – result in negative values of $\widehat{\psi}(\mu_0)$, so long as $\mu_0\sqrt{n/c_2}$ is larger than the proximity test statistic. Thus, $\widehat{\psi}(\mu_0)$ is significant if μ_0 is large enough, or if the sample size is large enough. In general, large proximity p-values result in small disparity p-values, and vice versa, so long as a μ_0 is suitably large with respect to the sample size. In fact, (8) can be used to determine the choice of the threshold μ_0 in disparity testing. Letting α and δ be the significance levels of the proximity and disparity procedures respectively, we obtain the relation

$$z_\delta = \frac{\sqrt{c_2}z_{1-\alpha} - \mu_0\sqrt{n}}{\sqrt{8c_1\mu_0 + c_2}}. \tag{9}$$

If the disparity procedure were significant at the 5% level (i.e., $\delta = 0.05$), then we would like the proximity procedure to fail to reject with at least α probability, where now α is larger than 0.05 and can be selected by the user. For example, one might choose $\alpha = 0.50$ or $\alpha = 0.20$. One can easily show via (7), that for μ_0 satisfying (9), the disparity power is approximately $1 - \alpha$ when the alternative is white noise (i.e., $\mu_a = 0$). In other words, given a δ-level disparity test, we can choose μ_0 according to (9) to ensure approximate $1 - \alpha$ power against a white noise alternative, where α is chosen by the user. Solving (9) for μ_0 when $\delta = 0.05$, we obtain two roots by the quadratic formula. So long as $0.05 < \alpha < 0.95$, the smaller root is negative, and so we let μ_0 be given by the larger root:

$$\mu_0 = \frac{\sqrt{c_2}z_{1-\alpha}}{\sqrt{n}} + \frac{4c_1z_{0.05}^2}{n} + \sqrt{\frac{16c_1^2z_{0.05}^4}{n^2} + \frac{8c_1\sqrt{c_2}z_{1-\alpha}z_{0.05}^2}{n^{3/2}} + \frac{c_2z_{0.05}^2}{n}}. \tag{10}$$

We denote this choice of μ_0 by $\mu_0(n)$ (since it depends on sample size), when we use (10) to determine the parameter. To re-iterate the property of this choice of μ_0: when the disparity test statistic has p-value less than 0.05, the proximity will have p-value greater than α percent, and the disparity will have approximate power $1 - \alpha$ against the white noise alternative. Since this represents a fairly intuitive relationship between disparity and proximity testing, we recommend using (10) for determining μ_0. A depiction of this relation is given in Fig. 3 for $\alpha = 0.2$ and $\delta = 0.05$.

6. Empirical studies

In this section we evaluate both the proximity and disparity measures using Monte Carlo simulation and real time series data. Although our measure is similar to the diagnostics of Peña and Rodríguez [15], we do not make direct comparisons here. Instead we make comparisons

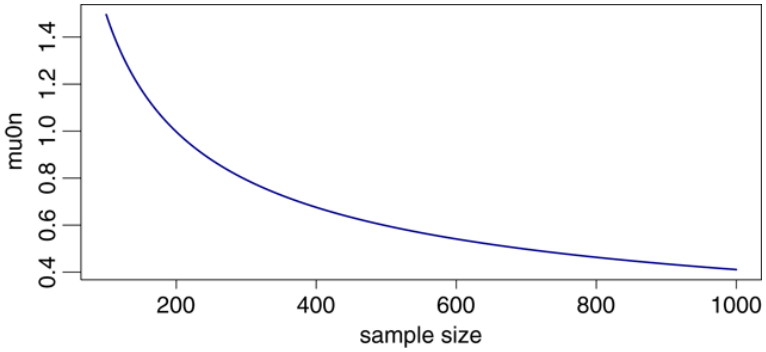

Fig. 3. This is a plot of μ_0 as a function of sample size such that $\alpha = 0.2$ and $\delta = 0.05$.

to the Ljung–Box (LB) statistic. Of course, LB is a gof procedure, and can be used to indicate the inadequacy of a model class; in contrast, our proximity test could be significant when the model class is correct but the parameter values are faulty. Another reason we do not make a direct comparison to [15] is that we were unable to duplicate the results presented in their simulation studies, and even assuming the results of the simulation study they conduct for their gof diagnostic are correct, they claim to beat the performance of the LB statistic. We acknowledge that the LB gof diagnostic out-performs our proximity diagnostic (although, as noted above, they assess somewhat different aspects of modeling time series data), and we present the results of a simulation study that quantifies to what extent its performance is superior.

6.1. Simulations

We determine the size and power of both our proximity/disparity diagnostics under several different departures from white noise residuals. First we consider the size of our proximity diagnostic under the Null Hypothesis of white noise. Additionally, we evaluate the distribution of our test statistic in finite samples. In order to do this we performed 10,000 Gaussian simulations of various samples sizes ($n = 150, 250, 350, 500$) and calculated the mean, standard deviation and α-level of our diagnostic under a nominal α-level of $\alpha = 0.05$. Further, we investigated the size of the LB under identical conditions using $m = 5, 10$, and 20 autocorrelations in the calculation of the statistic (see Table 1). Although the size of the LB statistic is moderately better than ours, both diagnostics are fairly close to the nominal level and approach 0.05 as the sample size increases, as expected. Similarly the mean and standard deviation of our test statistic under the Null Hypothesis approach the correct mean and standard deviation as the sample size increases (see Table 1). Although it is crucial that the mean and standard deviation approach 0 and 1 respectively, it is equally important that the distribution be normal. As can be seen from Fig. 4, the distribution of the proximity statistic is well-approximated by the normal for sample size 500.

Next, we compare the power of our proximity diagnostic with the power of the LB statistic (at $m = 5, 10, 20$) and the turning point diagnostic for independence (see [3], pp. 312–313), defined by the asymptotic distribution of the number of turning points in the series of model residuals. To assess the performance we simulated from an MA(1) data generating process with $\theta = 0.9$. We then computed model residuals obtained from an MA(1) model with θ ranging between 0.1 and 0.8. The residuals were then tested for whiteness, as discussed in Section 5. In our simulation as θ decreases from 0.9, the departure from whiteness in the estimated residuals increases, and it should be easier to reject the H_0. This simulation was conducted at the nominal α-level of 0.05 with 1000 simulations of various sample sizes (see Table 2). In general our power does not

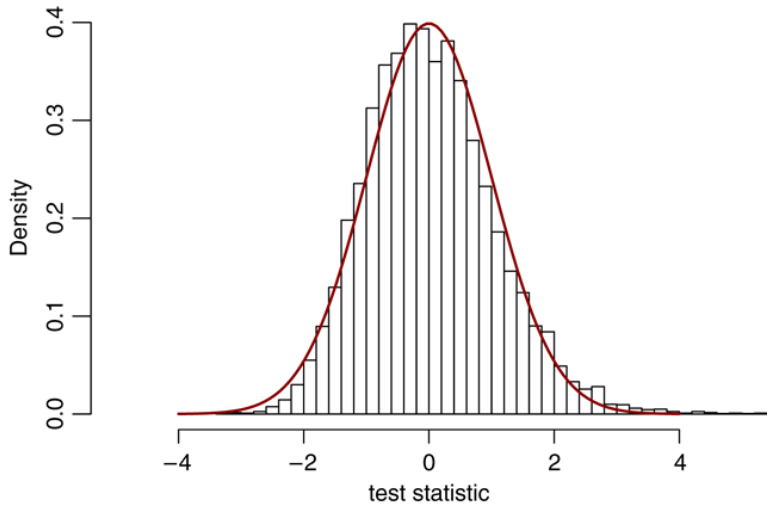

Fig. 4. This figure contains a histogram for the distribution of the proximity statistic from a simulation with 10,000 repetitions of sample size 500, under a white noise Null Hypothesis. Note that the theoretical Normal(0, 1) pdf is superimposed for convenience.

Table 1
Comparison of level for proximity diagnostic to Ljung–Box

n	ψ			Ljung–Box: α-level		
	Mean	Std.	α-level	$m = 10$	$m = 15$	$m = 20$
150	0.081	1.091	0.0775	0.0541	0.0605	0.0646
250	0.060	1.056	0.0725	0.0555	0.0597	0.0652
350	0.042	1.033	0.0697	0.0528	0.0559	0.0592
500	0.026	1.032	0.0666	0.0493	0.0554	0.0588

Note that the number of Monte Carlo simulations was 10,000, at a nominal α-level of 0.05. Distributional results are presented in the form of mean and standard deviation.

perform as well as the LB or turning point statistics. However, one advantage of our proximity diagnostic (incidentally, shared by Drouiche [7]) over LB is that only one test statistic is formed and only one p-value is produced. So the practitioner is freed from having to choose m (and deal with multiple testing issues) when testing for model adequacy.

We now turn our attention to the disparity diagnostic. For this diagnostic we evaluated the power under several formulations. In the first simulation we chose the threshold value, $\mu_0(n)$, adaptively based on the sample size using (10) with $\alpha = 0.2$ and $\delta = 0.05$. This method of choosing μ_0 induces a Null Hypothesis $\psi(\tilde{f}) = \mu_0$ with an Alternative hypothesis $\psi(\tilde{f}) < \mu_0$, while simultaneously controlling the power. Further in this first power study we suppose that the model residual process follows an MA(1) with parameter θ between 0 and approximately 0.42. We simulated Gaussian MA(1) processes with $\theta = 0.4, 0.3, 0.2, 0.1, 0$ for various sample sizes between $n = 100$ and $n = 500$ and determined power using 10,000 Monte Carlo replications and nominal level equal to 0.05 (see Table 3). As a result of choosing μ_0 via (10) we find that the power is rather good even for sample sizes as small as $n = 100$. In fact, we see that the power of our test remains fairly constant across sample size and is approximately equal to 0.8 for a white noise alternative. Here it is $\mu_0(n)$ that changes with sample size. Specifically, $\mu_0(n)$ is a decreasing function of n, and as such when the sample size increases so does our assurance that rejecting the Null Hypothesis of a "bad" model in the direction of a better fitting model actually results in a "good" model, and not just a model superior to that postulated under the

Table 2
This table compares the power of our proximity diagnostic (ψ) with the Ljung–Box (LB) and Turning Point diagnostics, by considering the residuals as a function of the MA(1) parameter θ

θ	ψ	LB $m = 5$	LB $m = 10$	LB $m = 20$	Turning Point
Power for sample size $n = 150$					
0.8	0.102	0.106	0.112	0.117	0.106
0.7	0.193	0.241	0.203	0.206	0.235
0.6	0.306	0.485	0.417	0.377	0.467
0.5	0.470	0.705	0.598	0.554	0.662
0.4	0.675	0.899	0.793	0.735	0.834
0.3	0.820	0.981	0.915	0.884	0.940
0.2	0.921	0.999	0.987	0.968	0.981
0.1	0.961	1	0.997	0.990	0.996
Power for sample size $n = 250$					
0.8	0.115	0.173	0.169	0.156	0.154
0.7	0.228	0.494	0.398	0.373	0.391
0.6	0.413	0.828	0.731	0.665	0.678
0.5	0.607	0.978	0.925	0.874	0.913
0.4	0.849	0.998	0.993	0.977	0.972
0.3	0.948	1	0.999	0.996	0.999
0.2	0.987	1	1	1	1
0.1	0.998	1	1	1	1
Power for sample size $n = 350$					
0.8	0.116	0.260	0.236	0.204	0.195
0.7	0.275	0.731	0.593	0.526	0.518
0.6	0.469	0.986	0.944	0.849	0.843
0.5	0.779	1	1	0.991	0.982
0.4	0.935	1	1	1	0.998
0.3	0.997	1	1	1	1
0.2	0.999	1	1	1	1
0.1	1	1	1	1	1

Note that the number of Monte Carlo simulations was 1000, at a nominal α-level of 0.05 and the data generating process was a MA(1) with $\theta = 0.9$.

Table 3
This table examines the power of the disparity diagnostic under several different departures from white noise

Power with μ_0 as a function of sample size – $\mu_0(n)$								
H_a: θ	100 (1.495)	150 (1.177)	200 (0.997)	250 (0.878)	300 (0.793)	350 (0.728)	400 (0.676)	500 (0.598)
0.4	0.5733	0.5160	0.4705	0.4286	0.3837	0.3571	0.3111	0.2636
0.3	0.6826	0.6610	0.6223	0.6100	0.5888	0.5550	0.5395	0.5070
0.2	0.7496	0.7283	0.7332	0.7192	0.7160	0.6980	0.6971	0.6854
0.1	0.7878	0.7877	0.7799	0.7736	0.7801	0.7767	0.7814	0.7768
WN	0.8025	0.8016	0.7919	0.7987	0.7974	0.7944	0.7992	0.7951

Note that the number of Monte Carlo simulations was 10,000, at a nominal α-level of 0.05. Also μ_0 is determined as a function of sample size with $\alpha = 0.2$ and $\delta = 0.05$. The second row of the table gives sample sizes n, with the corresponding $\mu_0(n)$ in parenthesis.

Table 4
This table examines the power of the disparity diagnostic under several different departures from white noise

H_a: θ	H_0: $\theta \approx 0.66$ or $\mu_0 = 1$			H_0: $\theta \approx 0.48$ or $\mu_0 = 0.5$		
	$n = 150$	$n = 250$	$n = 350$	$n = 150$	$n = 250$	$n = 350$
0.4	0.3803	0.5475	0.6724	0.0648	0.0945	0.1167
0.3	0.5200	0.7122	0.8359	0.1271	0.2081	0.2694
0.2	0.6231	0.8119	0.9025	0.2066	0.3178	0.4090
0.1	0.6684	0.8524	0.9318	0.2653	0.4002	0.5022
WN	0.7011	0.8654	0.9400	0.2836	0.4297	0.5483

Note that the number of Monte Carlo simulations was 10,000, at a nominal α-level of 0.05.

Null. Additionally, our simulations confirm that the power under a white noise alternative turns out to be approximately $1 - \alpha$.

In our second power study we mapped the different values of μ_0 into equivalent MA(1) processes; if θ is the MA parameter, then $\mu_0 = 2\sum_{j=1}^{\infty} \theta^{2j}/j^2$ as shown in Example 2 of Section 3. For a graph of this mapping see Fig. 1. Thus here we are keeping μ_0 fixed across different sample sizes. First we suppose that the model residual process follows an MA(1) with parameter θ between 0 and 0.66, which corresponds to $\psi(\tilde{f})$ values between 0 and 1. Our Null Hypothesis states that $\psi(\tilde{f}) = 1$, and the Alternative Hypothesis states that $\psi(\tilde{f}) < 1$, or equivalently that $\theta < 0.66$. Again, we simulated Gaussian MA(1) processes with $\theta = 0.4, 0.3, 0.2, 0.1, 0$, with three different sample sizes, and determined the power using 10,000 Monte Carlo replications and an α level of 0.05. Secondly, we let $\mu_0 = 0.5$, which for an MA(1) corresponds to $\theta = 0.48$. So we simulated Gaussian MA(1) processes with $\theta = 0.4, 0.3, 0.2, 0.1, 0$ with three different sample sizes, and determined the power. The results are reported in Table 4. Even though, in this case, the power depends on what size departures one is willing to accept under the Null Hypothesis, this is still a very sensible way to test for model inadequacy. This flexibility allows the practitioner the control in deciding what degree of "badness" is acceptable for a given application. If μ_0 is chosen equal to 0.5 a priori, large samples are needed to achieve high power if one only wishes to consider slight departures from whiteness. On the other hand, for $\mu_0 = 1$ a priori fair power is achieved for relatively moderate sample sizes.

6.2. Case studies

Next, we consider the diagnostics on several time series: *m00110*, *m00100*, *France*, and *Shoe*. The first two time series are from the Foreign Trade Division of the US Census Bureau; the first series is Imports of Meat Products, and the second series is Imports of Dairy Products and Eggs. Both of these series are for the time period from January 1989 to December 2003. The *France* series refers to the sales volume for Grands Magasins produced by the Chamber of Commerce and of Industry of Paris (CCIP), from January 1990 through March 2004. The *Shoe* series is US Retail Sales of Shoe Stores data from the monthly Retail Trade Survey of the Census Bureau, from 1984 to 1998.

In order to illustrate our diagnostics usefulness in practice we fit models to the data using the "automodl" command of the 2007 update (version 0.3) of X-12-ARIMA ([17], which follows closely the automatic modeling procedure of Gómez and Maravall [8]. We adjusted for regression effects (such as outliers and trading day) when applicable. Next we obtained the estimated residuals from the fitted model and calculated our proximity and disparity diagnostic tests using

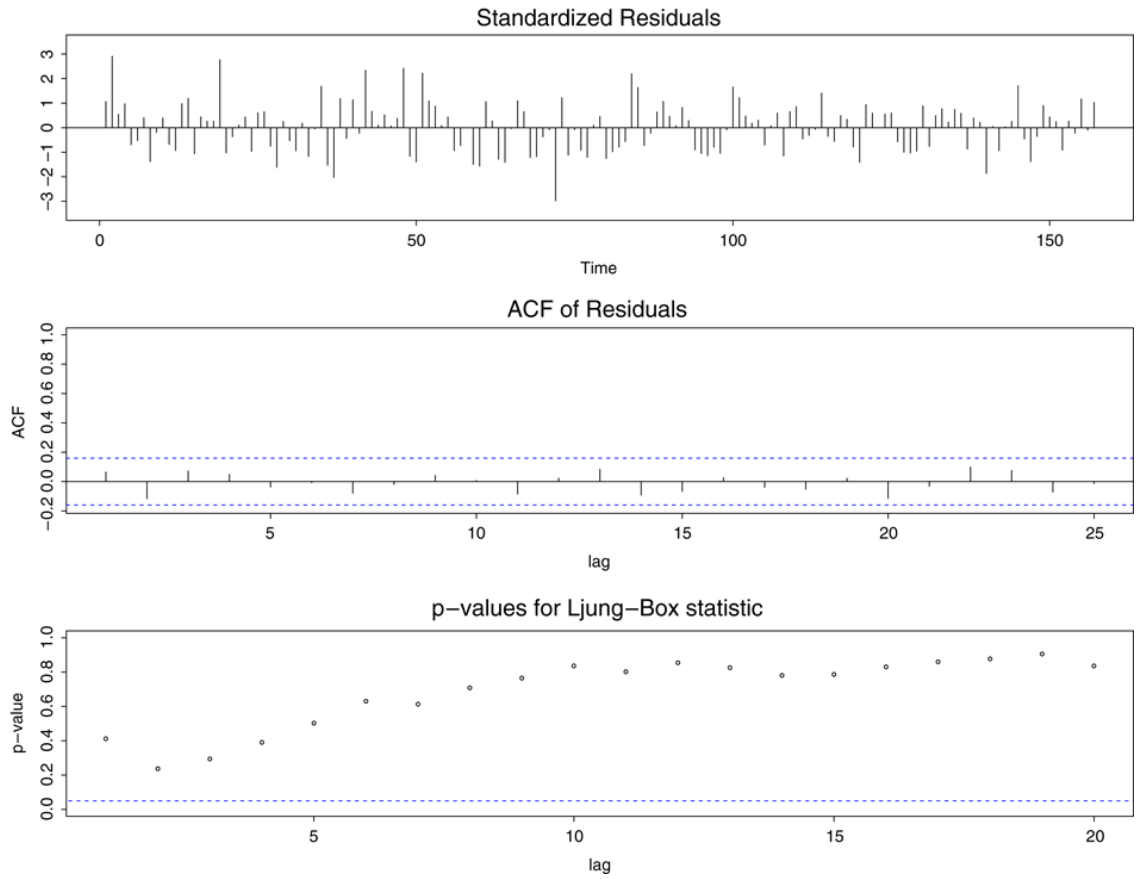

Fig. 5. This figure contains a time series plot of the residuals from the model fit to the *Shoe* data using the automodel feature in X-12-ARIMA, along with a plot of the acf of the residuals. Finally, the p-values of the Ljung–Box statistic are plotted for lags up to 20.

$\mu_0(n)$ with $\alpha = 0.2$ and $\delta = 0.05$. Below, n is the effective number of observations (the sample size of the differenced series). For comparison we constructed a time series plot and acf plot of the residuals along with the p-values through lag 20 for the LB statistic.

The first series we consider is the *Shoe* series ($n = 157$). The automodl procedure of X-12-ARIMA provided the following model:

$$(1 - B)(1 - B^{12})X_t = (1 - 0.572B)(1 - 0.336B^{12})\varepsilon_t.$$

The p-value from our proximity diagnostic was 0.489, while the p-values for our disparity diagnostic with $\mu_0(n) = 1.146$ was 0.011. This can be contrasted with various other measures constructed from the residuals (see Fig. 5). It appears that the acf plot of the residuals and LB statistics seem to indicate an adequate fit. In this case our results agree with this assessment.

The next series we consider is the *France* series ($n = 158$). The automodl procedure of X-12-ARIMA obtained the following model for the log transformed data:

$$(1 - B)(1 - B^{12})X_t = (1 - 1.0382B + 0.3381B^2)\varepsilon_t.$$

The p-value from our proximity diagnostic was 0.927, while the p-values for disparity with $\mu_0(n) = 1.142$ was < 0.001. Next we compared our analysis with the acf plot of the residuals and the LB statistic out to lag 20 (see Fig. 6). In this case our results corroborate the results of the acf plot and the LB test, namely that the model is very good.

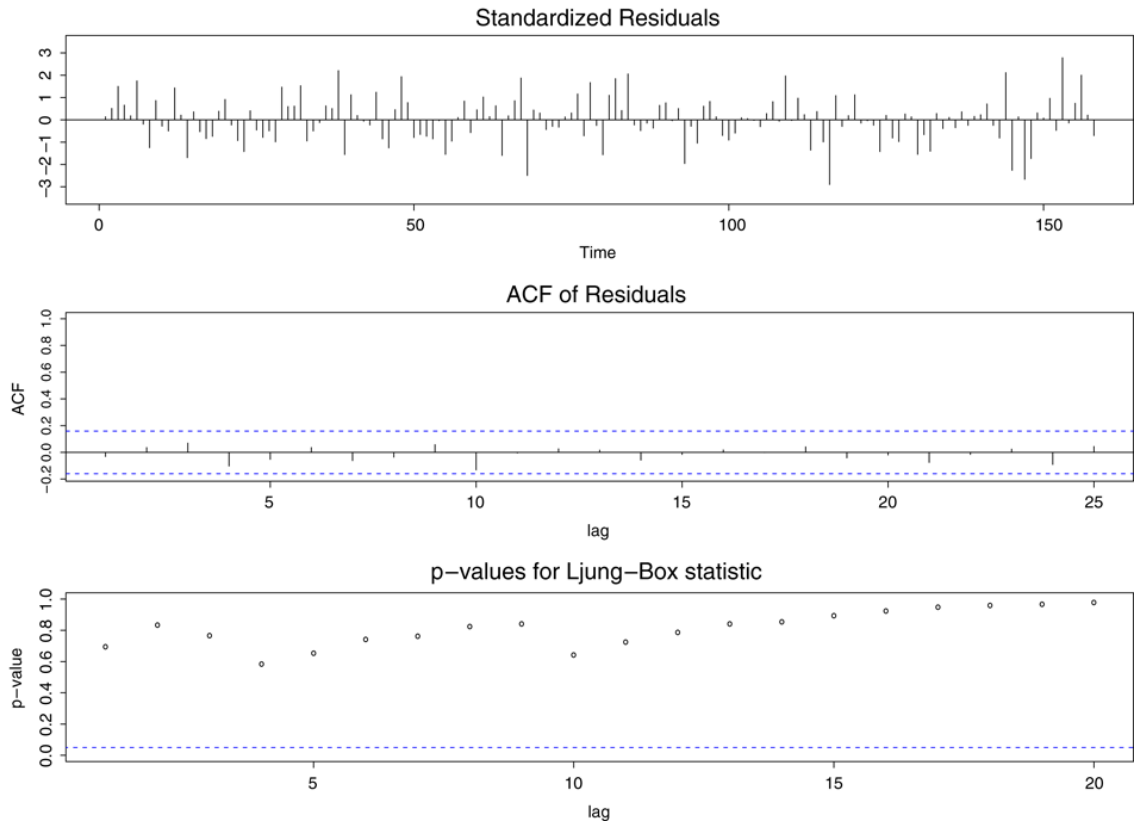

Fig. 6. This figure contains a time series plot of the residuals from the model fit to the *France* data using the automodel feature in X-12-ARIMA, along with a plot of the acf of the residuals. Finally, the p-values of the Ljung–Box statistic are plotted for lags up to 20.

We next focus our attention on the *m00100* series ($n = 167$). The automodl feature of X-12-ARIMA fitted the following model to the log transformed data:

$$(1 - B)(1 - B^{12})X_t = (1 - 0.8390B^{12})\varepsilon_t.$$

The p-value from our proximity diagnostic was 0.134, while the p-values for disparity with $\mu_0(n) = 1.106$ was 0.075. We can compare this with the time series plot and acf of the residuals along with the LB out to lag 20 (see Fig. 7). This example illustrates an instance when using LB and acf plots is somewhat ambiguous. When looking at the LB p-values, the decision rule is dependent on which lag is being considered. Moreover, it requires the practitioner to make a choice as to what constitutes a high p-value.

Finally, we examine the *m00110* series ($n = 167$). The automodl procedure of X-12-ARIMA obtained the following model for the log transformed data:

$$(1 - 0.5907B)(1 - B)(1 - B^{12})X_t = (1 - 0.9380B)(1 - 0.9377B^{12})\varepsilon_t.$$

The p-value from our proximity diagnostic was 0.541, while the p-values for our disparity diagnostics with $\mu_0(n) = 1.106$ was 0.008. This series illustrates an instance where the determination of model adequacy is less definitive. Specifically, the decision to accept a model is markedly different, depending on whether the number of correlations used in the LB test are less than or greater than 14 (see Fig. 8). However, our disparity statistic avoids this complication. The practitioner can use these diagnostics to assess what type of departure from whiteness is deemed acceptable, and take action accordingly.

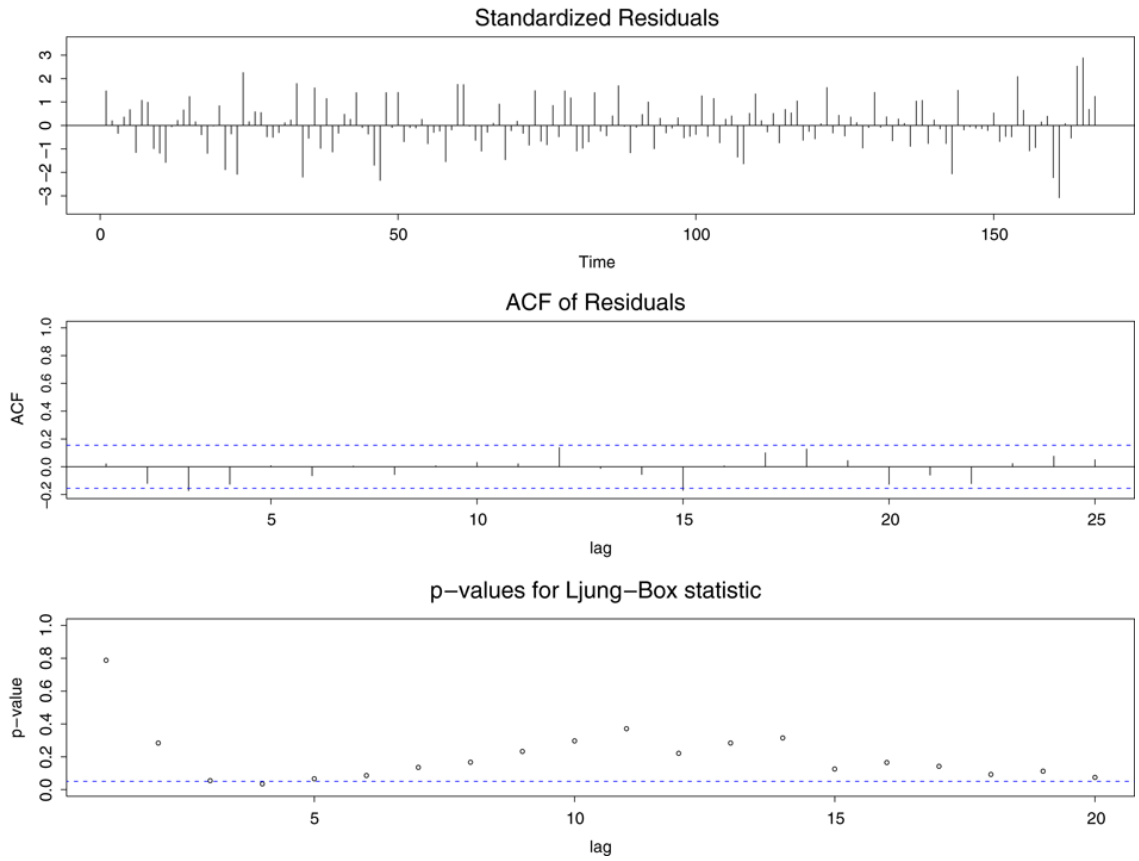

Fig. 7. This figure contains a time series plot of the residuals from the model fit to the *m00100* data using the automodel feature in X-12-ARIMA, along with a plot of the acf of the residuals. Finally, the p-values of the Ljung–Box statistic are plotted for lags up to 20.

6.3. Discussion

This paper introduces the concept of disparity testing for time series modeling, using a convenient measure of whiteness of the spectrum. We explicitly demonstrate how proximity and disparity testing can be implemented, and the relationship between them; we also describe how the crucial μ_0 parameter can be chosen in an intuitive fashion to ensure any degree of statistical power. Our method is illustrated on four economic time series, and the results are compatible with the information from acf and pacf plots.

A potential criticism of disparity testing raises the question of the choice of μ_0. Given a significant rejection of H_0 in the disparity test, can we be assured that the residuals are truly white? How close to white are they then, and are they "close enough"? However, similar questions could be leveled at proximity testing: given that whiteness is rejected with significance, can it be that model residuals are still close enough to being white noise, such that the model fit might be deemed decent – perhaps through some other assessment such as out-of-sample forecasting performance? We have attempted to resolve the issue of the choice of μ_0 as follows: one chooses an α – associated with a proximity test – such that the asymptotic power of the disparity test is $1 - \alpha$ for a white noise alternative, so long as μ_0 is chosen according to (10). In our simulations and case studies we have taken $\alpha = 0.2$, since this gives high power (about 80%) against the white noise alternative. Clearly, other choices are available.

Graphs such as Fig. 9 give an indication of the relationship between $\psi(\tilde{f})$ and flatness of the spectrum for MA(1) models. Furthermore, the four case studies presented illustrate that low p-

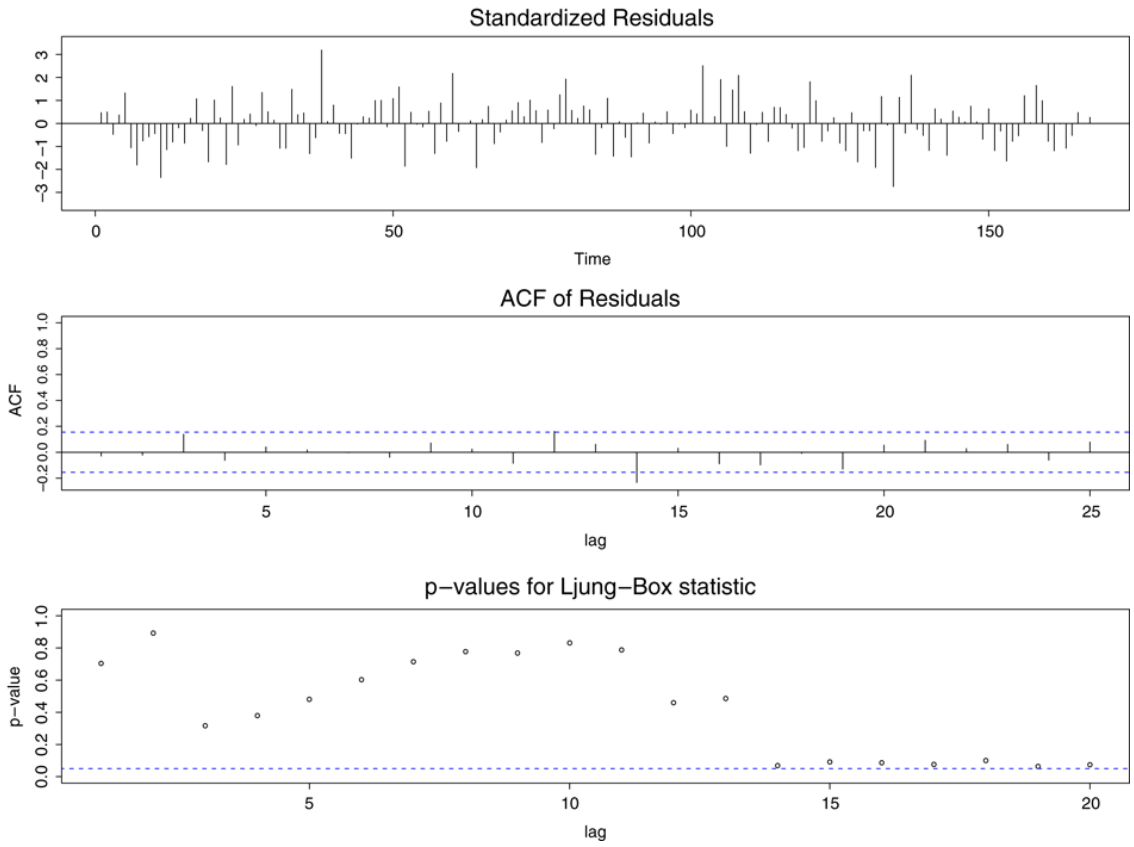

Fig. 8. This figure contains a time series plot of the residuals from the model fit to the *m00110* data using the automodel feature in X-12-ARIMA along with a plot of the acf of the residuals. Finally, the p-values of the Ljung–Box statistic are plotted for lags up to 20.

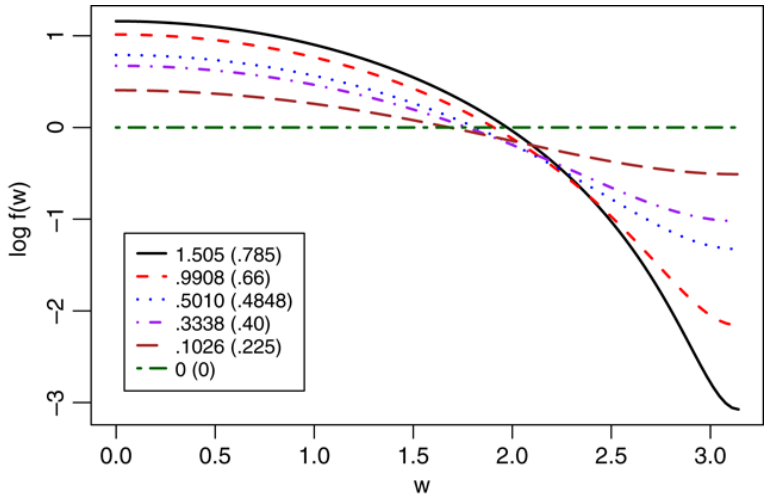

Fig. 9. This figure contains a graph of the theoretical log spectral density for several values of μ_0. Further we display the parameters μ_0, along with their associated MA(1) parameters in parenthesis.

values (the *France* series) indeed correspond to white residuals, whereas moderate p-values (the *Shoe* and *m00110* series) still indicate residuals that are close to being uncorrelated — witness the plots of standardized residuals in Figs. 5 and 8. The high p-value for the *m00100* series seems to indicate that the residuals are too far from whiteness; now one should flip things around, and

do proximity testing to reject the given model (this is a borderline case, because the proximity p-value is 0.134 and only one LB statistic is significant at the 5% level).

In practice, plots of residual ACFs and standardized residuals will be helpful in determining model adequacy; our disparity testing procedure allows for quantization of this concept via an appropriate measure of whiteness, allowing one to "accept" fitted models with statistical significance.

Acknowledgements

Holan's research was supported by an ASA/NSF/BLS research fellowship.

Disclaimer

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, and operational issues are those of the authors and not necessarily those of the US Census Bureau.

Appendix

Proof of Theorem 1. We use the notation for Riemann sums introduced in the beginning of Section 4. We wish to consider the convergence, for any real numbers a and c, of

$$\frac{1}{n}\sum_{j=-n/2}^{n/2}\left(a\log\widehat{f}(\lambda_j)+c\log^2\widehat{f}(\lambda_j)\right)=\frac{1}{n}\sum_{j=-n/2}^{n/2}\left(a\log\widehat{f}(\lambda_j)+c\log^2\widehat{f}(\lambda_j)\right). \quad \text{(A.1)}$$

This can be written as

$$\frac{1}{n}\sum_{j=-n/2}^{n/2}\zeta(\widehat{f}(\lambda_j)),$$

where $\zeta(x)=a\log x+c\log^2 x$. The convergence of this type of functional follows from Theorem 6.4.3 of [16]. Then the asymptotic mean is

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\int_0^{\infty}\zeta(\tilde{f}(\lambda)r)\mathrm{e}^{-r}\,\mathrm{d}r\,\mathrm{d}\lambda.$$

Applying this, we see that the asymptotic mean of (A.1) is given by

$$a\left(\theta(\log\tilde{f})+\dot{\Gamma}(1)\right)+c\left(\theta(\log^2\tilde{f})+2\theta(\log\tilde{f})\dot{\Gamma}(1)+\ddot{\Gamma}(1)\right).$$

Taking $a=1,c=0$ and $a=0,c=1$ respectively gives the means stated in the theorem. Letting

$$v_1(\lambda)=\int_0^{\infty}\zeta^2(\tilde{f}(\lambda)r)\mathrm{e}^{-r}\,\mathrm{d}r-\left(\int_0^{\infty}\zeta(\tilde{f}(\lambda)r)\mathrm{e}^{-r}\,\mathrm{d}r\right)^2$$

the asymptotic variance is given by $\frac{2}{2\pi}\int_{-\pi}^{\pi}v_1(\lambda)\,\mathrm{d}\lambda$. We next compute this variance function:

$$\begin{aligned}v_1(\lambda)=a^2&\left(\ddot{\Gamma}(1)-\dot{\Gamma}^2(1)\right)+2ac\left(2\log\tilde{f}(\lambda)\ddot{\Gamma}(1)\right.\\&\left.-2\log\tilde{f}(\lambda)\dot{\Gamma}^2(1)+\dddot{\Gamma}(1)-\dot{\Gamma}(1)\ddot{\Gamma}(1)\right)+c^2\left(4\log^2\tilde{f}(\lambda)\ddot{\Gamma}(1)\right.\\&\left.-4\log^2\tilde{f}(\lambda)\dot{\Gamma}^2(1)+4\log\tilde{f}(\lambda)\left(\dddot{\Gamma}(1)-\dot{\Gamma}(1)\ddot{\Gamma}(1)\right)+\ddddot{\Gamma}(1)-\ddot{\Gamma}^2(1)\right).\end{aligned}$$

It follows that the asymptotic variance of (A.1) is

$$2a^2\left(\ddot{\Gamma}(1)-\dot{\Gamma}^2(1)\right)+4ac\left(2\theta(\log\tilde{f})\left(\ddot{\Gamma}(1)-\dot{\Gamma}^2(1)\right)+\left(\dddot{\Gamma}(1)-\dot{\Gamma}(1)\ddot{\Gamma}(1)\right)\right)$$
$$+c^2 8\theta(\log^2\tilde{f})\left(\ddot{\Gamma}(1)-\dot{\Gamma}^2(1)\right)+8\theta(\log\tilde{f})\left(\dddot{\Gamma}(1)-\dot{\Gamma}(1)\ddot{\Gamma}(1)\right)$$
$$+2\left(\ddddot{\Gamma}(1)-\ddot{\Gamma}^2(1)\right).$$

Finally, setting $a = 1$ and $c = 0$ yields the asymptotic variance V_{11} of the first log moment, while $a = 0$ and $c = 1$ yield V_{22}, the second log moment. If we set $a = 1 = c$, then we should subtract off $V_{11} + V_{22}$ from the resulting quantity, which yields $2V_{12}$. In this way the asymptotic covariance matrix V is obtained, and the log moments are asymptotically normal with the indicated mean and covariance matrix V. □

Proof of Corollary 1. We first observe that

$$\tilde{\theta}(\log^2\widehat{f})-\tilde{\theta}^2(\log\widehat{f})-\theta(\log^2\tilde{f})+\theta^2(\log\tilde{f})-\ddot{\Gamma}(1)+\dot{\Gamma}^2(1)$$
$$=\left(\tilde{\theta}(\log^2\widehat{f})-\theta(\log^2\tilde{f})-2\dot{\Gamma}(1)\theta(\log\tilde{f})-\ddot{\Gamma}(1)\right)$$
$$-\left(\tilde{\theta}(\log\widehat{f})-\theta(\log\tilde{f})-\dot{\Gamma}(1)\right)\left(\tilde{\theta}(\log\widehat{f})+\theta(\log\tilde{f})+\dot{\Gamma}(1)\right).$$

Now since $\tilde{\theta}(\log\widehat{f}) \xrightarrow{P} \theta(\log\tilde{f}/b)+\dot{\Gamma}(1)$, we use Theorem 1 to deduce that

$$\sqrt{n}\left(\tilde{\psi}(\log\widehat{f})-\psi(\log\tilde{f})-\ddot{\Gamma}(1)+\dot{\Gamma}^2(1)\right)\stackrel{\mathcal{L}}{\Longrightarrow} G_2-2\left(\theta(\log\tilde{f})+\dot{\Gamma}(1)\right)G_1,$$

where $[G_1, G_2]'$ is bivariate normal with covariance matrix V from Theorem 1. Hence the limiting variance W is given by

$$W=8\left(\ddot{\Gamma}(1)-\dot{\Gamma}^2(1)\right)\theta(\log^2\widetilde{f})+8\left(\dddot{\Gamma}(1)-\dot{\Gamma}(1)\ddot{\Gamma}(1)\right)\theta(\log\widetilde{f})$$
$$+2\left(\ddddot{\Gamma}(1)-\ddot{\Gamma}^2(1)\right)-8\left(\dot{\Gamma}(1)+\theta(\log\widetilde{f})\right)\left(2(\ddot{\Gamma}(1)-\dot{\Gamma}^2(1))\theta(\log\widetilde{f})\right.$$
$$\left.+(\dddot{\Gamma}(1)-\dot{\Gamma}(1)\ddot{\Gamma}(1))\right)+8\left(\dot{\Gamma}(1)+\theta(\log\widetilde{f})\right)^2(\ddot{\Gamma}(1)-\dot{\Gamma}^2(1)).$$

This simplified to the stated value for W. □

References

[1] J. Beran, Statistics for Long Memory Processes, Chapman and Hall, New York, 1994.
[2] P. Bloomfield, An exponential model for the spectrum of a scalar time series, Biometrika 60 (1973) 217–226.
[3] P. Brockwell, R. Davis, Time Series: Theory and Methods, Springer-Verlag, New York, 1991.
[4] W. Chen, R. Deo, A generalized portmanteau goodness-of-fit test for time series models, Econometric Theory 20 (5) (2004) 382–416.
[5] S. Chiu, Weighted least squares estimators on the frequency domain for the parameters of a time series, The Annals of Statistics 16 (1988) 1315–1326.
[6] R. Deo, W. Chen, On the integral of the squared periodogram, Stochastic Processes and Their Applications 85 (2000) 159–176.
[7] K. Drouiche, A test for spectrum flatness, Journal of Time Series Analysis 28 (6) (2007) 793–806.
[8] V. Gómez, A. Maravall, Automatic modeling methods for univariate time series, in: D. Pẽna, G. Tiao, R. Tsay (Eds.), A Course in Time Series, John Wiley & Sons, New York, 2001, pp. 171–201.
[9] W. Li, Diagnostic Checks in Time Series, CRC Press, 2004.
[10] G. Ljung, G. Box, On a measure of lack of fit in time series models, Biometrika 65 (1978) 297–303.

[11] K. Miller, M. Rochwarger, Estimation of spectral moments of time series, Biometrika 57 (1970) 513–517.
[12] E. Paparoditis, Spectral density based goodness-of-fit tests for time series models, Scandinavian Journal of Statistics 27 (2000) 143–176.
[13] D. Peña, J. Rodríguez, A powerful portmanteau test of lack of fit for time series, Journal of the American Statistical Association 97 (2002) 601–610.
[14] D. Peña, J. Rodríguez, Descriptive measures of multivariate scatter and linear dependence, Journal of Multivariate Analysis 85 (2003) 361–374.
[15] D. Peña, J. Rodríguez, The log of the determinant of the autocorrelation matrix for testing goodness of fit in time series, Journal of Statistical Planning and Inference 136 (2006) 2706–2718.
[16] M. Taniguchi, Y. Kakizawa, Asymptotic Theory of Statistical Inference for Time Series, Springer-Verlag, New York City, New York, 2000.
[17] US Census Bureau, X-12 ARIMA Reference Manual, Version 0.2.10, Washington, DC, 2002.