Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **481,** 4311–4322 (2018)
Advance Access publication 2018 September 10

doi:10.1093/mnras/sty2487

An irregular discrete time series model to identify residuals with autocorrelation in astronomical light curves

Susana Eyheramendy,[1,2,3]★ Felipe Elorrieta[1,2] and Wilfredo Palma[1,2]

[1]*Department of Statistics, Faculty of Mathematics, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, 7820436 Macul, Santiago, Chile*
[2]*Millennium Institute of Astrophysics, Santiago, Chile*
[3]*Max-Planck-Institut für Astronomie, Konigstuhl 17, 69117 Heidelberg, Germany*

Accepted 2018 September 5. Received 2018 September 2; in original form 2018 May 24

ABSTRACT

Time series observations are ubiquitous in astronomy and are generated, for example, to distinguish between different types of supernovae to detect and characterize extrasolar planets and to classify variable stars. These time series are usually modelled using a parametric and/or physical model that assumes independent and homoscedastic errors, but in many cases, these assumptions are not accurate and there remains a temporal dependence structure on the errors. This can occur, for example, when the proposed model cannot explain all the variability of the data or when the parameters of the model are not properly estimated. In this work, we define an autoregressive model for irregular discrete-time series based on the discrete time representation of the continuous autoregressive model of order 1. We show that the model is ergodic and stationary. We further propose a maximum likelihood estimation procedure and assess the finite sample performance by Monte Carlo simulations. We implement the model on real and simulated data from Gaussian as well as other distributions, showing that the model can flexibly adapt to different data distributions. We apply the irregular autoregressive model to the residuals of a transit of an extrasolar planet to illustrate errors that remain with temporal structure. We also apply this model to residuals of an harmonic fit of light curves from variable stars to illustrate how the model can be used to detect incorrect parameter estimation.

Key words: methods: data analysis – methods: statistical – planets and satellites: general – stars: variables: general.

1 INTRODUCTION

An irregular time series is a sequence of observational times and values (t_n, y_n) such that the series $t_1, \ldots, t_N$ is strictly increasing, the distance between consecutive times, $t_j - t_{j-1}$ in general differs, and $y_1, \ldots, y_N$ is a sequence of real numbers. Irregular time series are commonly observed in many disciplines. For example, natural disasters, such as earthquakes, floods, or volcano eruptions, occur with different time gaps. In the health science, patients can be observed irregularly in time, and in astronomy, observations are usually obtained at irregular time gaps due to, for example, its dependence on clear skies to be able to get observational data from optical telescopes.

The analysis and modelling of time series are common and there exists a vast amount of theory and methods, most of which assume equally spaced measurements (e.g. Brockwell & Davis 1991, 2016; Box et al. 2015). In practice, often the analysis of irregularly spaced data is performed by ignoring the irregularity of the times and assuming regular spaced data. This practice can introduce bias in the parameter estimation leading to inaccurate predictions. Another common practice is to transform the irregular time series into a regular time series by performing interpolation, usually linear, and then apply methodology for equally spaced data [see e.g. Adorf (1995) for a survey of such methods in the context of astronomical data]. This again can introduce significant bias in the parameter estimation, specially when the time gap differences vary a lot [see e.g. Eckner (2014) for more details].

Some efforts have been made in trying to develop models for irregular time series. For example, Erdogan et al. (2005) introduces two models, one that assumes stationarity that can be seen as an extension of the classical autoregressive model of order one [AR(1)], while the second model does not assume stationarity, allowing some flexibility. Eckner (2014) attempts to develop a general framework for modelling irregular time series consistent with existing methods on equally spaced time series, but does not consider model specification and estimation. Other authors have suggested to embed irregular time series into continuous diffusion processes (e.g. Jones 1985) and use the Kalman filter to estimate the parameters and to carry out predictions (e.g. Parzen 1984; Belcher, Hampton & Wilson 1994).

In astronomy, considerable effort has been put in the estimation of spectrum of an irregular time series (e.g. Lomb 1976; Scargle 1982;

★ E-mail: susana@mat.uc.cl

© 2018 The Author(s)
Published by Oxford University Press on behalf of the Royal Astronomical Society

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Catï¿½lica de Chile user on 08 October 2018

Thiebaut & Roques 2005) and some effort in the modelling. For example, Tuomi et al. (2013) developed a first-order autoregressive model, a first order moving average model and a general ARMA model, but these models do not have desirable statistical properties, as they are neither stationary[1] nor ergodic.[2] Bailer-Jones (2011) developed Bayesian models for terrestrial impact cratering to assess periodicity on the impact ages, and generalize the models for different types of data in Bailer-Jones (2012). Kelly et al. (2014) follows a different approach by proposing to use continuous-time autoregressive moving average (CARMA) models to fit irregular time series.

Other methods have been developed that attempt to estimate the autocorrelation of a time series, which in general do not depend on a model but estimate the autocorrelation directly from the data (for a review of such methods, see e.g. Rehfeld et al. 2011). But in general, for fitting light curves for example, there are two main approaches followed by astronomers that account for irregular spaced time series. One is to use Monte Carlo simulations to forward model the periodogram as a function of a model power spectrum, and the other approach is to fit the light curves in the time domain fitting usually Gaussian processes (e.g. Rasmussen 2006; Foreman-Mackey et al. 2017). Both general methodologies can be computationally very expensive (e.g. Done et al. 1992; Uttley, McHardy & Papadakis 2002; Kelly, Bechtold & Siemiginowska 2009; Brewer et al. 2011; Emmanoulopoulos, McHardy & Papadakis 2013; Kelly et al. 2014).

Exceptions can be found on models that can be represented as state space models, such as the CARMA(p, q) models. These models overcome the computational burden by using Kalman filter to estimate the likelihood function.

In this study, we consider the continuous autoregressive model of order 1, the so-called CAR(1) model or CARMA(1,0). Based on the discrete-time representation of this model, we define the *irregular autoregressive (IAR) model*, derive its statistical properties, and develop statistical tests to assess the significance of the parameter of the model. We further show that this discrete representation of the autoregressive model allows for Gaussian and non-Gaussian-distributed data, leading to increase flexibility.

We focus on applications of the IAR model in astronomy, but the model could be applied to any other field as well. Models for irregular time series are particularly relevant in astronomy as current and future time domain optical surveys, such as SDSS Stripe 82 Supernova Survey (Frieman et al. 2008), Palomar Transient Factory (PTF; Law et al. 2009), the Catalina Real-Time Transient Survey (CRTS; Drake et al. 2009), Pan-STARRS (Kaiser et al. 2002), and the Large Synoptic Survey Telescope (LSST; Ivezic et al. 2008), will provide a huge amount of data in the form of irregular time series.

In this article, the models and its properties are shown in Section 2. In Section 2.1, the CAR(1) model is described, while in Section 2.2 the IAR model is defined and its statistical properties derived. We assess the finite sample performance of the maximum likelihood estimator of the parameter of the IAR model via Monte Carlo simulations and show the results in Section 3. We compare the performance of the IAR model with the regular autoregressive model of order one and the ARFIMA models, and show the results in Section 4. In Section 5, we illustrate how the IAR model can fit a Gamma and a Student's t-distributed sequence. Further, in order to illustrate some possible uses of this model in astronomy, we implement the IAR model in the context of two astronomical data set (Section 7). We implement the model on light curves of variable stars obtained from the OGLE (Udalski et al. 1999) and *Hipparcos* (Perryman et al. 1997) surveys (Section 7.1) and on a light curve from a star with a transiting exoplanet (Section 7.3). We develop statistical tests to assess the significance of the single parameter of the model, which allows to check whether there remain significant autocorrelation on the time series. We develop an algorithm for maximum likelihood estimation. We implement code in the R statistical software and PYTHON to estimate the model and to perform the statistical test that assess significance. We end this paper with a discussion in Section 8.

[1]A stationary process is a stochastic process whose unconditional joint probability distribution does not change when shifted in time.

[2]A stochastic process is said to be ergodic if its statistical properties can be deduced from a single, sufficiently long, random sample of the process.

2 TIME SERIES MODELS AND THEIR PROPERTIES

We consider astronomical time series that can be fit using a parametric model that is represented as

$$z_t = g(t, \theta) + \delta_t, \tag{1}$$

where z_t is the astronomical observation at time t, $g(t, \theta)$ is the mean of the model at time t, that depends on the vector of parameters θ, and δ_t is the error of the model at time t.

For example, in fitting light curves of periodic variable stars, the usual approach is to use an harmonic model where

$$g(t, \theta) = \alpha + \beta t + \sum_{j=1}^{4} (a_j \sin(2\pi f_1 j t) + b_j \cos(2\pi f_1 j t)) \tag{2}$$

and $\theta = (\alpha, \beta, f_1, a_1, \ldots, a_4, b_1, \ldots, b_4)$. In this case, z_t represents the flux measurement of the variable star at time t, α and β are the parameters of a linear trend, $1/f_1$ is the period of the star, $\{a_j\}$ and $\{b_j\}$ are the parameters of the harmonic model. For transient or variable phenomena, such as supernovae or planets, $g(t, \theta)$ is fit using a deterministic statistical or astrophysical model.

These errors (i.e. $\{\delta_t\}$) are usually assumed independent with a Gaussian distribution with mean zero and variance σ^2. In many cases, neither the independence of the errors nor the homoscedasticity (or equal variance) of the errors is achieved. To identify and overcome these problems, the continuous autoregressive model (CAR), for example, can be implemented on $z_t - g(t, \hat{\theta})$, i.e. the residuals of the model in equation (1), in order to assess whether a correlation structure remains after fitting such model.

In the following two sections, we describe the CAR(1) model and define the IAR model. These models would typically be used to identify autocorrelation in the residuals.

2.1 Continuous autoregressive model of order 1

The continuous autoregressive model of order 1 [CAR(1)] attempts to solve a stochastic differential equation of order one, driven by white noise. White noise is the name used in time series analysis for an independent series of random variables (when the data is assumed to be Gaussian). The problem is that continuous time white noise exists only in the sense that its integral is a continuous time random walk, commonly referred as Brownian motion or Wiener process. A continuous time random walk is the limit of a discrete time random walk as the time interval gets small. The path function of a Wiener process can be simulated, and will be continuous with a very wiggly appearance and its derivative does not exist. Moreover, a finite segment of this curve has infinite path length. Despite all

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Catï¿½lica de Chile user on 08 October 2018

these undesirable properties the Wiener process is still the key to get random input into a continuous time process (Jones 1993).

The mathematical formulation of the process $\varepsilon(t)$ corresponding to a CAR(1) model is

$$\frac{\mathrm{d}}{\mathrm{dt}}\epsilon(t) + \alpha_0\epsilon(t) = \sigma_0\nu(t) + \beta, \tag{3}$$

where $\nu(t)$ is the continuous time white noise, and α_0 and β are unknown parameters of the model. It can be shown that the process $\varepsilon(t_k)$ that is a solution of equation (3) is also a solution of the difference equation given by

$$\epsilon(t) - \frac{\beta}{\alpha_0} = \mathrm{e}^{-\alpha_0(t-s)}(\epsilon(s) - \frac{\beta}{\alpha_0}) + \mathrm{e}^{-\alpha_0 t}(I(t) - I(s)), \tag{4}$$

where $I(t) = \sigma_0 \int_0^t \mathrm{e}^{\alpha_0 u}\mathrm{d}w(u)$ is an Itô integral[3] (Brockwell & Davis 2016). See Appendix A for a full derivation of this result. Based on the last equation, we define the discrete time series model for irregularly sampled observations and derive its statistical properties (shown on the following sections).

2.2 IAR model

Denote y_{t_j} an observation measured at time t_j, and consider an increasing sequence of observational times $\{t_j\}$ for $j = 1, \ldots, n$. We define the IAR process by

$$y_{t_j} = \phi^{t_j - t_{j-1}}\, y_{t_{j-1}} + \sigma\, \sqrt{1 - \phi^{2(t_j - t_{j-1})}}\, \varepsilon_{t_j}, \tag{5}$$

where ε_{t_j} are independent random variables with zero mean and unit variance. Note that by replacing $\mathrm{e}^{-\alpha_0}$ with ϕ (and setting $\beta = 0$) in equation (4), we get to the Gaussian IAR model, because the CAR(1) model assumes Gaussian data. The connection between equations (4) and (5) is completed by defining $\sigma^2 = \frac{\sigma_0^2}{2\alpha_0}$.

Importantly, the model described by equation (5) can also be established without assuming Gaussian errors. From now on, we do not assume Gaussian data to derive the statistical properties of the model unless we explicitly mention a distribution assumption.

Observe that

$$E(y_{t_j}) = 0 \text{ and } \mathrm{Var}(y_{t_j}) = \sigma^2 \text{ for all } y_{t_j}, \tag{6}$$

and the covariance between y_{t_k} and y_{t_j} is $E(y_{t_k}\, y_{t_j}) = \sigma^2\, \phi^{t_k - t_j}$ for $k \geq j$.

Thus, for any two observational times $s < t$, we can define the autocovariance function as

$$\gamma(t-s) = E(y_t\, y_s) = \sigma^2\, \phi^{t-s}, \tag{7}$$

as well as the autocorrelation function (ACF), $\rho(t-s) = \frac{\gamma(t-s)}{\gamma(0)} = \phi^{t-s}$.

Given the results above, the sequence $\{y_{t_j}\}$ corresponds to a second-order or weakly stationary[4] process. We show in the next theorem that, in addition, under some conditions the process is stationary and ergodic.

Theorem 1: Consider the process defined by equation (5) and assume that the input noise is an i.i.d. sequence of random variables with zero mean and unit variance. Furthermore, suppose that $t_j - t_{j-n} \geq C\log n$ as $n \to \infty$, $0 < \phi < 1$, where C is a positive constant that satisfies $C\log\phi^2 < -1$. Then, there exists a solution to the process defined by equation (5), and the sequence $\{y_{t_j}\}$ is stationary and ergodic. See Appendix B for a proof of this theorem.

Note that, if $t_j - t_{j-1} = 1$ for all j, then equation (5) becomes

$$y_{t_j} = \phi\, y_{t_{j-1}} + \sigma\, \sqrt{1 - \phi^2}\, \varepsilon_{t_j} \quad \text{for } j = 2, \ldots, n, \tag{8}$$

which corresponds to the autoregressive model of order 1 [AR(1)] for regularly space data. Therefore, the IAR model is an extension of the regular autoregressive model. As mentioned previously, it is also an extension of the continuous autoregressive model of orden 1.

Note also that for the regular AR(1) model the two assumptions on the theorem are satisfied: $t_j - t_{j-n} = n$, $n > \log(n)$ is achieved, and $\phi^2 < 1$ is part of the assumptions of the regular autoregressive model. Therefore, the AR(1) is ergodic and stationary.

Corollary: Let $\bar{y}_n = \frac{1}{n}\sum_{j=1}^n y_{t_j}$ and $\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{j=1}^n (y_{t_j} - \bar{y}_n)^2$ be the sample mean and the sample variance of the IAR process, respectively. Then, we have that $\bar{y}_n \to E(y_{t_j})$ and $\tilde{\sigma}_n^2 \to \sigma^2$, in probability, as $n \to \infty$.

2.3 Estimation

The likelihood of the data $\{y_{t_1}, \ldots, y_{t_n}\}$ can be expressed as

$$f(y_{t_1}, \ldots, y_{t_n}; \theta) = f(y_{t_1}; \theta) f(y_{t_2}|y_{t_1}; \theta) \times \ldots \times f(y_{t_n}|y_{t_{n-1}}; \theta), \tag{9}$$

where $\theta = (\sigma^2, \phi)$ is the parameter vector of the model. To describe clearly the estimation process, we assume here that the marginal and conditional distributions of the time series are Gaussian. Note that this assumption is not necessary to obtain the statistical properties stated in Theorem 1. In Section 5, we show an example where the conditional distribution is assumed to be Gamma, and in Section 6, we show an example where the conditional distribution is assumed to be a Student's t-distribution.

Assume that

$$f(y_{t_1}; \sigma^2, \phi) \sim N(0, \sigma^2) \text{ and} \tag{10}$$

$$f(y_{t_j}|y_{t_{j-1}}; \sigma^2, \phi) \sim N(\phi^{t_j - t_{j-1}}\, y_{t_{j-1}}, \sigma^2\,(1 - \phi^{2(t_j - t_{j-1})}) \tag{11}$$

for $j = 2, \ldots, n$. Based on equation (5), minus the log-likelihood of this process can be written as

$$\ell(\theta) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^n \log \nu_{t_j} + \frac{1}{2}\sum_{j=1}^n \frac{e_{t_j}^2}{\nu_{t_j}}, \tag{12}$$

where we define $e_{t_1} = y_{t_1}$, $e_{t_j} = y_{t_j} - \phi^{t_j - t_{j-1}}\, y_{t_{j-1}}$ for $j > 1$ and their variances as $\nu_{t_j} = \mathrm{Var}(e_{t_j})$.

Observe that the finite past predictor of the process at time t_j is given by

$$\widehat{y}_{t_1} = 0, \text{ and } \widehat{y}_{t_j} = \phi^{t_j - t_{j-1}}\, y_{t_{j-1}}, \text{ for } j = 2, \ldots, n. \tag{13}$$

Therefore, $e_{t_j} = y_{t_j} - \widehat{y}_{t_j}$ is the prediction error with variance $\nu_{t_1} = \mathrm{Var}(e_{t_1}) = \sigma^2$,

$$\nu_{t_j} = P\mathrm{Var}(e_{t_j}) = \sigma^2[1 - \phi^{2(t_j - t_{j-1})}], \text{ for } j = 2, \ldots, n. \tag{14}$$

By direct maximization of the log-likelihood (12), we can obtain the maximum likelihood estimator of σ^2,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^n \frac{(y_{t_j} - \widehat{y}_{t_j})^2}{\tau_{t_j}}, \text{ where } \tau_{t_j} = \nu_{t_j}/\sigma^2. \tag{15}$$

But it is not possible to find $\widehat{\phi}$, the maximum likelihood estimator of ϕ, by direct maximization of the likelihood, but iterative methods

[3]The integral is an extension of the Riemann–Stieltjes integral, where the integrands and the integrators are now stochastic processes.

[4]A weakly stationary process is a random sequence of random variables that requires that the first moment (i.e. the mean) and the autocovariance do not vary with respect to time.

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Católica de Chile user on 08 October 2018

Table 1. Maximum likelihood estimation of simulated IAR series with mixture of Exponential distribution for the observational times, with $\lambda_1 = 130$, $\lambda_2 = 6.5$, $w_1 = 0.15$, and $w_2 = 0.85$.

Case	n	ϕ	$\widehat{\phi}$	SD($\widehat{\phi}$)	$\sigma(\widehat{\phi})$	$\widehat{\sigma}$
1	50	0.900	0.887	0.044	0.034	1.013
2	50	0.990	0.985	0.008	0.008	1.039
3	50	0.999	0.996	0.004	0.003	1.155
4	100	0.900	0.894	0.029	0.024	1.005
5	100	0.990	0.988	0.005	0.006	1.015
6	100	0.999	0.998	0.002	0.002	1.049

Table 2. Maximum likelihood estimation of simulated IAR series of size n, with exponential distribution mix observation times, $\lambda_1 = 300$ and $\lambda_2 = 10$.

Case	n	ϕ	$\widehat{\phi}$	SD($\widehat{\phi}$)	$\sigma(\widehat{\phi})$	$\widehat{\sigma}$
1	40	0.900	0.8843	0.058	0.038	1.011
2	40	0.990	0.9854	0.009	0.007	1.037
3	40	0.999	0.9969	0.003	0.002	1.120
4	80	0.900	0.8929	0.034	0.027	1.006
5	80	0.990	0.9876	0.005	0.005	1.018
6	80	0.999	0.9980	0.001	0.002	1.046

can be used (for details, see Chapter 5 of Palma 2016). We developed scripts in the statistical language/software R, and also in PYTHON, to estimate ϕ.

Lemma 1: Consider the process defined by equation (5) and suppose that $t_j - t_{j-n} = h\,n$, for a positive constant h, $0 < \phi < 1$. Let $\widehat{\phi}_n$ be the maximum likelihood estimator of ϕ. Then, the maximum likelihood estimator (MLE) satisfies the following asymptotic normal distribution:

$$\sqrt{n}\,(\widehat{\phi}_n - \phi) \to N(0, \sigma_\phi^2), \tag{16}$$

as $n \to \infty$, where

$$\sigma_\phi^2 = \frac{1 - \phi^{2h}}{h^2 \phi^{2h-2}}. \tag{17}$$

See Appendix C for a proof of this lemma.

Similar to the continuous time autoregressive models, the IAR can be represented using state-space models from which the Kalman filter (Kalman 1960) can be implemented allowing fast and scalable estimation of parameters.

3 SIMULATION STUDY TO ASSESS THE MAXIMUM LIKELIHOOD ESTIMATORS OF THE IAR MODEL

This section shows the results of Monte Carlo experiments assessing the finite sample performance of the proposed maximum likelihood estimator.

The simulated processes correspond to the model (5) where the observational times follow a mixture of two exponential distributions with means $1/\lambda_1$ and $1/\lambda_2$, respectively, and random weights w_1 and w_2, respectively. We find that this choice for the observational times corresponds to a reasonable representation for the observational times of a multiyear large time series survey such as the Vista Variable of the Via Lactea (Minniti et al. 2010). Table 1 shows a summary of the simulations based on 1000 repetitions with $\lambda_1 = 130$, $\lambda_2 = 6.5$, $w_1 = 0.15$, and $w_2 = 0.85$. Table 2 shows a summary of the simulations based on 1000 repetitions with $\lambda_1 = 300$, $\lambda_2 = 10$, $w_1 = 0.15$, and $w_2 = 0.85$.

Table 3. Maximum likelihood estimation of simulated IAR series with quasi-periodic behaviour in the observational times.

Case	n	ϕ	$\widehat{\phi}$	SD($\widehat{\phi}$)	$\widehat{\sigma}$
1	60	0.900	0.887	0.039	1.011
2	60	0.990	0.985	0.008	1.022
3	60	0.999	0.996	0.003	1.108
4	100	0.900	0.890	0.032	1.009
5	100	0.990	0.986	0.008	1.013
6	100	0.999	0.996	0.003	1.076

The Monte Carlo simulations suggest that the finite-sample performance of the proposed methodology is accurate. In particular, the estimation bias is small even for the smaller sample sizes used in Tables 1 and 2. Note that we restrict to high values of the parameter ϕ. The reason for this is the choice of the distribution of the observational time gaps, which tend to be large. Observe that an approximate asymptotic estimation of the standard deviation $\sigma(\widehat{\phi})$ obtained by an application of Lemma 1 is also provided in these tables. Notice that the approximation seems to work well for larger sample sizes (e.g. 80 or 100) and high values of ϕ (e.g. 0.9900 or 0.9990).

To assess whether the observational time distribution has any effect on the parameter estimation, we perform another Monte Carlo experiment using a quasi-periodic distribution. To generate a sample of size n of these times, we use the following scheme. First, we assume a year of 365 d and then we randomly select $\frac{n}{10}$ observations from the uniform distribution $U(180, 210)$ and another $\frac{n}{10}$ observations from the uniform distribution $U(240, 270)$. This is repeated for five consecutive years. In this way, we obtain n observational times at two fixed months a year (June and August), but on randomly picked days within the month.

The finite sample performance is assessed by a simulation experiment based on 1000 repetitions of the IAR process of sizes $n = 60$ and $n = 100$. The observational times are generated using the procedure mentioned above. Comparing the results in Tables 1 and 3, we can conclude that the accuracy of the proposed estimation method is not affected by a quasi-periodic sample of the observational times.

4 SIMULATION STUDY TO COMPARE THE IAR MODEL WITH OTHER MODELS FOR REGULAR TIME SERIES

We compare the IAR model with other standard models for regular time series. Fig. 1 shows the standard deviation of the prediction errors, i.e. the root of the series $\hat{v}_t$ in equation (14). Note that because in the IAR model the prediction errors e_{t_j} are unbiased, i.e. $E(e_{t_j}) = 0$, the standard deviation of the prediction errors are equivalent to the root mean squared error (RMSE).

To estimate the prediction errors, we generate the sequence $\{y_1, \ldots, y_n\}$ using the IAR model with $\phi = 0.99$ and $n = 100$. The red line corresponds to the standard deviation of the sequence, the blue and green lines correspond to the standard deviation of the regular autoregressive model of order one [AR(1)] and ARFIMA(1,d,0), respectively. These models assume regular spaced data. The observational times are generated using the density $f(t|p, \lambda_1, \lambda_2) = p\mathcal{E}(t|\lambda_1) + (1-p)\mathcal{E}(t|\lambda_2)$ with $p = 0.15$, $\lambda_1 = 130$, $\lambda_2 = 6.5$, where $\mathcal{E}(t|\lambda_1)$ denotes an exponential distribution with parameter λ_1.

Observe that the only model that changes the standard deviation at each observational time is the IAR model, corresponding to the

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Catï¿½lica de Chile user on 08 October 2018

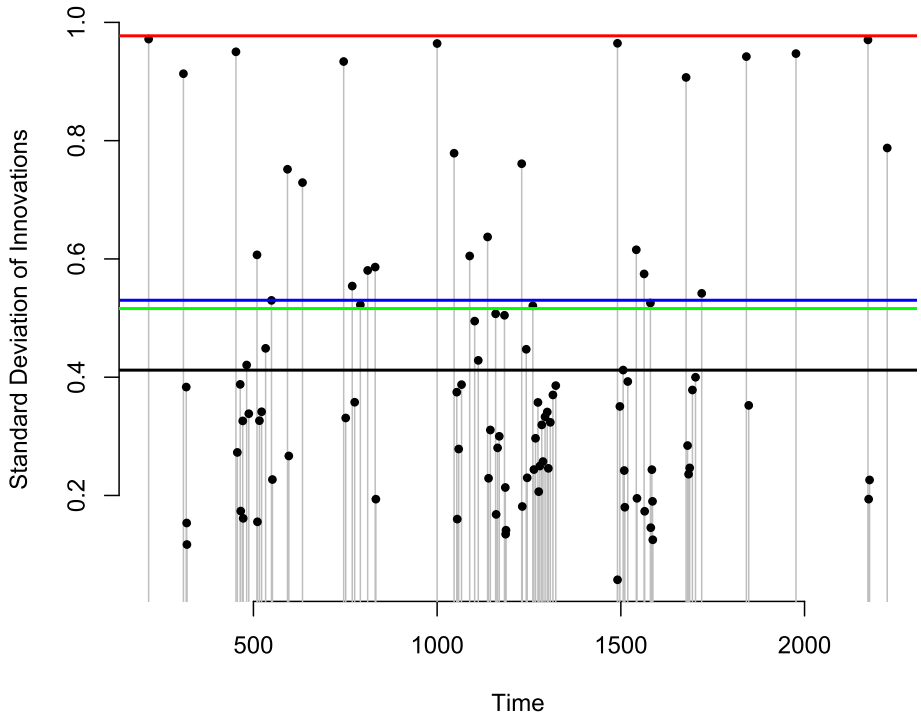


Figure 1. Comparison of standard deviation at each time of a sequence simulated with the IAR model with parameter $\phi = 0.99$ and length 100. The red line corresponds to the standard deviation of the sequence, the blue and green lines correspond to the standard deviation estimated assuming an AR(1) and ARFIMA(1,d,0) models, respectively. The black line corresponds to the average standard deviation of the IAR model, where the black dots are the individual standard deviations at each time.

black dots in Fig. 1, where larger values close to one are observed after a larger observational time gap. The average standard deviation of the IAR model is shown as the black line, and it is smaller than the standard deviation of any of the other models.

In the next two sections, we show simulation studies to describe how the IAR model can be implemented to fit Gamma and Student's *t*-distributed series and compare its performance with the continuous autoregressive model.

5 IMPLEMENTATION OF THE IAR MODEL ON SIMULATED GAMMA-DISTRIBUTED SERIES

We implement the IAR model on simulated conditional Gamma distributions following the procedure described at Palma & Zevallos (2011). Specifically, the conditional mean and variance of the IAR model are defined as

$$\begin{aligned}\mathbb{E}(y_{t_j}|y_{t_{j-1}}) &= \mu + \phi^{t_j - t_{j-1}}\, y_{t_{j-1}}\\ \mathbb{V}(y_{t_j}|y_{t_{j-1}}) &= \sigma^2\,(1-\phi^{2(t_j-t_{j-1})}).\end{aligned} \tag{18}$$

These moments are equivalent to the ones for the Gaussian case stated in equation (11), the only difference is the positive parameter μ that corresponds to the expected value of $y_{t_{j-1}}$. If $y_{t_j}|y_{t_{j-1}}$ follows a Gamma distribution, a positive value of μ is required in order to ensure that the process is positive. However, the process may be shifted, so that $y_{t_j} - \mu$ have zero mean, like the Gaussian IAR. For simplicity, we set $\mu = 1$.

In addition, note that under the assumption of stochastic times the marginal mean $\mathbb{E}(y_{t_j}) = \frac{\mu}{1-\mathbb{E}(\phi^{t_j-t_{j-1}})}$ and marginal variance $\mathbb{V}(y_{t_j}) = \sigma^2 + \frac{\mathbb{E}(y_{t_j})^2\mathbb{V}(\phi^{t_j-t_{j-1}})}{1-\mathbb{E}(\phi^{2(t_j-t_{j-1})})}$ are constants.

If $x_{t_j} \sim \text{Gamma}(\alpha_{t_j}, \beta_{t_j})$ follows a Gamma distribution with shape α_{t_j} and scale β_{t_j} parameters, it is well known that the expected value and the variance of x_{t_j} are $\mathbb{E}(x_{t_j}) = \alpha_{t_j}\,\beta_{t_j}$ and $\mathbb{V}(x_{t_j}) = \mathbb{E}(x_{t_j})\,\beta_{t_j}$, respectively. From the two equations,

$$\begin{aligned}\alpha_{t_j}\,\beta_{t_j} &= \mu + \phi^{t_j-t_{j-1}}\, y_{t_{j-1}},\\ \alpha_{t_j}\,\beta_{t_j}^2 &= \sigma^2\,(1-\phi^{2(t_j-t_{j-1})}),\end{aligned}$$

we obtain α_{t_j} and β_{t_j} as functions of the parameters ϕ and σ^2: $\alpha_{t_j} = \alpha_{t_j}(\phi, \sigma^2)$ and $\beta_{t_j} = \beta_{t_j}(\phi, \sigma^2)$. Thus, the log-likelihood of the conditional distribution of $y_{t_j}|y_{t_{j-1}}$ can be written as

$$\begin{aligned}\ell_j &= \log f_\theta\left(\phi, \sigma^2\right)\\ &= -\left(\alpha_{t_j}\right)\log\beta_{t_j} - \log\Gamma\left(\alpha_{t_j}\right) - \frac{1}{\beta_{t_j}}y_{t_j} + \left(\alpha_{t_j}-1\right)\log y_{t_j}.\end{aligned}$$

Here, we omit the dependence of α_{t_j} and β_{t_j} on ϕ and σ^2 to keep notation clear. If $y_{t_1} \sim \text{Gamma}(1, 1)$, then the full log-likelihood is,

$$\ell(\theta) = \sum_{j=2}^{N}\ell_j + \ell_1$$

where $\ell_1 = -y_{t_1}$. The unknown parameters of the model are ϕ and σ that can be estimated using iterative methods.

We perform Monte Carlo experiments, based in 1000 repetitions, and we assess the accuracy in parameter estimation on simulated conditionally Gamma-distributed time series. We implement the Gamma-distributed IAR model as well as the Gaussian-distributed IAR model in the statistical software package R and PYTHON. The Gaussian-distributed IAR model [i.e. samples from a CAR(1) model] is implemented using the R package `cts` and the PYTHON script developed by Pichara et al. (2012).

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Católica de Chile user on 08 October 2018

In Table 4, $\widehat{\phi}$ corresponds to the estimator using the correct Gamma distributed and $\widehat{\phi}^C$ is the estimator using the mismatched Gaussian-distributed IAR model. The performance of the Gaussian-distributed IAR model using PYTHON and R does not vary significantly. In both cases, performance assuming the mismatched Gaussian-distributed IAR model is substantially inferior to assuming the correct Gamma-distributed IAR model.

6 IMPLEMENTATION OF THE IAR MODEL ON STUDENT'S *T*-DISTRIBUTED SERIES

Another implementation of a non-Gaussian IAR process is on a heavy-tailed distribution such as the Student's *t*-distribution. This kind of distribution are useful to address the problem of possible outliers in a time series. Following the procedure mentioned in Section 5, we implement an IAR model with a Student's *t* conditional distribution. If $x_{t_j} \sim t_\nu(\lambda_{t_j}, \tau^2_{t_j})$ follows a non-standardized Student's *t*-distribution with mean λ_{t_j}, variance $\tau^2_{t_j}$ and ν degrees of freedom, the expected value of x_{t_j} is $\mathbb{E}(x_{t_j}) = \lambda_{t_j}$ and the variance is $\mathbb{V}(x_{t_j}) = \tau^2_{t_j} \frac{\nu}{\nu-2}$. From the conditional mean and variance of IAR model (18), we obtain,

$$\lambda_{t_j} = \phi^{t_j - t_{j-1}}\, y_{t_{j-1}}$$
$$\tau^2_{t_j} = \frac{\nu-2}{\nu}\,\sigma^2\,\left(1-\phi^{2(t_j-t_{j-1})}\right).$$

Thus the log-likelihood of the conditional distribution of $y_{t_j}|y_{t_{j-1}}$ can be written as

$$\begin{aligned}\ell_j &= \log f_\nu\left(\lambda_{t_j}, \tau^2_{t_j}\right)\\ &= \log\left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\right)\\ &\quad -\frac{1}{2}\log\tau^2_{t_j} - \frac{\nu+1}{2}\log\left(1+\frac{1}{\nu}\left(\frac{y_{t_j}-\lambda_{t_j}}{\tau_{t_j}}\right)^2\right).\end{aligned}$$

Let $y_{t_1} \sim N(0,1)$, then the full log-likelihood is

$$\ell(\theta) = \sum_{j=2}^{N}\ell_j + \ell_1$$

where $\ell_1 = -\frac{1}{2}(\log(2\pi) + y^2_{t_1})$.

In order to assess the accuracy in the parameter estimation procedure, we also perform Monte Carlo experiments with 1000 repetitions. We use two different values for the degrees of freedom $\nu = 3$ and $\nu = 5$. Table 5 shows that the estimation of the parameters ϕ and σ is precise in both cases. As expected, the estimation performance of the Gaussian IAR model is similar to the one obtained with the Student's *t*-distribution model.

7 EXAMPLES OF THE IAR MODEL IN ASTRONOMICAL TIME SERIES

In this section, we illustrate two implementation of the IAR model in Astronomical time series. The first implementation is to detect model misspecification, i.e. a model with incorrectly estimated parameters or that is not sufficiently complex to describe the data at hand. The second implementation is to identify the presence of time-correlated structure in model residuals. For the model misspecification case, we use variable star light curves from the OGLE and *Hipparcos* survey, and for the time-correlation structure we use a light curve of an exoplanetary transit.

7.1 Application to variable stars from the OGLE and *Hipparcos* surveys

The harmonic model described in equation (2) is used to model light curves from variable stars. This model requires first to find the period of the variable star, which can be estimated, for example, using the generalized Lomb–Scargle periodogram (Zechmeister & Kürster 2009). Then the remaining parameters are estimated using techniques for maximizing the likelihood. For more details on the procedure of the modelling of periodic light curve, see for example Debosscher et al. (2007), Richards et al. (2011), or Elorrieta et al. (2016).

Denote the residuals after subtracting a linear trend and an harmonic model with one frequency and four components as y_t, i.e.

$$y_t = z_t - \hat{\alpha} - \hat{\beta}t - \sum_{j=1}^{4}(\hat{a}_j \sin(2\pi f_1 j t) + \hat{b}_j \cos(2\pi f_1 j t)),$$

where $\hat{a}$ represents a maximum likelihood estimator. We implement the IAR model on these residuals.

First, we show that the model can be used to identify wrongly estimated periods. We select 40 variable stars from the OGLE and *Hipparcos* surveys for which the harmonic model gives a precise fit of the light curve. In selecting these stars, we can be certain that the periods are well estimated. These variable stars are selected from a group of 250 stars that have the highest R^2 values in the harmonic fit. The multiple correlation coefficient, R^2, is a standard statistical measure for assessing goodness of fit. In order to take a representative sample of the classes and frequencies values observed in OGLE and HIPPARCOS, we binned the frequencies in five groups, and select eight light curves from each bin and try at the same time to keep the representation of the classes of the original data set. Figs 2(a)–(c) show three examples of such set of light curves and Table 6 the distribution of classes over the different frequency bins.

We apply the IAR model to the residuals of the best harmonic model, shown in equation (2). For the 40 chosen light curves, we obtain small values close to zero for the parameter ϕ, as shown in the boxplot on the left of Fig. 3. This is expected given that the model fits the light curves very well and thus the residuals are consistent with white noise. We then vary the frequency in the interval $(f_1 - 0.5f_1, f_1 + 0.5f_1)$ taking a total of 38 frequencies equally space $g_1, \ldots, g_{38}$, 19 to the right of the correct frequency f_1 and 19 to the left. After doing so we fit the harmonic model with each wrong frequency g_j taken from the interval. The residuals of the harmonic model have now temporal structure that can be captured with the IAR model, and in particular by the inferred value of ϕ. For each 'incorrect' frequency g_j, we obtain a $\hat{\phi}_j$. The second row of Fig. 2 shows the plot of the pairs $(g_j, \hat{\phi}_j)$ (with the right frequency f_1 at the centre of the plot). Note that as we move away from the correct frequency, the value of ϕ generally increases in a non-monotonic way. Fig. 3 shows in the boxplot on the right the distribution of $\hat{\phi}$ for the light curves with the incorrect frequency. This distribution is spread-out, taking in general large values away from zero, which reflects the correlation structure that remains.

Summarizing, for a given variable star and a period we fit an harmonic model and we then apply the IAR model to check whether there's any evidence of temporal structure which in this case would arise from period misspecification. If we obtain a $\hat{\phi} \neq 0$, we want to assess whether it is possible to conclude that there is significant temporal structure or not. In order to do that, we propose the following statistical test.

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Catï¿½lica de Chile user on 08 October 2018

Table 4. Implementation of the Gamma-distributed IAR model and the CAR(1) model on simulated Gamma-IAR series in R and PYTHON. For the observational times, we use a mixture of two exponential distributions with parameters $\lambda_1 = 130$, $\lambda_2 = 6.5$, $w_1 = 0.15$, and $w_2 = 0.85$.

	N	ϕ	σ	$\widehat{\phi}$	SD($\widehat{\phi}$)	$\widehat{\phi}^C$	SD($\widehat{\phi}^C$)	$\widehat{\sigma}$	SD($\widehat{\sigma}$)
R	100	0.9	1	0.899	0.014	0.418	0.306	0.984	0.170
R	100	0.99	1	0.990	0.001	0.890	0.201	0.985	0.161
R	200	0.9	1	0.899	0.010	0.355	0.286	0.993	0.122
R	200	0.99	1	0.990	0.001	0.900	0.184	0.998	0.120
PYTHON	100	0.9	1	0.899	0.013	0.449	0.318	0.990	0.169
PYTHON	100	0.99	1	0.990	0.001	0.919	0.169	0.981	0.200
PYTHON	200	0.9	1	0.899	0.010	0.393	0.299	0.985	0.127
PYTHON	200	0.99	1	0.990	0.001	0.927	0.163	0.996	0.332

Table 5. Implementation of the t-distributed IAR model and the CAR(1) model on simulated Student's t IAR series. For the observational times, we use a mixture of two Exponential distributions with parameters $\lambda_1 = 130$, $\lambda_2 = 6.5$, $w_1 = 0.15$, and $w_2 = 0.85$.

N	ϕ	ν	$\widehat{\phi}$	SD($\widehat{\phi}$)	$\widehat{\phi}^C$	SD($\widehat{\phi}^C$)	$\widehat{\sigma}$	SD($\widehat{\sigma}$)
100	0.9	3	0.895	0.025	0.884	0.068	1.010	0.231
100	0.99	3	0.988	0.005	0.983	0.045	0.979	0.360
200	0.9	3	0.898	0.016	0.889	0.054	1.003	0.163
200	0.99	3	0.989	0.003	0.987	0.005	0.991	0.258
100	0.9	5	0.896	0.028	0.892	0.037	1.010	0.225
100	0.99	5	0.989	0.005	0.986	0.005	1.017	0.395
200	0.9	5	0.897	0.018	0.895	0.023	1.006	0.157
200	0.99	5	0.989	0.003	0.988	0.003	1.007	0.274

7.1.1 Statistical test for assessing significance of the parameter ϕ

In the second row of Fig. 2, we observe the relationship between frequency of the variable stars versus the parameter ϕ of the autoregressive model. At zero in the x-axis lies the correct frequency for which we obtain the smaller ϕ value in the three examples shown. This is expected because the light curves are chosen such that the harmonic model attains an accurate fit. Note that even though the smaller ϕ is obtained at the estimated frequency f_1, this value of ϕ relative to the neighbouring values differ substantially. Note also that while the graph on the left has values of ϕ above 0.75, in the middle the values are around 0.18 and in the figure on the right all values are between 7.5×10^{-5} and 7.8×10^{-5}, with the exception of the value of $\hat{\phi}$ at its minimum in f_1 which as expected is close to zero. Therefore, just from the value of $\hat{\phi}$ it is not always possible to discriminate between a correct period with residuals without temporal dependence and an incorrect period with residuals with temporal dependence. We propose to evaluate whether the minimum $\hat{\phi}$ is significantly smaller than the remaining $\hat{\phi}$ by assuming that the $\log(\hat{\phi})$ distributes as a Gaussian. The bottom row of Fig. 2 shows the density of the $\log(\hat{\phi})$ values at the incorrect periods, and the red triangle shows the $\log(\hat{\phi})$ values at the correct period. The p-values for the three $\log(\hat{\phi})$ at the correct period are 0, 1.62×10^{-281}, 2.86×10^{-19}, respectively, indicating that they are all statistically significantly smaller that their neighbours.

7.2 Study on simulated and real multiperiodic variable stars

Several classes of variable stars can have multiperiodic stars, for example, double-mode Cepheids and double-mode RR-Lyrae. For those stars fitting an harmonic model with only one period produces errors in the model that are not independent, but correlated. Therefore, we expect that when fitting the IAR model to the residuals of this harmonic model, the estimate of the parameter ϕ will be large. We show with simulated and real data that this is indeed the case, illustrating a case where the model lacks the complexity to describe the time series at hand.

We simulate multiperiodic light curves with two periods using the harmonic model. We show an example in which the light curve is simulated using the harmonic model with two periods and four components for each period. Specifically, at time t the value simulated is $y(t) = \sum_{i=1}^{2}\sum_{j=1}^{4}(\sin(2\pi f_i jt) + \cos(2\pi f_i jt)) + \tau_t$, where τ_t is generated from a standard Gaussian distribution with mean zero and variance one and $f_1 = 1/3, f_2 = 1/12$. The observational times are simulated using a mixture of two exponential distributions, i.e. $f(t|\lambda_1, \lambda_2, w_1, w_2) = w_1 g(t|\lambda_1) + w_2 g(t|\lambda_2)$, where $\lambda_1 = 130$, $\lambda_2 = 6.5$, $w_1 = 0.15$, and $w_2 = 0.85$. In Fig. 4, we show on the top plot the residuals after fitting an harmonic model with one period. The header has the value of $\phi = 0.5447$, which is the value of ϕ estimated from these residuals. The bottom plot has the residuals after fitting the harmonic model with two periods. From this series the estimated of ϕ has decreased to a small value close to zero ($\leq$0.0001).

From the set of real light curves observed in the OGLE and *Hipparcos* surveys, we identified some multiperiodic variable stars. Fig. 5(a) shows the residuals of a double model Cepheid after fitting an harmonic model with one period, and Fig. 5(b) shows the residuals of the same variable star after fitting an harmonic model with two periods. The $\hat{\phi}$ of the IAR model at the residuals after fitting an harmonic model with one period is 0.5411, while $\hat{\phi} = 0.033$ at the residuals after fitting an harmonic model with two periods.

7.3 Exoplanet transit light curve

A planet orbiting a star will block part of the signal if it transits in front of it as seen from our vantage point. The observed flux can then be modelled by multiplying the approximately constant flux of the star with the transit signal, which can be modelled with the formalism described in Mandel & Agol (2002). We have again a

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Catï¿½lica de Chile user on 08 October 2018

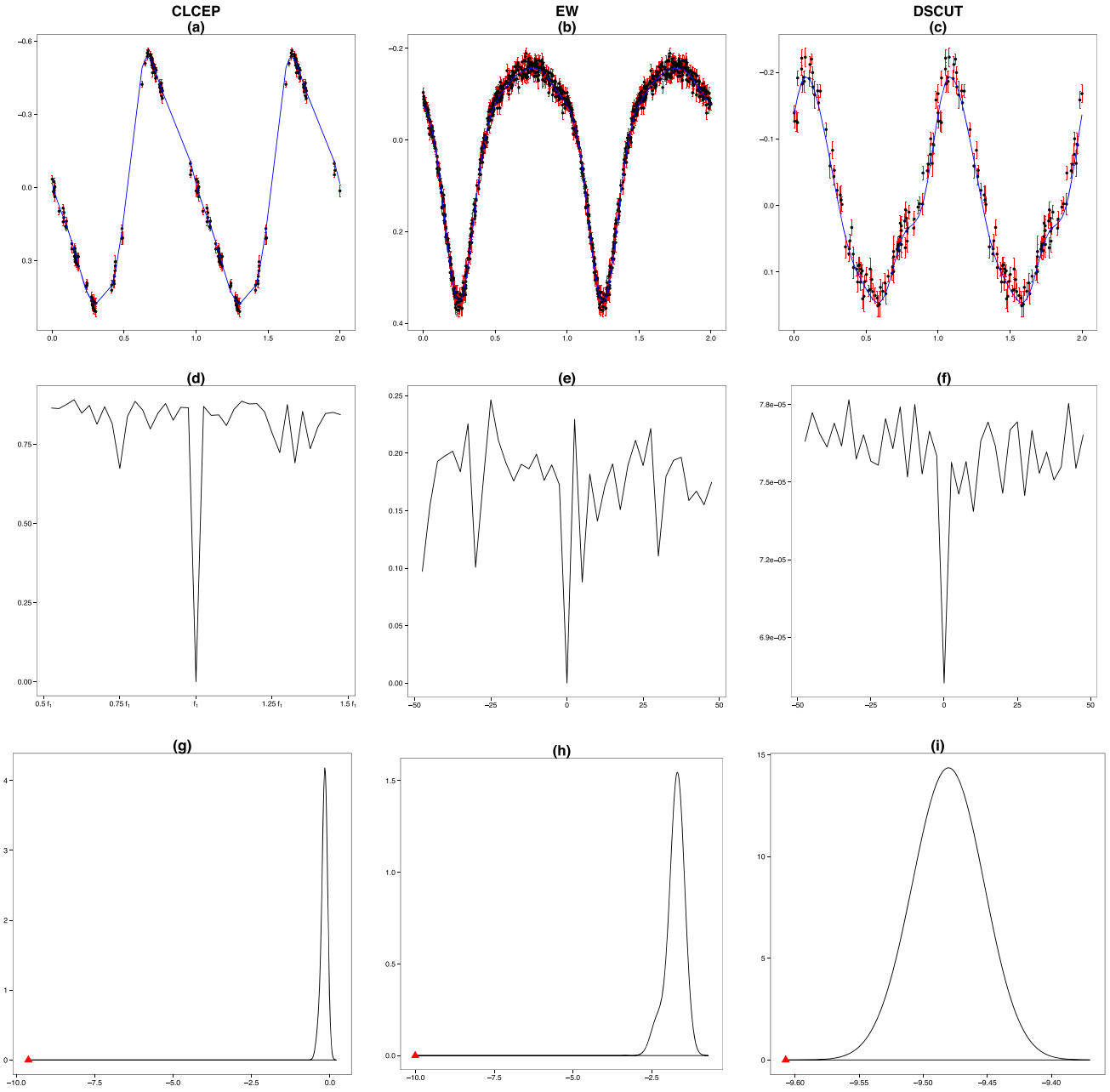


Figure 2. In the first row, the light curves of a Classical Cepheid, EW, and DSCUT are shown in Figs (a)–(c), respectively. The continuous blue line is the harmonic best fit. On the second row (Figs d–f), for each of the variable stars, it is depicted on the x-axis the per cent of variation from the correct frequency, and on the y-axis is the estimate of the parameter ϕ of the IAR model obtained after fitting an harmonic model with the wrong period (except at zero that corresponds to the right period). On the third row (Figs g–i), the distribution of the parameter ϕ of the IAR model is shown when each light curves is fit with the wrong period. The red triangle corresponds to the value of ϕ when the correct period is used in the harmonic model fitting the light curves.

structure for the model described by $z(t) = g(t, \theta) + \varepsilon(t)$, where $z(t)$ represents in this case the logarithm of the measurement flux of the star, $g(t, \theta)$ is the sum of a log constant flux and the transiting signal and $\varepsilon(t)$ is the error at time t assumed to be independent Gaussian with mean zero and variance σ^2. It is common that the residuals are not well modelled by white noise, and this can lead to biases in the estimation of transit parameters and their uncertainties e.g. Carter & Winn (2009). In Jordán et al. (2013), a transit of the exoplanet WASP-6b was observed with Magellan in order to estimate its transmission spectra. The white light curve (time series of the stellar flux integrated over wavelength) was fit with a transit model and via a model-comparison process it was assessed that the residual structure was best described by a flicker model with power spectral density $\propto 1/f$, indicating a long memory time dependence. Other models tried where a white noise model and an ARMA(2,2) model. All models tried assumed that the observational times are equally spaced, which in their case is a good approximation to the data, although it is not exact. In Fig. 6, we show some statistics of the time gaps between the observations for this data, which we will use to illustrate the performance of our model on a data set that should be very well suited for methods that assume constant cadence but that does have some small departures from such behaviour.

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Catï¿½lica de Chile user on 08 October 2018

Table 6. Distribution of the 40 examples selected by its frequency range and class of variable stars.

Class	$f_1 \leq 0.1$	$0.1 < f_1 \leq 0.5$	$0.5 < f_1 \leq 1$	$1 < f_1 \leq 2$	$f_1 > 2$
Classical Cepheid (CLCEP)	2	4	–	–	–
Chem. Peculiar (CP)	–	1	–	–	–
Double Mode Cepheid (DMCEP)	–	–	1	2	–
Delta Scuti (DSCUT)	–	–	–	–	2
Beta Persei (EA)	–	1	4	2	–
Beta Lyrae (EB)	1	–	2	2	–
W Ursae Maj (EW)	–	1	1	1	2
Mira (MIRA)	4	–	–	–	–
PV Supergiants (PVSG)	–	1	–	–	–
RR Lyrae, FM (RRAB)	–	–	–	1	1
RR Lyrae, FO (RRC)	–	–	–	–	2
Semireg PV (SR)	1	–	–	–	–
SX Phoenicis (SXPHE)	–	–	–	–	1
Total	8	8	8	8	8

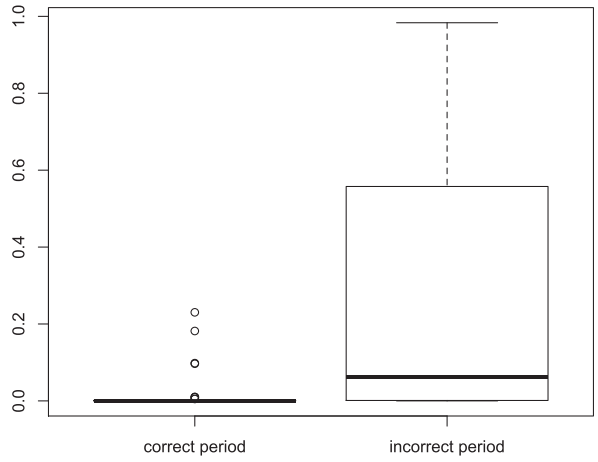


Figure 3. Boxplot of the distribution of ϕ, for the light curves using the correct frequency (on the left) and for the light curves using the incorrect frequency (on the right).

After fitting the model, described above, to a star with an exoplanet orbiting around it, we implement the IAR model on the residuals, which are shown in Fig. 7(a). These residuals correspond to the same data utilized at Jordán et al. (2013) and are shown in the left-bottom panel of fig. 6 in Jordán et al. (2013). The red triangle in Fig. 7(b) corresponds to $\log(\hat{\phi})$, where $\hat{\phi}$ is the estimator of the parameter of the IAR model. To evaluate whether this value of the parameter could have been obtained from a series with no temporal dependence, we perform a randomized experiment. In this experiment, we fixed the observation times of the time series, but shuffled the flux measurements a hundred times to obtain hundred estimates of the parameter ϕ, which allow us to have an estimate of the ϕ values that are expected to be observed when there is no temporal dependence in the time series. This distribution is shown in Fig. 7(b). Note that the actual value of $\hat{\phi}$ is very unlikely to have arisen from this distribution, having a p-value of 5.64×10^{-5}. This result is consistent with the results of Jordán et al. (2013), where they also find temporal structure on this data using a flicker-noise and an ARMA model.

REMARK 1. Observe that the *residuals* of the fitted model, defined as $y_t = z_t - g(t, \widehat{\theta})$, are not necessarily equal to the model *errors* δ_t, say. However, under the assumption that the estimator of $g(t, \theta)$, $g(t, \widehat{\theta})$ is consistent we have that asymptotically, $y_t \sim \delta_t$. Note that due to the irregularity of the observation times, the residuals do not share the same variance. A well-known procedure for assessing that the residuals are indeed white noise is the Ljung–Box test. Thus, we suggest to apply first this test to the adequately standarized residuals for whiteness, taking in consideration the sensibility of this test to the sample size. If the null hypothesis of white noise is rejected, then proceed to model the serial dependence observed in the residuals. Notice that the ultimate goal of this modelling approach is to obtain white noise residuals, that is, to remove all systematic error components. In the normal case, the theoretical residuals are correlated when the covariates are not orthogonal, which is standard in multiple linear regression. But in this case there are statistical tests, such as the Durbin–Watson type of test or Breusch–Godfrey test or Ljung–Box test, that assess whether the residuals remain correlated/autocorrelated. These tests have been extensively used in multiple linear regression. The purpose of the IAR model is to test whether there remain significant correlation on y_t and to model it.

8 DISCUSSION

In this work, we present an autoregressive model for irregularly observed time series (IAR), and we show that it is weakly stationary, and under some conditions, it is stationary and ergodic, providing a solid statistical framework to assess autocorrelation in the residuals of a model sampled at irregular times. We show that this model is not limited by Gaussian time series. We develop examples with samples from a Gamma and a Student's t-distributed series, in which the IAR model under the correct distribution outperforms the model under the Gaussian distribution. We further develop statistical tests to assess significance of the parameter of the model that measures autocorrelation of the time series. We have developed a maximum likelihood procedure to estimate the model and provide code in the R statistical software and in PYTHON.

We have illustrated two implementations of the model on astronomical data set to show some possible applications in this field for the identification of misspecified models and the assessment of the presence of time-correlated structure in time series. In both examples, we follow a two-stage approach for parameter estimation, i.e. first the parameter of the harmonic model are estimated and to the residuals of this model we implement and estimate the IAR model. This is certainly not ideal, as it would be more appealing to jointly estimate the parameter of the IAR and the harmonic model. We have not presented it in that form because of the examples that we have chosen. Periodic light curves from variable stars require to

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Catï¿½lica de Chile user on 08 October 2018

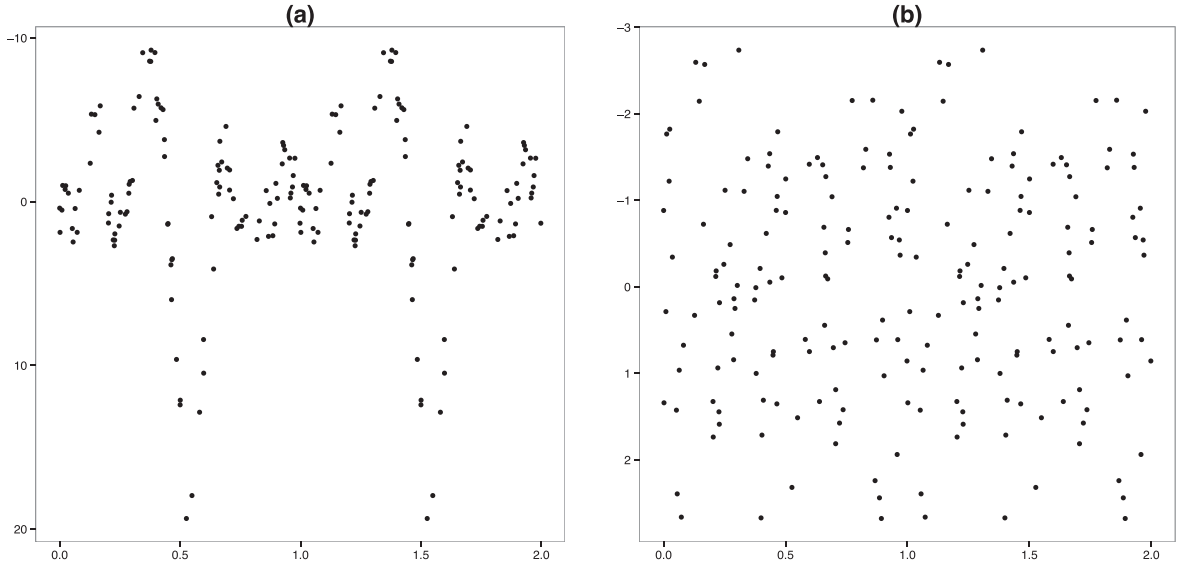


Figure 4. (a) Residuals of the best harmonic fit with one frequency for a simulated multiperiodic light curve; (b) residuals of the best harmonic best fit with two frequencies for the same simulated multiperiodic light curve.

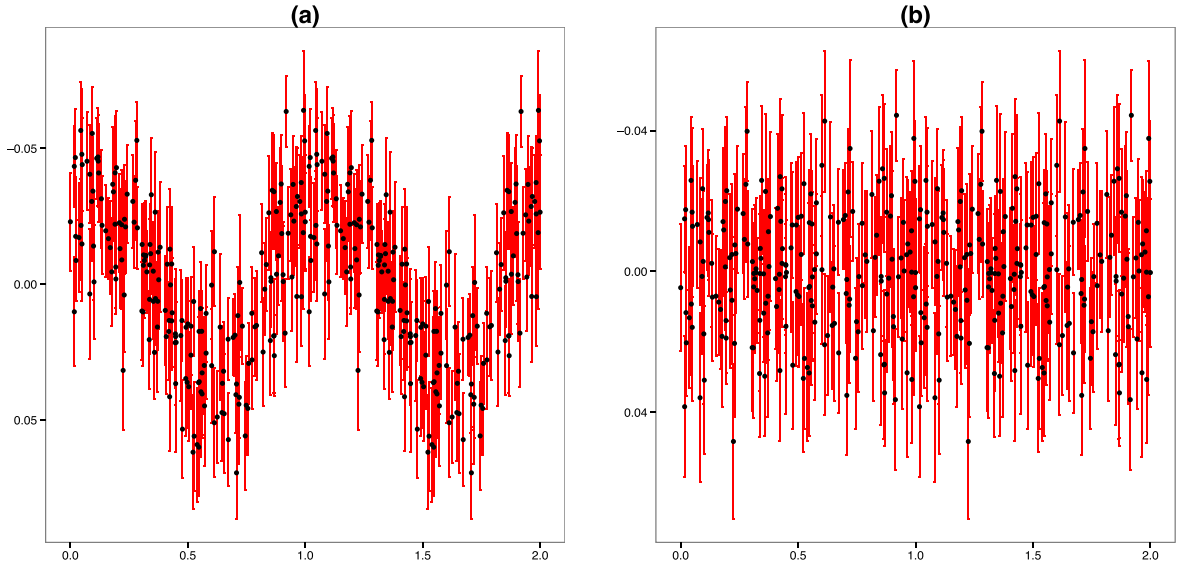


Figure 5. (a) Residuals of a double model Cepheid after fitting an harmonic model with one period; (b) residuals of the same variable star after fitting an harmonic model with two periods.

have a period estimated. While there are methodologies that estimate jointly the period and a parametric model (e.g. the coefficients of truncated Fourier series, Palmer 2009) by far the most common practice is to first estimate a period and then estimate the model parameters given a period (see e.g. Elorrieta et al. 2016 and references therein). We follow the same procedure with the light curve of the star with an orbiting exoplanet. For other implementations, we advocate simultaneous estimation of the parametric and IAR models.

The model presented here is a simple model that depends on one parameter that measures the autocorrelation of the series and another parameter that measures the size of the error of the model. Nevertheless, having correlated errors not accounted for in the specification of a model can have important consequences. For example, in the context of linear regression, the estimator of the error of the model can be biased toward zero. This can lead to confidence intervals that are too narrow, based on the t-statistic, and therefore, can produce falsely significant results.

A drawback that both the IAR model and the CAR(1) have is that they only allow to estimate positive autocorrelation, i.e. the parameter ϕ is constraint to be non-negative. In the case of the Gaussian and non-Gaussian IAR models, equation (5) would require a negative ϕ to the power of a real number which, in general, does not exist. In the case of the CAR(1) model, the autocorrelation is measured by $e^{-\alpha_0}$, where $\alpha_0 > 0$ for the process to be stationary, and therefore also takes only positive numbers. We are currently extending the IAR model to allow to estimate series with negative autocorrelation.

With this work, we try to entice the researchers to model time series with irregular times as series of discrete and not continuous

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Catï¿½lica de Chile user on 08 October 2018

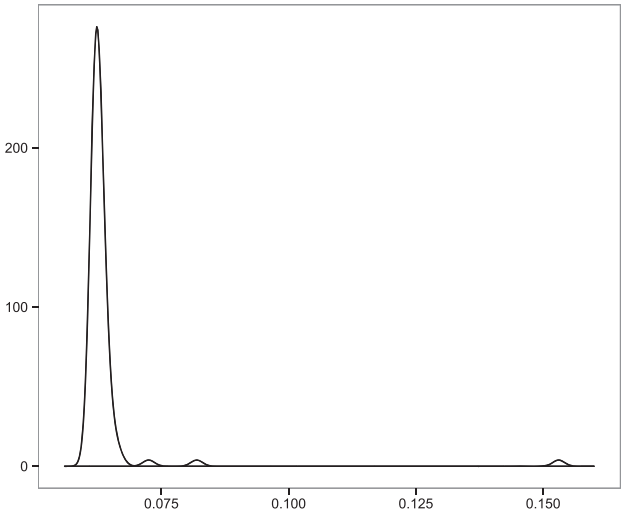


Figure 6. Density of the observational time gaps from WASP-6.

times as they have been commonly treated. This opens a new avenue for developing models that can fit irregular time series based on discrete times. These models can be simple but with sound statistical properties. We consider that the discrete representation for irregular time series is specially suitable for time series obtained from astronomical data sets because the gaps between observations can be very large, in the order of days, months or years. Whereas in disciplines where the time gaps between observations are tiny, a continuous model such as the continuous autoregressive model could be more suitable.

Software to implement the model and simulations are available in PYTHON and R upon request to the authors.

ACKNOWLEDGEMENTS

Support for this research was provided by grant IC120009, awarded to The Millennium Institute of Astrophysics (MAS) by Iniciativa Cientifica Milenio. WP acknowledges support from National Fund for Scientific and Technological Development (Fondecyt) grant 1160861. FE acknowledges support from National Commission for Scientific and Technological Research (CONICYT-PCHA) (Doctorado Nacional 2014-21140566).

REFERENCES

Adorf H.-M., 1995, in Shaw R. A., Payne H. E., Hayes J. J. E., eds, ASP Conf. Ser. Vol. 77, Astronomical Data Analysis Software and Systems IV. Astron. Soc. Pac., San Francisco, p. 460
Bailer-Jones C. A. L., 2011, MNRAS, 416, 1163
Bailer-Jones C. A. L., 2012, A&A, 546, A89
Belcher J., Hampton J. S., Wilson G. T., 1994, J. R. Stat. Soc. B (Methodology), 56, 141
Box G. E. P., Jenkins G. M., Reinsel G. C., Ljung G. M., 2015, Time Series Analysis: Forecasting and Control, 5th edn. John Wiley & Sons, Inc., Hoboken, New Jersey
Brewer B. J. et al., 2011, ApJ, 733, L33
Brockwell P., Davis R., 1991, Time Series: Theory and Methods: Theory and Methods. Springer Series in Statistics. Springer, New York
Brockwell P., Davis R., 2016, Introduction to Time Series and Forecasting, 3rd edn. Springer-Verlag, New York
Carter J. A., Winn J. N., 2009, ApJ, 704, 51
Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159
Done C., Mulchaey J. S., Mushotzky R. F., Arnaud K. A., 1992, ApJ, 395, 275
Drake A. J. et al., 2009, ApJ, 696, 870
Eckner A., 2014, A Framework for the Analysis of Unevenly Spaced Time Series Data. Working Paper
Elorrieta F. et al., 2016, A&A, 595, A82
Emmanoulopoulos D., McHardy I. M., Papadakis I. E., 2013, MNRAS, 433, 907
Erdogan E., Ma S., Beygelzimer A., Rish I., 2005, Statistical Models for Unequally Spaced Time Series. Proceedings of the fifth SIAM International Conference on Data Mining, SIAM, p. 626
Foreman-Mackey D., Agol E., Ambikasaran S., Angus R., 2017, AJ, 154, 220
Frieman J. A. et al., 2008, AJ, 135, 338
Ivezic Z. et al., 2008, preprint (arXiv:0805.2366)
Jones R. H., 1985, Handbook of Statistics, Vol. 5, Elsevier, p. 157
Jones R. H., 1993, Longitudinal Data with Serial Correlation: A State-Space Approach (Monographs on Statistics and Applied Probability). Chapman & Hall/CRC, London
Jordán A. et al., 2013, ApJ, 778, 184
Kaiser N. et al., 2002, in Tyson J. A., Wolff S., eds, Proc. SPIE Vol. 4836, Survey and Other Telescope Technologies and Discoveries. p. 154
Kalman R. E., 1960, Trans. ASME J. Basic Eng., 82, 35
Kelly B., Bechtold J., Siemiginowska A., 2009, ApJ, 698, 895
Kelly B. C., Becker A. C., Sobolewska M., Siemiginowska A., Uttley P., 2014, ApJ, 788, 33

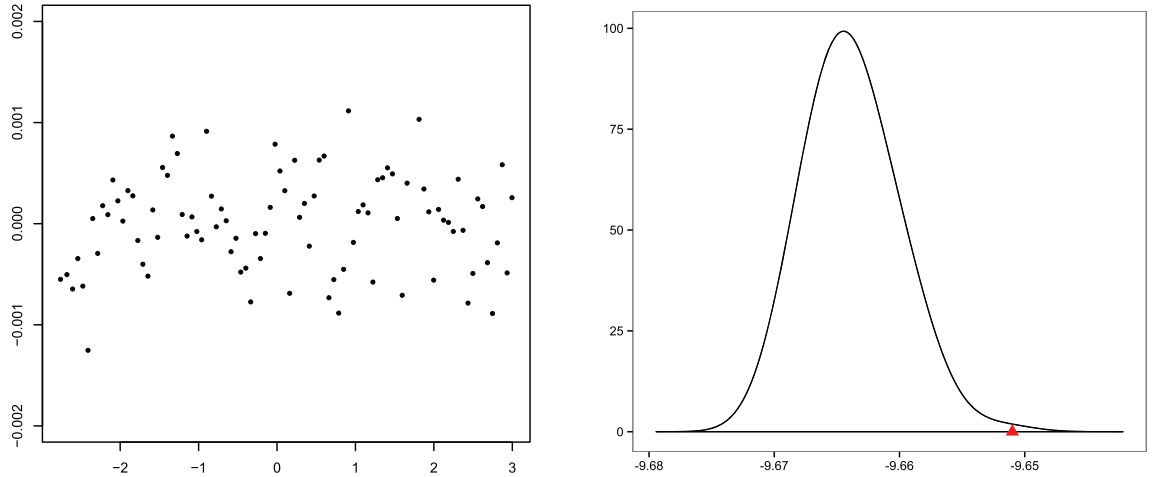


Figure 7. (a) Residuals after fitting the model for a transiting exoplanet; (b) the red triangle represent the $\log(\hat{\phi})$, where $\hat{\phi}$ is the parameter of the IAR model. The black line represents the density of the ϕ for the randomized experiment.

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Catï¿½lica de Chile user on 08 October 2018

Law N. M. et al., 2009, PASP, 121, 1395
Lomb N. R., 1976, Ap&SS, 39, 447
Mandel K., Agol E., 2002, ApJ, 580, L171
Minniti D. et al., 2010, New Astron., 15, 433
Palma W., 2007, Long Memory Time Series: Theory and Methods. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ
Palma W., 2016, Time Series Analysis. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ
Palma W., Zevallos M., 2011, Appl. Stoch. Models Business Ind., 27, 23
Palmer D. M., 2009, ApJ, 695, 496
Parzen E., 1984, Proceedings of a Symposium Held at Texas A&M University, College Station, Texas, February 10–13, 1983, Lecture Notes in Statistics: Time series analysis of irregularle observed data. Springer-Verlag, New York Berlin Heidleberg Tokyo
Perryman M. A. C. et al., 1997, A&A, 323, L49
Pichara K., Protopapas P., Kim D.-W., Marquette J.-B., Tisserand P., 2012, MNRAS, 427, 1284
Rasmussen C. E., 2006, Gaussian Processes for machine learning. MIT Press, Cambridge, USA
Rehfeld K., Marwan N., Heitzig J., Kurths J., 2011, Nonlinear Process. Geophys., 18, 389
Richards J. W. et al., 2011, ApJ, 733, 10
Scargle J. D., 1982, ApJ, 263, 835
Thiebaut C., Roques S., 2005, EURASIP J. Appl. Signal Process., 15, 2486
Tuomi M. et al., 2013, A&A, 551, A79
Udalski A., Soszynski I., Szymanski M., Kubiak M., Pietrzynski G., Wozniak P., Zebrun K., 1999, Acta Astron., 49, 223
Uttley P., McHardy I. M., Papadakis I. E., 2002, MNRAS, 332, 231
Zechmeister M., Kürster M., 2009, A&A, 496, 577

APPENDIX A: REPRESENTATION OF CAR(1) MODEL IN A FORM OF A DISCRETE IRREGULAR TIME SERIES

As mentioned in Section 2.1, the CAR(1) model is defined as the solution of the following stochastic differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\epsilon(t) + \alpha_0\epsilon(t) = \sigma_0\nu(t) + \beta, \tag{A1}$$

where $\nu(t) = \frac{\mathrm{d}}{\mathrm{d}t}w(t)$ and $w(t)$ is a Brownian motion or Wiener process. The derivative of $w(t)$ does not exist, so a proper way of writing equation (3) is as an Itô differential equation

$$\mathrm{d}\epsilon(t) + \alpha_0\epsilon(t)\mathrm{d}t = \mathrm{d}w(t) + \beta\mathrm{d}t, \tag{A2}$$

where $\mathrm{d}\varepsilon(t)$ and $\mathrm{d}w(t)$ denote the increments in ε and w in the time interval $(t, t + \mathrm{d}t)$, and $\varepsilon(0)$ is a random variable with finite variance and independent of $\{w(t)\}$.

The solution of equation (A2) can be written as

$$\mathrm{d}\epsilon(t) = \mathrm{e}^{-\alpha_0 t}\epsilon(0) + \mathrm{e}^{-\alpha_0 t}I(t) + \beta\mathrm{e}^{-\alpha_0 t}\int_0^t \mathrm{e}^{\alpha_0 u}\mathrm{d}u, \tag{A3}$$

where $I(t) = \sigma_0\int_0^t \mathrm{e}^{\alpha_0 u}\mathrm{d}w(u)$ is an Itô integral satisfying $E(I(t)) = 0$ and $\mathrm{Cov}(I(t+h), I(t)) = \sigma_0^2\int_0^t \mathrm{e}^{2\alpha_0 u}\mathrm{d}u$ for all $t \geq 0$ and $h > 0$. It can be shown that necessary and sufficient conditions for $\{\varepsilon(t)\}$ to be stationary are $\alpha_0 > 0$, $E(\varepsilon(0)) = \beta/\alpha_0$ and $\mathrm{Var}(\epsilon(0)) = \sigma_0^2/(2\alpha_0)$. Further, if $\epsilon(0) \sim N(\beta/\alpha_0, \sigma_0^2/(2\alpha_0))$, then the CAR(1) process is also Gaussian and stationary.

If $\alpha_0 > 0$ and $0 \leq s \leq t$, it follows from equation (A3) that $\varepsilon(t)$ can be expressed as

$$\epsilon(t) = \mathrm{e}^{-\alpha_0(t-s)}\epsilon(s) + \frac{\beta}{\alpha_0}(1 - \mathrm{e}^{-\alpha_0(t-s)}) + \mathrm{e}^{-\alpha_0 t}(I(t) - I(s)) \tag{A4}$$

or equivalently

$$\epsilon(t) - \frac{\beta}{\alpha_0} = \mathrm{e}^{-\alpha_0(t-s)}(\epsilon(s) - \frac{\beta}{\alpha_0}) + \mathrm{e}^{-\alpha_0 t}(I(t) - I(s)). \tag{A5}$$

APPENDIX B: PROOF OF THEOREM 1

For a given positive integer n, we can write

$$y_{t_j} = \phi^{t_j - t_{j-n}}\, y_{t_{j-n}} + \sigma\sum_{k=0}^{n-1}\phi^{t_j - t_{j-k}}\sqrt{1 - \phi^{2(t_{j-k} - t_{j-k-1})}}\,\varepsilon_{t_{j-k}}.$$

Notice that under the assumptions of the theorem the first term converges to zero in probability. On the other hand, we have that

$$\phi^{2(t_j - t_{j-k})} \leq k^{\alpha},$$

where

$$\alpha = C\log\phi^2.$$

Consequently,

$$\sum_{k=0}^{\infty}\phi^{2(t_j - t_{j-k})} \leq \sum_{k=0}^{\infty}k^{\alpha} < \infty,$$

since $\alpha < -1$ by assumption. Thus, the expression

$$y_{t_j} = \sigma\sum_{k=0}^{\infty}\phi^{t_j - t_{j-k}}\sqrt{1 - \phi^{2(t_{j-k} - t_{j-k-1})}}\,\varepsilon_{t_{j-k}} \tag{B1}$$

corresponds to a measurable transformation of the independent and identically distributed (i.i.d.) sequence $\{\varepsilon_{t_j}\}$. Therefore, due to Theorem 1.7 of Palma (2007), the sequence $\{y_{t_j}\}$ is stationary and ergodic.

Further, it is straightforward to see that the equation (B1) is a solution to the process defined by equation (5). This can be shown by plugging-in $y_{t_{j-1}}$, as defined in equation (B1), into the right side of equation (5). After some arithmetic one gets to y_{t_j}, showing that (B1) is indeed a solution to the process defined by (5).

APPENDIX C: PROOF OF LEMMA 1

It follows from Section 8.8 of Brockwell & Davis (1991). Observe that $t_j - t_{j-n}$ satisfies the condition of Theorem 1, consequently, given that $0 < \phi < 1$, the process $\{y_{t_j}\}$ is stationary and ergodic. Furthermore, the process satisfies the equation: $y_{t_j} = \phi^h\, y_{t_{j-1}} + \eta_{t_j}$, where η_{t_j} is a white noise sequence with variance $\sigma_\eta^2 = \sigma^2(1 - \phi^{2h})$. Consider the transformation $\theta = \phi^h$. Thus, an application of Brockwell & Davis (1991, p. 259) yields $\sqrt{n}\,(\widehat{\theta}_n - \theta) \to \mathrm{N}(0, \sigma_\theta^2)$, as $n \to \infty$, where $\sigma_\theta^2 = 1 - \theta^2$. Therefore, by defining $g(\theta) = \theta^{1/h}$, we have that $\phi = \theta^{1/h}$ and then by an application of the continuous mapping theorem we conclude that $\sqrt{n}\,(g(\widehat{\theta}_n) - g(\theta)) \to N(0, \sigma_\theta^2\,[g'(\theta)]^2)$, as $n \to \infty$. But,

$$\sigma_\theta^2[g'(\theta)]^2 = \frac{1 - \phi^{2h}}{h^2\,\phi^{2h-2}},$$

which completes the proof.

This paper has been typeset from a TEX/LATEX file prepared by the author.

Downloaded from https://academic.oup.com/mnras/article-abstract/481/4/4311/5094606 by Pontificia Universidad Católica de Chile user on 08 October 2018