Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **433,** 907–927 (2013)
Advance Access publication 2013 June 11

doi:10.1093/mnras/stt764

Generating artificial light curves: revisited and updated

D. Emmanoulopoulos,[1]★ I. M. McHardy[1] and I. E. Papadakis[2,3]

[1]*Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK*
[2]*Department of Physics and Institute of Theoretical and Computational Physics, University of Crete, 71003 Heraklion, Greece*
[3]*IESL, Foundation for Research and Technology, 71110 Heraklion, Greece*

Accepted 2013 April 30. Received 2013 April 29; in original form 2013 April 3

ABSTRACT

The production of artificial light curves with known statistical and variability properties is of great importance in astrophysics. Consolidating the confidence levels during cross-correlation studies, understanding the artefacts induced by sampling irregularities, establishing detection limits for future observatories are just some of the applications of simulated data sets. Currently, the widely used methodology of amplitude and phase randomization is able to produce artificial light curves which have a given underlying power spectral density (PSD) but which are strictly Gaussian distributed. This restriction is a significant limitation, since the majority of the light curves, e.g. active galactic nuclei, X-ray binaries, gamma-ray bursts, show strong deviations from Gaussianity exhibiting 'burst-like' events in their light curves yielding *long-tailed* probability density functions (PDFs). In this study, we propose a simple method which is able to precisely reproduce light curves which match both the PSD and the PDF of either an observed light curve or a theoretical model. The PDF can be representative of either the parent distribution or the actual distribution of the observed data, depending on the study to be conducted for a given source. The final artificial light curves contain all of the statistical and variability properties of the observed source or theoretical model, i.e. the same PDF and PSD, respectively. Within the framework of *Reproducible Research*, the code and the illustrative example used in this paper are both made publicly available in the form of an interactive MATHEMATICA notebook.

Key words: methods: statistical – galaxies: active – galaxies: individual: NGC4051, 3C454.3 – gamma-rays: galaxies – X-rays: binaries – X-rays: individual: CygX-1.

1 INTRODUCTION

Currently in astrophysics, artificial light curves are usually constructed using the procedure of Timmer & Koenig (1995, hereafter TK95). This method is able to produce ensembles of non-deterministic, normally distributed time series from a given underlying power spectral density (PSD) model, $\mathscr{P}(f)$, which represents the variability power as a function of temporal frequency, f. Resembling the method proposed by Davies & Harte (1987), it randomizes correctly both the phase and the amplitude of the Fourier components, thus advancing on the previous method of Done et al. (1992), which randomizes only the phase, assuming a deterministic amplitude which causes a long-term trend in the resulting simulated data sets.

There are numerous applications of the TK95 procedure in a plethora of astrophysical fields such as follows.

(i) Its use in the establishment of statistical confidence intervals during cross-correlation studies (e.g. Agudo et al. 2011; Bartlett et al. 2013).

(ii) Its application during microlensing studies (e.g. Ofek & Maoz 2003; Koptelova et al. 2010; Tewes, Courbin & Meylan 2012).

(iii) The detection of variability in large catalogues and surveys (Bauer et al. 2009; MacLeod et al. 2010; Villforth, Koekemoer & Grogin 2010; Primini et al. 2011), as well as for the detection of the smallest variability time-scales embedded in a data set (e.g. Aharonian et al. 2007).

(iv) Determination of the detection limits of astronomical instruments for a given type of astrophysical source [e.g. active galactic nuclei (AGN), gamma-ray bursts (GRBs)] (e.g. Greene et al. 2010; Primini et al. 2011; Khabibullin, Sazonov & Sunyaev 2012; Doro et al. 2013).

(v) Its use in the derivation of confidence intervals during the study of quasi-periodic oscillations of Galactic and extragalactic objects (e.g. Benlloch et al. 2001; Gierliński et al. 2008; Do et al. 2009), as well as for the statistical characterization of periodic and pulsed patterns in stellar photometric data (Stanishev et al. 2002; Grosso et al. 2010; Blomme et al. 2011).

★E-mail: D.Emmanoulopoulos@soton.ac.uk

(vi) Its central role in the estimation of the underlying PSD of irregularly sampled AGN light curves within the procedure proposed by Uttley, McHardy & Papadakis (2002).

(vii) Its vital use in the study of both the powers and limitations of a given statistical method (e.g. Zhang et al. 2004; Vaughan 2005; Emmanoulopoulos, McHardy & Uttley 2010; Góra, Bernardini & Cruz Silva 2011).

(viii) Its use in the fields of Solar astrophysics (Rajaguru, Hughes & Thompson 2004) and geophysics (Venema et al. 2006).

At this point, it is very important to note that the above-mentioned method of TK95 is appropriate for the production of Gaussian artificial light curves *only*. This means that the resultant surrogate data sets[1] preserve only the first two statistical moments of the original data set, i.e. the mean value, μ, and the variance, σ^2, ignoring potential higher order statistical moments, such as skewness, kurtosis, or multimodes found in the normalized flux distribution of the data, i.e. probability density function (PDF), corresponding either to the parent or observed distribution. Thus, Gaussian light curves show on average the same amplitude variations above and below the mean, resulting in a zero skewness distribution of data points. Another characteristic is that since the surrogates are normally distributed, there is always a finite probability for the artificial data points to become negative.[2] However, the light curve of any astronomical source must by default remain positive and so must the resulting PDF.

In this framework, TK95 methodology can be used to simulate observed data sets which are Gaussian distributed in the broad sense, i.e. having a negligible skewness and/or kurtosis. On a large number of occasions however, light curves exhibit a 'burst'-like behaviour e.g. in X-rays with *RXTE* and *XMM–Newton* (e.g. Chitnis et al. 2009; Vaughan et al. 2011), in γ-rays with the Large Area Telescope (LAT) onboard *Fermi* (e.g. Chatterjee et al. 2009; Agudo et al. 2011), in which the events are distributed following right *heavy-tailed* distributions. This implies occurrence probabilities of high flux values larger than those expected from Gaussian distributions, and thus the Gaussian TK95 products cannot be used for the establishment of confidence intervals e.g. in cross-correlation studies.

Furthermore, the rms–flux relation, i.e. the linear scaling of the fractional root-mean-square (rms) variability amplitude with the flux, observed in both AGN and X-ray binaries (XRBs; Uttley & McHardy 2001; Uttley, McHardy & Vaughan 2005; Gandhi 2009; McHardy 2010), cannot be reproduced by the TK95 algorithm. Uttley et al. (2005) show that such a behaviour can arise from a non-linear multiplicative variability process in which the parent distribution follows a log-normal distribution. The authors therefore suggest a modification to the TK95 products involving exponentiation (in base e) of the normally distributed artificial data sets, yielding light curves which both possess a log-normal distribution and exhibit the rms–flux relation. Although the normalization of the input PSD, $\mathscr{P}_{\rm rescale}(f)$, is selected in such a way that the variance of the final products matches that of the observed light curve, the actual shape of the PSD is distorted from the original one, $\mathscr{P}(f)\delta f$, in such a way that the actual variability power within a given frequency range, $\mathscr{P}_{\rm rescale}(f)\delta f$, differs from the genuine one, $\mathscr{P}(f)\delta f$.

[1] The term 'surrogate' data set (after Theiler et al. 1992) will be used throughout the paper in exactly the same way as the terms 'artificial', 'simulated' and 'synthetic' data set.

[2] For a given point this probability is equal to $0.5\,{\rm erfc}[\mu/(\sigma\sqrt{2})]$, where erfc denotes the complementary error function.

Apart from this PSD distortion, the exponentiation transformation cannot be generalized to arbitrary parent or observed distributions (depending on the type of study). The specification of the parent distribution requires either very large data set (e.g. for the case of Cyg X-1 253 144 data points are required to form the parent log-normal distribution; Uttley et al. 2005) or a theoretical model (e.g. Kelly, Sobolewska & Siemiginowska 2011). Employment of the parent distribution is of vital importance in comparing variability properties of data sets obtained over a long period of time which map the complete variability behaviour of the source. Nevertheless, it is sometimes crucial to establish the detection significance of a given result coming from a single observed data set. This approach has been used several times in the field of reverberation studies, in the form of flux redistribution or random subset selection (e.g. Peterson et al. 1998), or the detection of time lags in very high energy Cherenkov astronomy in the framework of Quantum Gravity (Aharonian et al. 2008). For the case of transient phenomena in particular, e.g. GRBs, in which the concept of a parent distribution is not applicable, only a single realization is available for each observation and thus this should be used as the PDF.

In this paper, we put forward a simple method which combines the routine of TK95 and the iterative amplitude adjusted Fourier transform algorithm of Schreiber & Schmitz (1996, hereafter SS96), which produces artificial light curves which possess exactly the same PSD and PDF as the originally observed light curve (or a theoretical model). Thus, the surrogates will have exactly the same variability and statistical properties as the observed light curve. Initially, in Section 2 we describe in detail the method. For illustrative purposes, in Section 3 we then apply it to the case of the well-studied type I Seyfert AGN NGC 4051, using *XMM–Newton* observations. Following that, in Section 4, we produce artificial light curves for the γ-ray blazar 3C 454.3, using *Fermi*-LAT observations, and the XRB Cyg X-1 using observation obtained by the All Sky Monitor (ASM), onboard *RXTE*. In Section 5, we present an application in cross-correlation analysis, and in Section 6 we reproduce the rms–flux relation for the case of log-normally distributed light curves. In Section 7, we discuss which properties of the light curves are preserved during our simulation process, and finally a discussion together with a summary of our results can be found in Section 8. In Appendix A, we give the basic definitions and properties of the various quantities, i.e. periodogram, PSD and PDF, as well as the various fitting procedures that will be used throughout this paper. In Appendix B, we elucidate the differences between statistical moments and cumulants, which are commonly confused in the astronomical literature.

Throughout the paper the error estimates for the various best-fitting model parameters correspond to the 90 per cent confidence intervals unless otherwise stated. The error bars of the plot points in all the figures indicate the 68.3 per cent confidence intervals.

2 METHODOLOGY

2.1 The algorithm

This method is a combination of TK95 and SS96, with some significant alterations and modifications which join the two together.

Consider an observed light curve $x_{\rm obs}(t)$ consisting of N uniformly sampled observations (sampling rate Δt), $\{t_i, x_{\rm obs}(t_i)\}$ for $i = 1, 2, \ldots, N$. The light curve has an underlying PSD, $\mathscr{P}(f)$, and an observed (or 'parent', depending on the purpose of the statistical study) PDF, PDF $[0 \leq x_{\rm obs}(t) < \infty]$. Note that both/either PSD and/or PDF can also originate from a theoretical model which we want to check

the statistical properties of its products, i.e. time series. Note that if one wishes to take into account the various spectral distortion effects (Section 2.3), then one should adjust both the simulation length N and the time resolution accordingly as described in Section 2.3.

(i) Using the TK95 procedure, a normally distributed time series[3] is produced, $x_{\rm norm}(t)$, consisting of N values and an underlying PSD identical to $\mathscr{P}(f)$. Then, for each Fourier frequency, f_j, the discrete Fourier transform (DFT), $DFT_{\rm norm}(j)$, is estimated and from this the corresponding amplitudes, $\mathscr{A}_{\rm norm}(j)$, phases, $\phi_{\rm norm}(j)$, and periodogram, $P_{\rm norm}(f_j)$ (equations A2, A3 and A4, respectively). Note that since the iteration algorithm aims to produce artificial source light curves, the input PSD should not contain the Poisson noise component.

(ii) From the PDF $[0 \leq x_{\rm obs}(t) < \infty]$, a series of N pseudo-random numbers is produced which form a white noise data set, $x_{\rm sim,\,1}(t)$. Then, at each Fourier frequency, the DFT of $x_{\rm sim,\,1}(t)$ is estimated, $DFT_{\rm sim,1}(j)$, and from that the corresponding amplitudes, $\mathscr{A}_{\rm sim,1}(j)$, phases, $\phi_{\rm sim,\,1}(j)$, f_j and periodogram, $P_{\rm sim,\,1}(f_j)$.[4]

(iii) *Spectral adjustment*. For each frequency, f_j, the amplitudes $\mathscr{A}_{\rm sim,1}(j)$ are replaced with the amplitudes $\mathscr{A}_{\rm norm}(j)$, whilst keeping the phases $\phi_{\rm sim,\,1}(j)$ unaltered. This yields the adjusted DFT of $x_{\rm sim,\,1}(t)$, $DFT_{\rm sim.adjust,1}(j)$, on which we then perform an inverse discrete Fourier transform (IDFT), yielding the time series, $x_{\rm sim.adjust,\,1}(t)$. This time series has an identical underlying PSD to the desired one, $\mathscr{P}(f)$, but with a distribution of measurements which has been altered from that of PDF $[0 \leq x_{\rm obs}(t) < \infty]$.

(iv) *Amplitude adjustment*. A new time series is created from the values of $x_{\rm sim,\,1}(t)$ ordered based on the ranking of $x_{\rm sim.adjust,\,1}(t)$. This means that the highest value of $x_{\rm sim.adjust,\,1}(t)$ is replaced by the highest value of $x_{\rm sim,\,1}(t)$, the second highest value of $x_{\rm sim.adjust,\,1}(t)$ is replaced by the second highest value of $x_{\rm sim,\,1}(t)$, and so on. The resulting data train, $x_{\rm sim,\,2}(t)$, is distributed exactly as PDF $[0 \leq x_{\rm obs}(t) < \infty]$ but its PSD differs from the target one, $\mathscr{P}(f)$.

(v) The same process is repeated in an iterative fashion κ times, starting from step (ii), until the resulting products remain the same, i.e. $x_{\rm sim,\,k\,+\,1}(t) \equiv x_{\rm sim,\,k}(t)$ (convergence):

(a) 2nd iteration: $x_{\rm sim,\,1}(t)$ is replaced by $x_{\rm sim,\,2}(t)$;
(b) 3rd iteration: $x_{\rm sim,\,2}(t)$ is replaced by $x_{\rm sim,\,3}(t)$;
(c) κth iteration: $x_{\rm sim,\,\kappa\text{-}1}(t)$ is replaced by $x_{\rm sim,\,\kappa}(t)$.

After a given number of iterations, e.g. $\lambda = \kappa + 1$, the synthetic light-curve products do not change (i.e. convergence) and thus the $x_{\rm sim,\,\lambda}(t)$ iterated product comprises the final artificial light-curve product. The exact number of iterations depends on the length of the original data set, the underlying input PSD and the input PDF. More about the convergence can be found in Sections 3.1.1 and 3.2.2 using Monte Carlo simulations. Note that for the case of a Gaussian PDF, the iteration process gives exactly equivalent results with the TK95 products since step (i) yields products which are already Gaussian distributed. The flowchart of the above-mentioned method is given in Fig. 1.

[3] Actually this is an asymptomatically normally distributed time series. Despite the fact that the TK95 procedure corresponds to a realization of a Gaussian process, individual artificial data set products (of finite length) may not be necessarily normally distributed since the Gaussianity of the process limits the asymptotic distribution only, $N \rightarrow \infty$.

[4] For $x_{\rm sim,\,1}(t)$ the periodogram, $P_{\rm sim,\,1}(f)$, corresponds by default to an underlying PSD with a slope of $\alpha = 0$ (since it represents a white noise process).

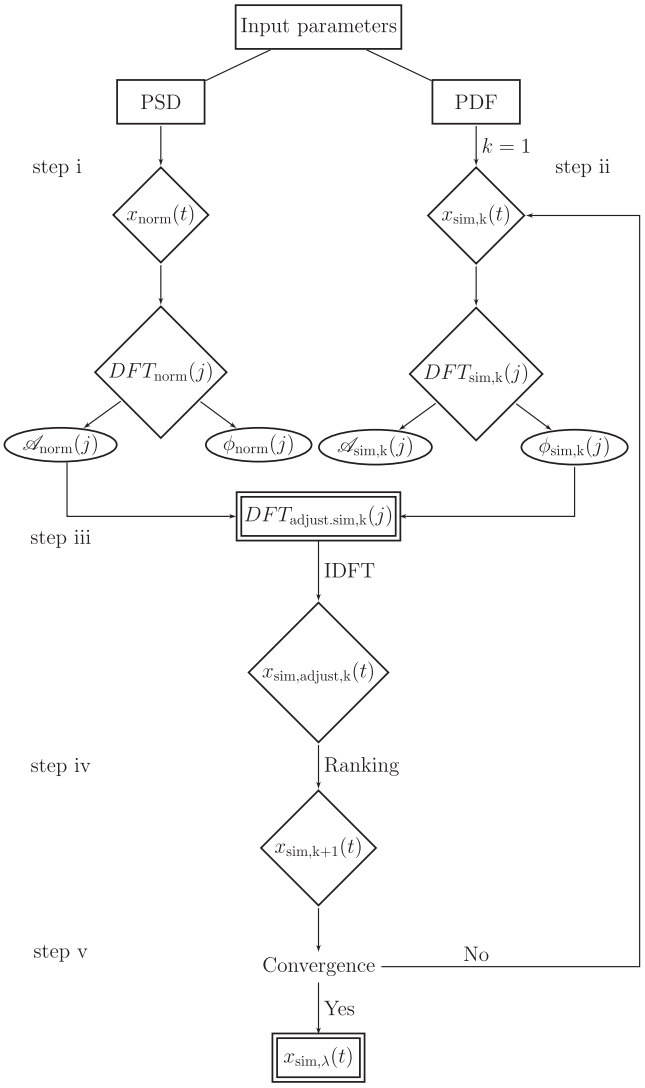

Figure 1. The flow chart diagram showing the various steps of the method.

2.2 Appropriate treatment of the Poisson noise

The final simulated product, $x_{\rm sim,\,\lambda}(t)$, has the desired distribution and PSD corresponding to the source light curve produced. Since the observed light curve is a product of a counting detector process, the observations are affected by Poisson noise, which is imprinted in the corresponding PSD as a constant component. In order to mimic this effect, each light-curve point $x_{\rm sim,\,\lambda}(t) = \{x_{\rm sim,\,\lambda}(t_1), x_{\rm sim,\,\lambda}(t_2), \ldots, x_{\rm sim,\,\lambda}(t_N)\}$ is replaced by an appropriate Poisson random variate

$$x_{\rm sim,Pois,\lambda}(t_i) \sim \frac{Pois[\mu = x_{\rm sim,\lambda}(t_i)\Delta t]}{\Delta t} \qquad \text{for } i = 1, \ldots, N, \qquad (1)$$

where $Pois[x_{\rm sim,\lambda}(t_i)\Delta t]$ dictates the probability mass function of the Poisson distribution with a mean value of $x_{\rm sim,\,\lambda}(t_i)\Delta t$.

2.3 Spectral distortions: red noise leak and aliasing

In the case of a non-white noise PSD (as in the case of AGN light curves), the periodogram estimates tend to be biased due to 'red noise leak' (the transfer of variability power from the low to high frequencies due to the finite length of observations; Deeter & Boynton 1982; Deeter 1984) and aliasing effects (fold-back of

variability power from high frequencies to lower frequencies due to the finite time resolution; Kirchner 2005).

These are two very well understood spectral distortions induced by the sampling properties of the data set, and they can be taken into account in the usual manner (e.g. Uttley et al. 2002). In order to take into account the 'red noise leak' effect, we produce surrogate data sets which are much longer than the observed data set (e.g. 100 times) and then we randomly select a subset having the desired length.

With respect to the aliasing, since we are dealing with data averaged over time intervals, $\Delta t_{\rm sample}$, rather than simply sampled data, its effect is very much reduced (van der Klis 1988). Nevertheless, if one also wishes to include the aliasing effect in the simulations, for the case of unbinned data (e.g. count data), then one should increase the time resolution of the simulations e.g. to 10 per cent of $\Delta t_{\rm sample}$.

With these two approaches, the dependences in both the Fourier amplitudes (equation A2) and the Fourier phases (equation A3) are taken into consideration. However, if one wishes to carry out statistical studies that deal only with the Fourier amplitudes (as in all the examples shown in this paper), then the length and binning adjustments can be applied during the first step of the above-mentioned method, i.e. during the application of the TK95 procedure. In this way, the TK95 artificial light curves, $x_{\rm norm}(t)$, carry all the spectral distortion effects that involve only the Fourier amplitudes, which will then be passed to the final surrogates during the spectral adjustment stage (step iii). A major advantage to this is that the whole process is much faster since the iteration process involves data sets which have lengths equal to that of the observed data set.

Note once again that for statistical studies that involve Fourier phases, e.g. phase-lag spectra studies, one should follow the initial recipe, i.e. carry out the whole simulation for longer and more finely sampled surrogate data sets, and select a subset which has the desired length from the final converged iteration product. The effect of red noise leak on the phases is rarely discussed in the literature despite the fact that it adds significant dependences to the phases.

For demonstrative purposes, we create an artificial light curve, using the TK95 procedure, consisting of 10^6 data points, binned in 100 s, with an input PSD which has a power-law model of slope -2.5. We then chop the light curve into 1000 segments each one consisting of 1000 consecutive points. For each data set and for each Fourier frequency, f_j, we estimate the DFT (equation A1) and from this its amplitude (via the periodogram, equation A4) and its phase (equation A3). Finally, for each f_j, we average the periodogram estimates and phases; the results are shown in the top panels of Fig. 2.

The top-left panel of Fig. 2 shows clearly the effect of red noise leak for the amplitudes, something that has been extensively discussed in the literature. The top-right panel of Fig. 2 shows vividly that the red noise leak also affects the phases in a very distinctive way. The onset of the effect is around 10^{-4} Hz (the same as it is for the amplitudes) and from then on all the phases follow an arched trend towards negative values. The last point in this plot corresponds to the phase estimate at the Nyquist frequency, $f_{N/2} = f_{\rm Nyq}$, for which the DFT is always a real number (positive or negative). This means that, for $f_{\rm Nyq}$, we average only the phases corresponding to the negative values, which are always π, over the total number of points, since for the positive values the arg is not formally defined (in this context one could consider it to be equal to 0). On average, one should get values around $\langle \phi_{N/2} \rangle \simeq \pi/2 \simeq 1.57$ (i.e. roughly equal numbers of positive and negative values).

We repeat the same process but this time with a constant underlying PSD, i.e. a power law with slope 0. As we can see from the bottom panels of Fig. 2, both the amplitudes (bottom-left panel) and the phases (bottom-right panel) are not affected by the red noise leak effect. It is clear that, as in the case of the Fourier amplitudes, the effect of red noise leak for the Fourier phases depends on the shape of the input PSD, e.g. the softer the power law of the underlying PSD, the greater is the effect of the red noise leak. As we discussed previously, our methodology correctly takes this effect into account by extending the total length of the surrogate data set and then chopping the converged final synthetic light curve to the desired length.

2.4 Basic differences and advantages from previous works

Our method is essentially a marriage of TK95 and SS96 algorithms. The former remains exactly the same during the application of the method but the latter (i.e. the iterative amplitude adjusted Fourier transform) contains several key differences, from the SS96, which makes it suitable particularly for the needs of astronomical data sets.

(i) We use a pseudo-random data set following the estimated distribution, rather than a shuffled version of the original observed data set.

(ii) We replace the Fourier phases of the TK95 products rather than those of the original data set.

(iii) All the spectral distortion effects due to the finite length and sampling rate of the observed data set (i.e. red noise leak and aliasing) are taken into account.

The coupling of the two methods allows us to study not only observed light curves but also theoretical models that give predictions about the PSD and the PDF of a given astrophysical object. The TK95 is carrying the spectral information (PSD) and the SS96 is distributing the various measurements (PDF) accordingly. In this way, we can produce, based on a theoretical model, realistic and positively defined non-Gaussian synthetic light curves as opposed to only Gaussian light curves coming from TK95.

The problem of generating stochastic sequences of numbers with specified properties is extensively analysed in the literature since the early 1970s (for a complete reference guide, see Sowey 1986). In particular, Liu & Munson (1982) proposed a white Gaussian noise input to a linear digital filter followed by a zero-memory non-linearity (ZMNL). The ZMNL is chosen so that the desired distribution is exactly realized and the digital filter is designed so that the desired autocovariance is closely approximated. Hunter & Kearney (1983) proposed a method for the generation of random number sequences with an arbitrarily specified first-order probability distribution function (PDF) and an arbitrarily specified first-order autocorrelation function (ACF). The procedure involves a stochastic optimization algorithm which minimizes the squared sum between the desired (output) and the actual (observed) ACF estimates.

An iterative method was developed by Yamazaki & Shinozuka (1988) which generates Gaussian distributed samples with a given periodogram which are then mapped into non-Gaussian distributed numbers. This is achieved by employing the invert expression of the target PDF (distribution distortion method), and the iteration process aims to correct the altered periodogram estimates (as they come out from the mapping process) to match the *desideratum* periodogram. The correlation distortion method was used by Johnson (1994), and consists of a non-linear transformation which is applied to construct non-Gaussian correlated features from correlated Gaussian random draws. Finally, Gurley, Kareem & Tognarelli (1996) presented a series of mathematical approaches using Volterra series

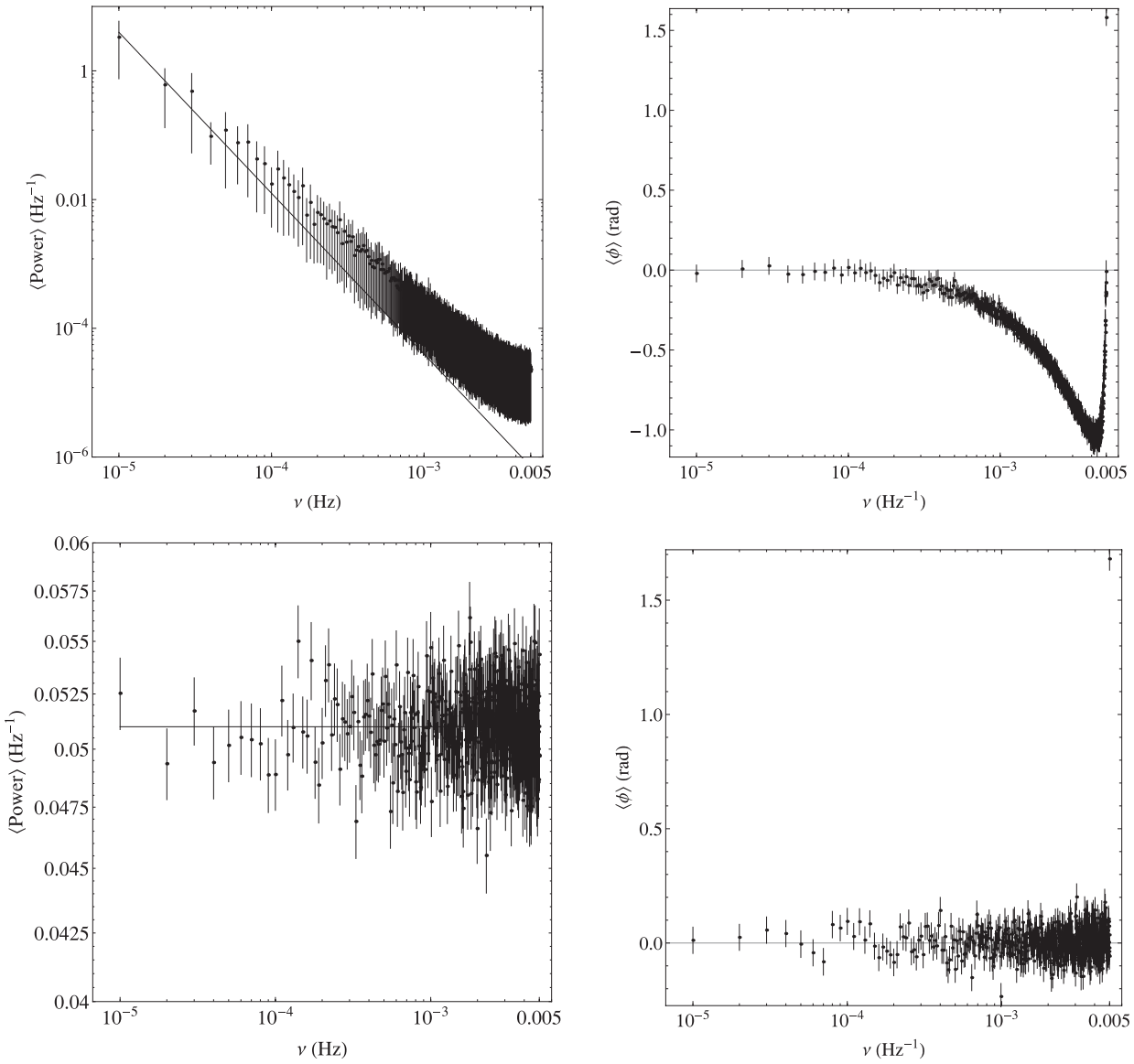

Figure 2. The effect of red noise leak in the Fourier amplitudes and phases. Top-left panel: the averaged periodogram estimates for an input PSD with a power-law shape of a slope −2.5. Top-right panel: the averaged Fourier phases for an input PSD with a power-law shape of a slope −2.5. Bottom-left panel: the averaged periodogram estimates for an input PSD with a power-law shape of a slope 0. Bottom-right panel: the averaged Fourier phases for an input PSD with a power-law shape of a slope 0.

and analytical kernels to achieve bispectral matching. In the same work, a neural network system identification model is employed for simulation also demonstrating the ability to match higher order spectral characteristics.

Besides the above-mentioned differences, our method differs fundamentally from all the previous methodologies with respect to the matching process of the PSD. We are not interested in matching the individual periodogram estimates (derived from the observed data set), but instead in the underlying PSD. In this way, at a given Fourier frequency f_j, the various periodogram estimates, $P(f_j)$, are distributed asymptotically around $\mathscr{P}(f_j)$ as a gamma distribution, $\Gamma[\nu/2, \mathscr{P}(f_j)]$ (equation A9)[5] with ν degrees of freedom (d.o.f.) corresponding to $\nu = 1$ for the Nyquist frequency and $\nu = 2$ for all other frequencies.

[5] In the literature this is usually referred to as 'scaled χ^2 distribution' with 2 d.o.f. (equation A8).

2.5 A publicly available code in the form of an active document

In the spirit of *Reproducible Results* and *Active Documents* (Claerbout 1990), we provide an interactive MATHEMATICA notebook (created with the version: 9.0.1.0) which contains the complete numerical code together with the example presented in Section 3. In detail the notebook contains

(i) the *XMM–Newton* data set of the AGN NGC 4051 which is used in Section 3,

(ii) a version of the TK95 code taking into consideration (if needed) the spectral distortions described in Section 2.3,

(iii) the iteration algorithm (SS96),

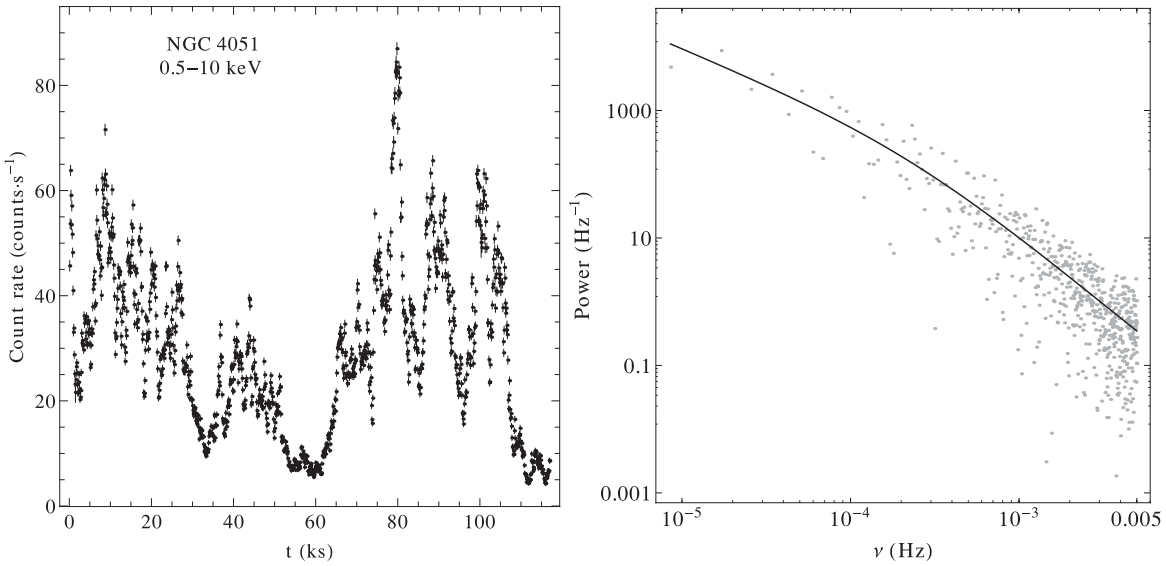

Figure 3. The *XMM–Newton* data set of NGC 4051. Left-hand panel: the EPIC pn and MOS combined light curve in the 0.5–10 keV energy band in bins of 100 s (obs ID: 0109141401, revolution: 0263). Right-hand panel: the corresponding periodogram estimates, $P_{\rm obs}(f)$ (grey points), and the underlying best-fitting PSD model, $\mathscr{P}(f;\vec{\gamma}_{\rm bf}, c_{\rm bf})$ (black line).

(iv) the addition of the Poisson noise as described in Section 2.2 and

(v) an animation of the simulated products at the various iteration steps.

It can be found on the web[6] as part of this paper (see Supporting Information in the online version) or at http://www.astro.soton.ac.uk/~de1e08/ArtificialLightCurves/. By changing the two random seeds (used for steps i and ii), the whole document is automatically updated and a new artificial light curve is produced. The simple numerical code, provided in the MATHEMATICA notebook, can be written in a much more compact form (i.e. much more computationally efficient), but for clarity purposes we have split it up into various programming lines. Due to its simple nature, the code can be implemented in any programming language.

3 APPLICATION: THE CASE OF NGC 4051

3.1 Step-by-step procedure for a single realization

We will now apply the method to a single X-ray data set of the type I Seyfert galaxy NGC 4051 ($z = 0.023\,36$) obtained by the European Photon Imaging Camera (EPIC) aboard *XMM–Newton* observatory (obs ID: 0109141401, revolution: 0263). The observed 0.5–10 keV light curve of NGC 4051 (sample light curve) is shown in the left-hand panel of Fig. 3 consisting of $N = 1170$ data points in bins of 100 s.

The following process is carried out for illustrative purposes only and aims to simulate a single artificial light curve with the same underlying PSD as the sample light curve, and an identical *observed* PDF (as opposed to the parent PDF). Nevertheless, depending on the purpose of the study one can select an appropriate PDF depicting either the underlying or the observed statistical properties. The method, of course, can also produce artificial light curves coming directly from a theoretical model, which specifies an underlying PDF and PSD, without the requirement of having an actual observed light curve. Since the PSD and the PDF of the observed data set are the two input parameters of the method, we first estimate these. Note that since we will not be performing any studies involving Fourier phases, the effect of 'red noise leak' has been taken into account during (step i), i.e. by producing a TK95 artificial data set 1000 times longer than the original NGC 4051 data set. For this particular data set (something which is generally true for *XMM–Newton* data sets), the aliasing effect is insignificant since we are dealing with averaged consecutive measurements.

Initially we derive the periodogram of the sample light curve, $P_{\rm obs}(f)$ (Fig. 3, right-hand panel, grey points), and then we estimate the underlying PSD by fitting the smoothly bending power-law model plus a constant, c, representing the Poisson noise level:

$$\mathscr{P}(f;\boldsymbol{\gamma},c) = \frac{Af^{-\alpha_{\rm low}}}{1+(f/f_{\rm bend})^{\alpha_{\rm high}-\alpha_{\rm low}}} + c, \qquad (2)$$

in which $\boldsymbol{\gamma} = \{A, f_{\rm bend}, \alpha_{\rm low}, \alpha_{\rm high}\}$, with components as the source's PSD model parameters, i.e. normalization, bend frequency, low- and high-frequency slopes, respectively. During the fit we fix the $\alpha_{\rm low}$ to 1.1 (as derived from long-term *RXTE* data; McHardy et al. 2004) and the best-fitting model parameters are $\boldsymbol{\gamma}_{\rm bf} = \{0.030 \pm 0.004\ {\rm Hz}^{-1}, 2.3^{+1.2}_{-0.9} \times 10^{-4}\ {\rm Hz}, 1.1, 2.20^{+0.07}_{-0.04}\}$ and $c_{\rm bf} = 9.2^{+0.7}_{-0.8} \times 10^{-3}\ {\rm Hz}^{-1}$ (Fig. 3, right-hand panel, black line). Note that the derived best-fitting values agree entirely with the ones given by Vaughan et al. (2011) but that the best-fitting $f_{\rm bend}$ differs from the value $8^{+4}_{-3} \times 10^{-4}$ Hz estimated by McHardy et al. (2004) (note that the error estimates correspond to the 90 per cent confidence intervals). This best-fitting model will be the target PSD which should be matched by the surrogate data sets.

In order to assess the probability distribution of the sample data set, we then form its probability density function histogram (Fig. 4, left-hand panel, black line). The latter exhibits two clear modes: the first narrow mode corresponds to the low-count-rate regimes (e.g. the regions around 35, 55 and 110 ks in the left-hand panel of Fig. 3), and the second broader mode to the high source states. We parametrize the observed distribution of the sample data set by fitting a probability model. For this particular data set of NGC 4051, we select a mixture distribution model consisting of a gamma

[6] You can also request the MATHEMATICA notebook via e-mail to D.Emmanoulopoulos@soton.ac.uk.

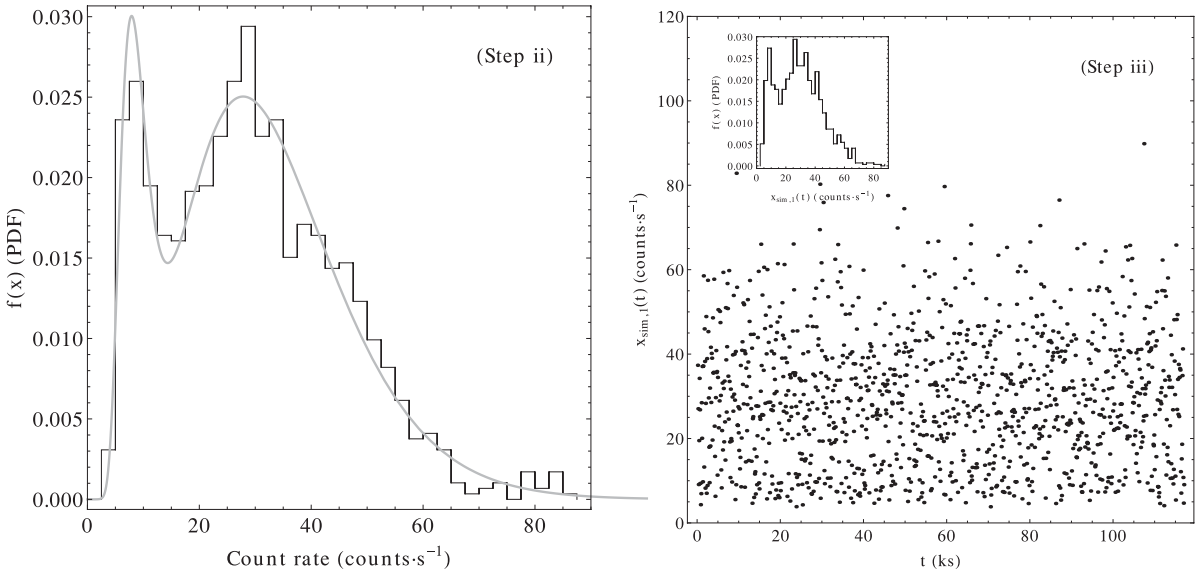

Figure 4. (Step ii) Left-hand panel: the PDF histogram of the observed data (black line) together with the best-fitting mixture distribution model, $\mathfrak{f}_{\rm mix}(x;\boldsymbol{\eta}_{\rm bf})$ (grey line, equation 3). Right-hand panel: an ensemble of N pseudo-random variates produced from equation (3) and the inset shows its corresponding PDF histogram.

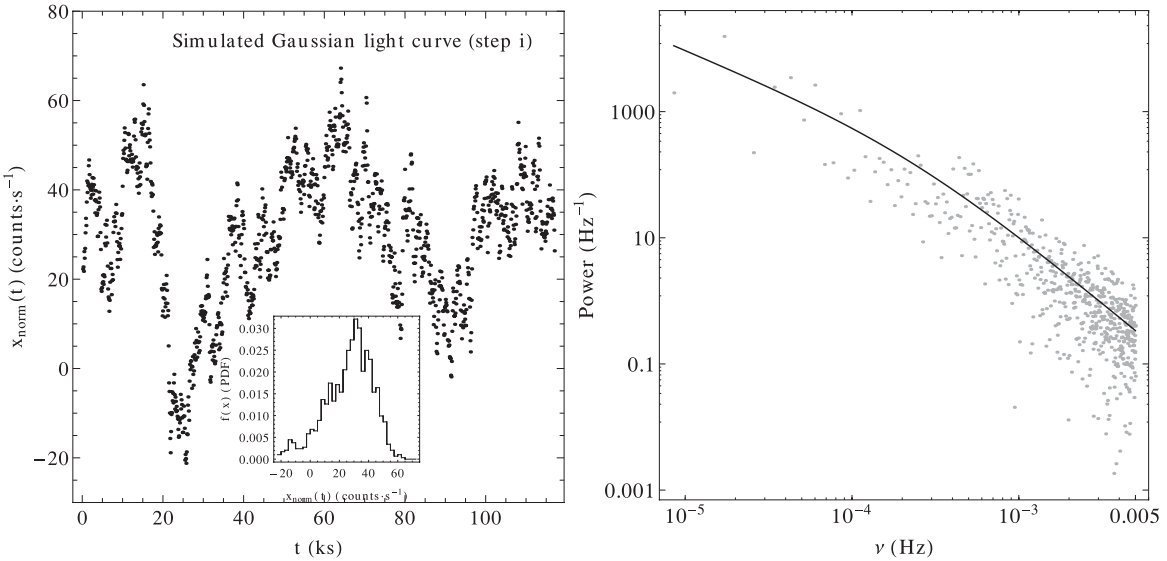

Figure 5. (Step i) Left-hand panel: a normally distributed simulated light curve created using the original data NGC 4051's best-fitting PSD ($c = 0$) and its PDF histogram (inset). Right-hand panel: the corresponding periodogram estimates (grey points) and the underlying target PSD model, $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)$ (black line).

distribution, $\Gamma(\kappa,\theta)$, with κ and θ being the shape and the scale parameters, and a log-normal distribution, $\ln\mathcal{N}(\mu,\sigma^2)$, with μ and σ^2 being the mean and the variance of the count rate's (variable x) natural logarithm (equation 3). Finally, each of these component distributions contributes to the overall PDF of the mixture distribution, $\mathfrak{f}_{\rm mix}(x;\boldsymbol{\eta})$, with a weight of w_Γ and $w_{\ln\mathcal{N}} = 1 - w_\Gamma$, respectively, with $\boldsymbol{\eta}$ being a vector consisting of the model parameters $\boldsymbol{\eta} = \{\kappa,\theta,\mu,\sigma,w_\Gamma\}$:

$$\mathfrak{f}_{\rm mix}(x;\boldsymbol{\eta}) = w_\Gamma\frac{\theta^{-\kappa}\mathrm{e}^{-x/\theta}x^{\kappa-1}}{\Gamma(\kappa)} + w_{\ln\mathcal{N}}\frac{\mathrm{e}^{-(\ln x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi}x\sigma}. \quad (3)$$

The best-fitting PDF model is shown superimposed on the data sample histogram in the left-hand panel of Fig. 4 (grey line) with best-fitting model parameters of $\boldsymbol{\eta}_{\rm bf} = \{5.67^{+0.04}_{-0.03}, 5.96^{+0.06}_{-0.04}, 2.14 \pm 0.06, 0.31^{+0.05}_{-0.04}, 0.82^{+0.05}_{-0.04}\}$. Having as a null hypothesis, H_0, that the data set is drawn from the derived best-fitting model distribution and an alternative hypothesis, $H_{\rm a}$, that it was not drawn from that distribution, the Anderson–Darling test (Anderson & Darling 1952) yields a statistic value of 0.34 corresponding to an H_0 probability of 0.89 which depicts the good representation of the data by the given model.

Having defined the best-fitting PSD and PDF we continue to the actual production of the artificial light curves.

(i) The best-fitting PSD model with $c = 0$ is then used to create the normally distributed time series with a periodogram, $P_{\rm norm}(f_j)$. A simulated light curve of this kind together with its periodogram is shown in Fig. 5 (left- and right-hand panels, respectively). By

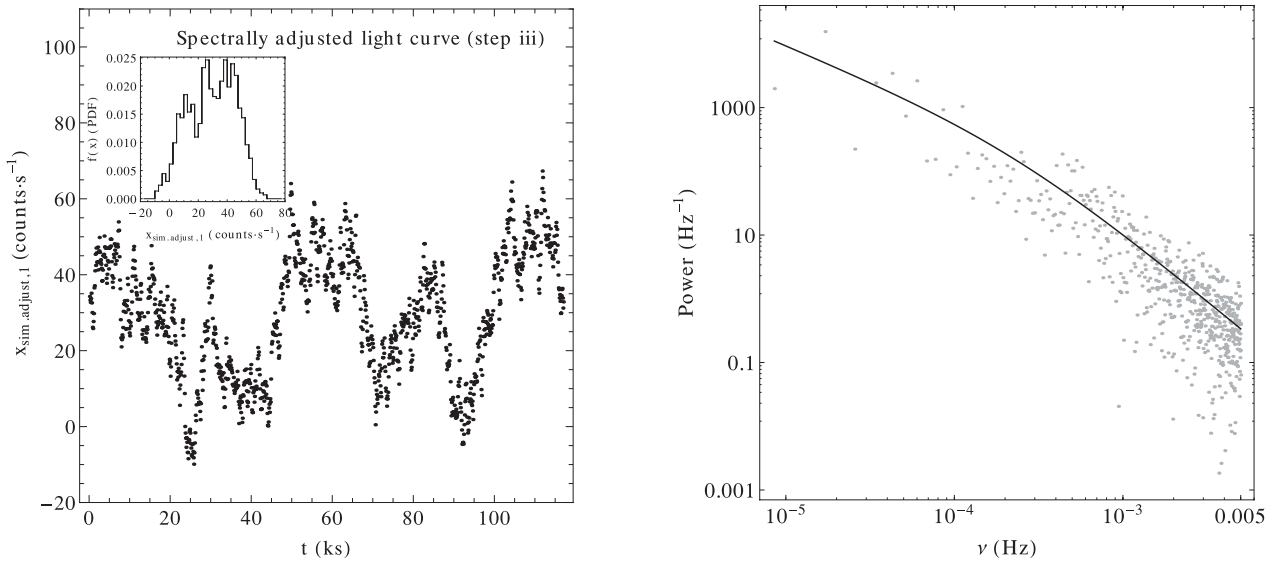

Figure 6. (Step iii) Left-hand panel: the spectrally adjusted light curve together with its PDF histogram (inset). Right-hand panel: the corresponding periodogram estimates (grey points) are by construction identical to those shown in the right-hand panel of Fig. 5, and the underlying target PSD, $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)$ (black line).

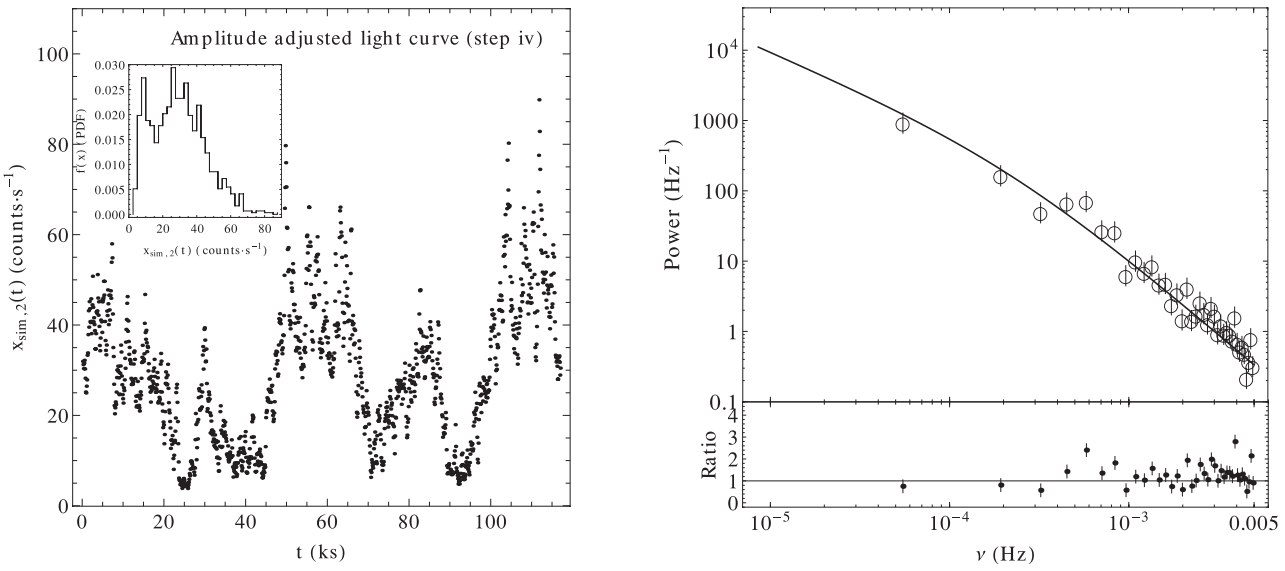

Figure 7. (Step iv) Left-hand panel: the amplitude adjusted light curve together with its PDF histogram (inset). Right-hand panel: the corresponding binned logarithmic periodogram estimates (open circles) and the underlying target PSD model, $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)$ (black line), having attached in the bottom of the ratio plot, data/model.

estimating the DFT of $x_{\rm norm}(t)$, for each Fourier frequency, f_j, we derive the corresponding amplitudes and phases, $\mathscr{A}_{\rm norm}(j)$ and $\phi_{\rm norm}(j)$, respectively.

(ii) The best-fitting PDF model is used to generate a list of N pseudo-random variates, $x_{\rm sim,\,1}(t)$, shown in the right-hand panel of Fig. 4. Then, by estimating for each Fourier frequency the DFT of $x_{\rm sim,\,1}(t)$, $DFT_{\rm sim,1}(j)$, we derive the corresponding amplitude and phases, $\mathscr{A}_{\rm sim,1}(j)$ and $\phi_{\rm sim,\,1}(j)$, respectively.

(iii) *Spectral adjustments*. $\mathscr{A}_{\rm sim,1}(j)$ are replaced by $\mathscr{A}_{\rm norm}(j)$, keeping the $\phi_{\rm sim,\,1}(j)$ unaltered, yielding an adjusted version of $DFT_{\rm sim,1}(j)$, $DFT_{\rm sim.adjust,1}(j)$. By performing an IDFT, we obtain the light curve $x_{\rm sim.adjust,\,1}(t)$ (Fig. 6, left-hand panel) with an identical periodogram to $P_{\rm norm}(f)$ (Fig. 6, right-hand panel, grey points), but now with measurements which are not longer distributed as $\mathfrak{f}_{\rm mix}(x;\boldsymbol{\eta}_{\rm bf})$ (Fig. 6, left-hand panel, inset).

(iv) *Amplitude adjustments*. Finally, the values of $x_{\rm sim.adjust,\,1}(t)$ are replaced by the values of $x_{\rm sim,\,1}(t)$, based on the ranking of the former. The resulting light curve, $x_{\rm sim,\,2}(t)$ (Fig. 7, left-hand panel), has an identical histogram with the sample light curve, but this time the periodogram estimates do not correspond to the target underlying PSD, $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)$. The right-hand panel of Fig. 7 shows the binned logarithmic periodogram estimates (in bins of 15 consecutive periodogram estimates) of the amplitude adjusted light curve together with $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)$ and the ratio plot, i.e. data/model. From the ratio plot it is obvious that particularly the high-frequency periodogram estimates, in particular above 10^{-3} Hz, are systematically

larger than the corresponding $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)$ values by a factor of 1.5.

All the above-mentioned procedure is a single iteration step of the method. Exactly the same process is repeated iteratively from step (iii), by replacing $x_{\rm sim,\,1}(t)$ with the amplitude adjusted light curve $x_{\rm sim,\,2}(t)$, $x_{\rm sim,\,2}(t)$ with $x_{\rm sim,\,3}(t)$, and so on. For this particular case, after the 55th iteration the synthetic light curves remain the same.

3.1.1 Convergence of a single artificial light curve

In order to check the convergence of the method, we fit the bending power-law model (equation 2) to the corresponding periodogram estimates of each iteration step. The resulting values for the $\alpha_{\rm low}$, $\alpha_{\rm high}$ and $f_{\rm bend}$ are shown in Fig. 8[7] which we can see form a plateau after the 55th iteration. Fig. 9 shows the corresponding results for the 56th iteration; the synthetic light curve (left-hand panel) follows the exact distribution of the observed data, and the binned logarithmic periodogram estimates (in bins of 15 consecutive periodogram estimates) as expected follow, within the 68.3 per cent confidence levels depicted by the error bars, the corresponding $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)$. This means that the 56th synthetic light curve is the final simulated data set, with all the desired statistical and variability properties of the original data set. The best-fitting PSD model for the 56th surrogate yields $\alpha_{\rm low}=1.36^{+0.22}_{-0.38}$, $\alpha_{\rm high}=2.24^{+0.08}_{-0.05}$ and $f_{\rm bend}=3.4^{+1.1}_{-1.3}\times10^{-4}$ Hz, and the resulting histogram is by construction identical to the original one since it is drawn from its best-fitting PDF model (Fig. 4, right-hand panel).

Finally, in order to take into account the Poisson statistics (Section 2.2) we re-sample the 56th surrogate data set according to equation (1). The resulting artificial light curve is shown in Fig. 10. This single random synthetic data set encloses all the information of our initial data set and thus can be used in any sort of statistical study.

3.2 Overall procedure for an ensemble of realizations

3.2.1 Proposed methodology

In this section we repeat the above-mentioned procedure for an ensemble of 1000 realizations and we compare the statistical properties of the final products to those of the original data set of NGC 4051, i.e. the light curve and underlying PSD (Fig. 3). Initially, we perform a goodness-of-fit Kolmogorov–Smirnov hypothesis test (Press et al. 1992) for the distribution of each artificial light curve, with H_0 that the surrogate data set is drawn from the best-fitting model distribution of NGC 4051 (Fig. 4, left-hand panel) and $H_{\rm a}$ that it was not drawn from that distribution. The mean Kolmogorov–Smirnov statistic derived from the ensemble of light curves is $D_n=0.025^{+0.008}_{-0.006}$ and the mean H_0 probability derived is $0.51^{+0.28}_{-0.22}$, depicting the high degree of accordance between the distribution of the artificial data sets and that of the original data set (the error estimates correspond to the 68 per cent confidence intervals). Thus, this method assures that the resulting simulated data sets have the same statistical moments as the observed light curve of NGC 4051.

We then fit the PSD model of equation (2) to the periodogram estimates of each artificial light curve. The distributions of both the low- and the high-frequency PSD slopes, as well as the bending frequency, are shown in the left- and right-hand panels of Fig. 11, respectively. The sample mean values, together with their 68.3 per cent confidence limits (i.e. standard deviation of the sample mean) and the 68.3 per cent confidence intervals of the distributions for $\alpha_{\rm low}$, $\alpha_{\rm high}$ and $f_{\rm bend}$, are given in Table 1. The simulation results, which come from the proposed method, are entirely consistent with those derived from the original data set of NGC 4051, indicating that there are no biases towards the PSD model parameters which could cause systematic deviations from the targeted values. Thus, the artificial light curves produced as an ensemble with this algorithm have the same variability power, as a function of Fourier frequency, as that of NGC 4051.

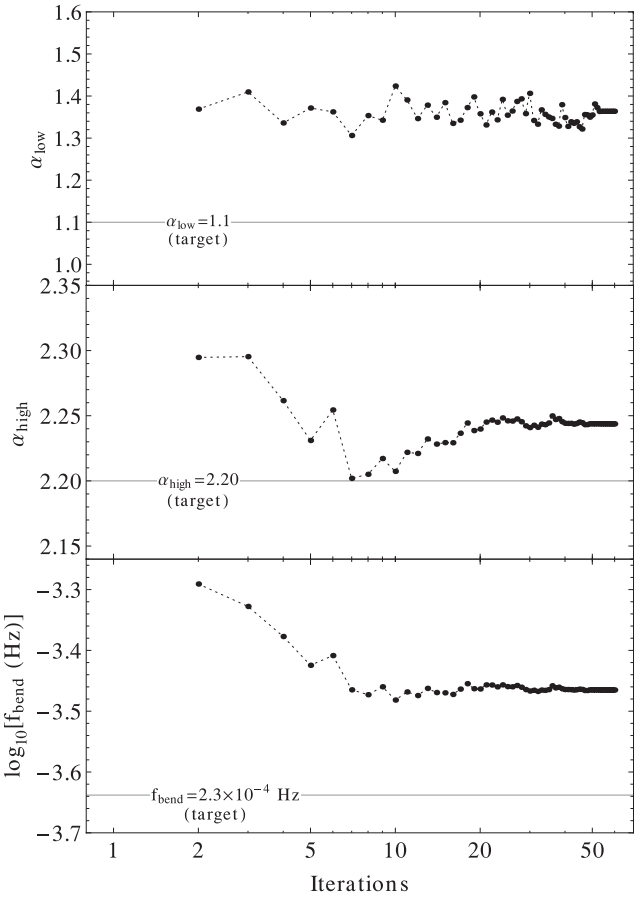

Figure 8. Convergence of the iteration process (the first iteration is not shown; see footnote 7). The horizontal solid grey lines indicate the corresponding target values, i.e. the best-fitting values as derived from the observed data. The dotted line among the various iterations shows a linear interpolation intended only to guide the eye. Top panel: the $\alpha_{\rm low}$ estimates as a function of the iteration number (in logarithmic scale), stabilizing at 1.36. Middle panel: the $\alpha_{\rm high}$ estimates as a function of the iteration number (in logarithmic scale), stabilizing at 2.24. Bottom panel: the logarithm of $f_{\rm bend}$ estimates as a function of the iteration number (in logarithmic scale), stabilizing at −3.47.

[7] The results of the first iteration are excluded from the panels in order to cover better the variations of the other iterated products. The omitted values are $\alpha_{\rm low}=2.56$, $\alpha_{\rm high}=3.21$ and $\log_{10}[f_{\rm bend}\,({\rm Hz})]=-1.90$), respectively. These large deviations result from the fact that the minimization routine does not localize the minimum (after 500 iteration steps) for the initial periodogram estimates, $P_{\rm sim,\,1}(f)$, corresponding to a flat PSD (see footnote 4). Naturally, by increasing the number of iterations we can correct for this artefact, but in this context it is unnecessary since for the next steps the localization of the minimum occurs in less than 30 iterations as the degeneracy $\alpha_{\rm high}=\alpha_{\rm low}\simeq0$ does not exist any more.

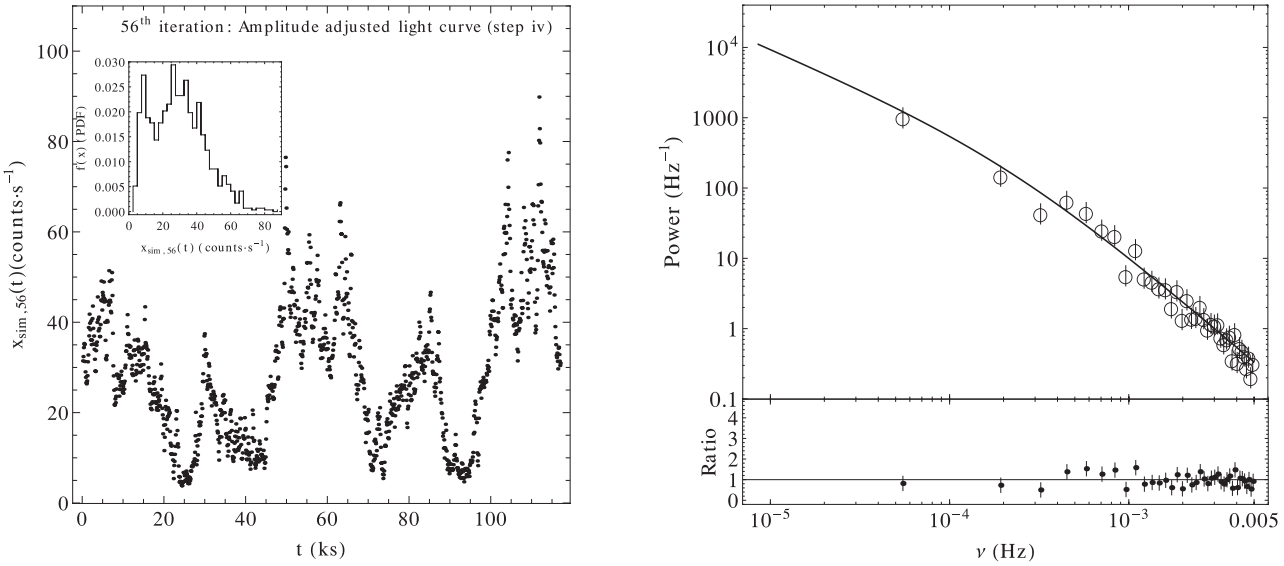

Figure 9. Step iv for the 56th iteration. Left-hand panel: the amplitude adjusted light curve together with its PDF histogram (inset). Right-hand panel: the corresponding binned logarithmic periodogram estimates (open circles) and the underlying target PSD model, $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)$ (black line), having attached in the bottom of the corresponding ratio plot, data/model.

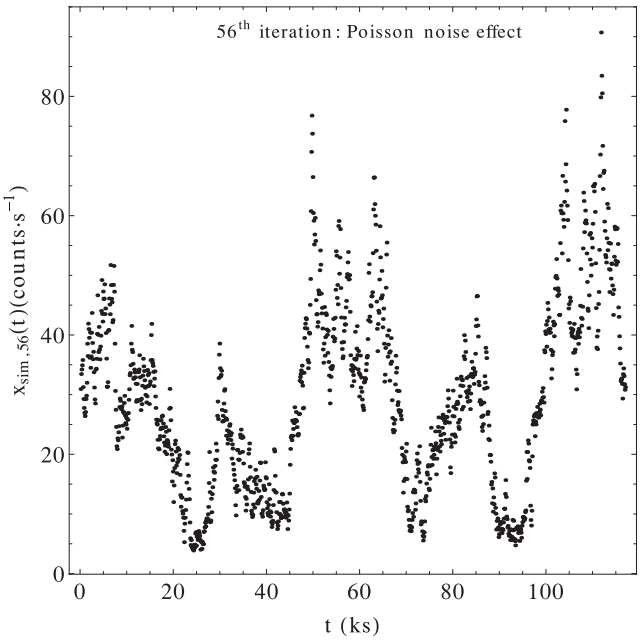

Figure 10. The 56th surrogate data set re-sampled from a Poisson distribution as dictated by equation (1) ($\Delta t = 100$ s).

3.2.2 *Convergence of the ensemble of artificial light curves*

The scattering in the various estimated PSD model parameters, coming from the 1000 simulated light curves, originates from the asymptotic distribution of the various periodogram estimates, $P(f_j)$, around the input PSD, $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)$. As we discussed in Section 2.4 (see for details Appendix A2) at a given Fourier frequency f_j, $P(f_j)$ is distributed asymptotically around $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)$ as a gamma distribution, $\Gamma[\nu/2, \mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)]$ with ν d.o.f. This behaviour is depicted in the left-hand panel of Fig. 12 which shows the distribution of the 1000 periodograms around the target PSD. As a sanity check for our simulations, we test whether for a given Fourier frequency, f_j, the distribution of their various periodogram estimates is indeed in accordance with equation (A9). Thus, we derive for each f_j the distribution of points and we perform an Anderson–Darling test goodness-of-fit test with H_0: the periodogram estimates at a given f_j are drawn from the $\Gamma[\nu/2, \mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf},0)]$ distribution and $H_{\rm a}$ that they are not drawn from this distribution. The mean value of the statistic is $2.92^{+0.24}_{-0.13}$ yielding a mean H_0 probability $0.18^{+0.09}_{-0.06}$ depicting the high degree of accordance between the estimated distribution of the simulated products and the expected ones (the error estimates correspond to the 68 per cent confidence intervals).

Finally, depending on the type of the statistical study, it is not necessary always for each surrogate data set to carry out the iteration process up to the convergence point (as shown in Fig. 8). Stopping the process in an intermediate step, e.g. at the 5th iteration step, will yield surrogate data sets which will still have accurate PSD parameters (i.e. they will be distributed correctly around the target values without systematic trends), but the various estimates will be less precise than those derived from the final converged products (i.e. they will exhibit larger scatter around the target values). Nevertheless, the differences are very small and for this particular example (5th iteration step) are on average of the order of 5 per cent. In the right-hand panel of Fig. 12, we show this effect by plotting the convergence in the PSD parameter $\alpha_{\rm high}$ for 15 synthetic data sets (as we did in the top panel of Fig. 8). The publicly available MATHEMATICA notebook (Section 2.5) contains an animation showing these small differences between all the iteration steps for a single surrogate.

3.2.3 *Exponential light curves*

In this section we follow the recipe of Uttley et al. (2005) and exponentiate (in base e) the TK95 products which are produced by the renormalized PSD model (using equations 13 and 14 in Uttley et al. 2005). In this case, the artificial data sets always follow by construction a log-normal distribution which differs intrinsically from the observed statistical properties (described by equation 3) which we are interested in reproducing for the given illustrative purposes. Thus, we do not need to perform a goodness-of-fit

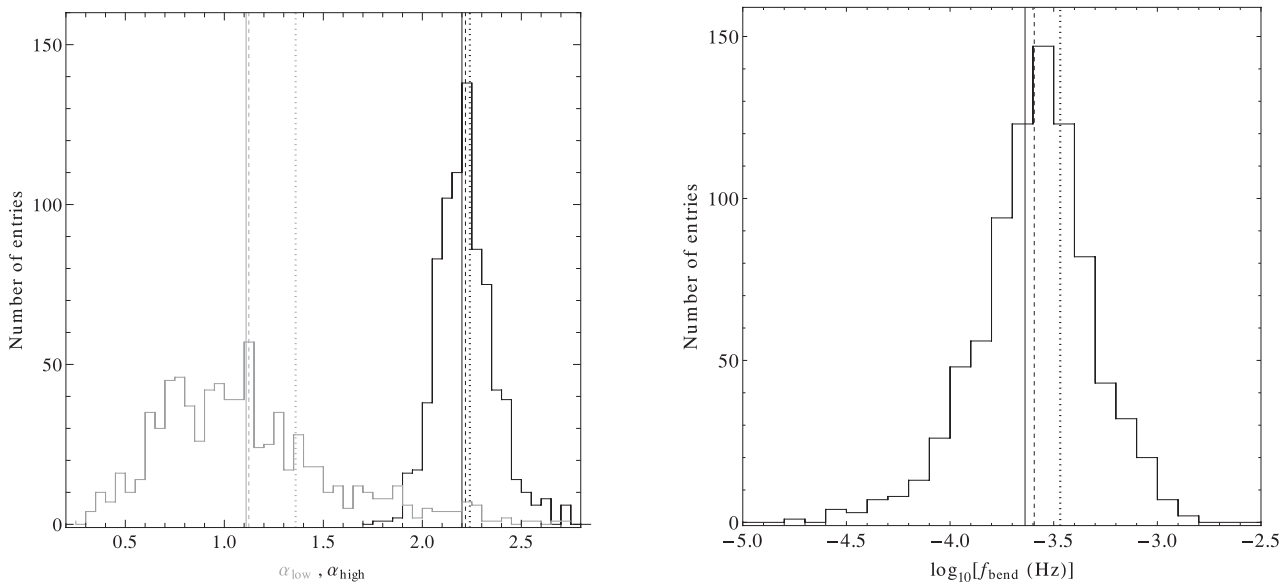

Figure 11. Overall simulation results for 1000 artificial light curves. Left-hand panel: the histogram of the best-fitting $\alpha_{\rm low}$ (grey lines) and $\alpha_{\rm high}$ (black lines). The solid lines correspond to the target values coming directly from the observed data, 1.1 and 2.2, respectively, the dashed line corresponds to the mean estimate of the distribution, 1.12 and 2.21 respectively, and the dotted line corresponds to the best-fitting value as derived from the 56th surrogate of the single realization, 1.36 and 2.24, respectively (Fig. 8, top and middle panels, respectively). Right-hand panel: the histogram of the best-fitting logarithms of $f_{\rm bend}$. The solid line corresponds to the target value coming directly from the observed data $f_{\rm bend} = 2.3 \times 10^{-4}$ Hz (or -3.64 in log scale with base 10), the dashed line corresponds to the mean estimate of the distribution, 2.4×10^{-4} Hz (or -3.62 in log scale with base 10) and the dotted line corresponds to the best-fitting value as derived from the 56th surrogate of the single realization, 3.4×10^{-4} Hz (or -3.47 in log scale with base 10, Fig. 8, bottom panel).

Table 1. Global simulation results for the PSD parameters.

Model parameter	Target values[a]	Proposed method[b]	Exponential function method[b]
$\alpha_{\rm low}$	1.1 (fixed)	$1.123^{+0.002}_{-0.007}$ [0.87, 1.20]	$0.973^{+0.004}_{-0.006}$ [0.86, 1.03]
$\alpha_{\rm high}$	$2.20^{+0.07}_{-0.04}$	$2.213^{+0.002}_{-0.001}$ [2.15, 2.26]	2.063 ± 0.002 [2.00, 2.10]
$f_{\rm bend}$ ($\times 10^{-4}$ Hz)	$2.3^{+1.2}_{-0.9}$	2.4 ± 0.1 [2.1, 3.3]	3.9 ± 0.1 [3.2, 5.0]

[a]These are the values of NGC 4051 derived in Section 3.1.
[b]The first value is the sample mean together with its 68.3 per cent confidence limits, and the second value in the square brackets corresponds to the 68.3 per cent confidence intervals of the distribution around the mean.

hypothesis test for the distributions, since we know *a priori* that they are by construction different.

In the next step, we repeat the PSD model fitting procedure for the periodogram estimates of the exponential light curves. The results for the distribution of the best-fitting parameters of $\alpha_{\rm low}$, $\alpha_{\rm high}$ and $f_{\rm bend}$ are shown in Fig. 13. Finally, as above, the sample mean values together with their 68.3 per cent confidence limits (i.e. the standard deviation of the sample mean) and the 68.3 per cent confidence intervals of the distributions for $\alpha_{\rm low}$, $\alpha_{\rm high}$ and $f_{\rm bend}$ are given in Table 1. We can see that the PSD becomes systematically softer by around 8 per cent something which is also shown in fig. B1 in Uttley et al. (2005). Most importantly for this particular case the most noticeable distortion appears in the bend frequency which systematically shifts towards higher frequencies, deviating in this way by 70 per cent from the target value.

Using these simulated data sets for the recovery of the bend frequency of an irregularly sampled light curve (using the procedure of Uttley et al. 2002) will yield systematic deviations from the true underlying value. Note that the degree of the various PSD distortions of the exponential light curves depends on the particular variability properties of the light curves, as well as the actual values of the underlying PSD model.

A potential solution to these spectral alterations could be the following: to consider the logarithm of the observed data set (which is Gaussian distributed for the case of a log-normal distribution) and estimate its PSD which is then going to be used as the input PSD for the TK95 simulation. The exponentially transformed TK95 products should follow the original PSD of the observed data set which is log-normally distributed. Before following this recipe, further investigation of this approach should be carried out, something which is out of the scope of this paper.

4 COMPLIMENTARY APPLICATIONS: *FERMI* AND *RXTE* DATA SETS

In order to show the wide applicability of our newly proposed method, we further apply it to two radically different looking data sets: a γ-ray *Fermi*-LAT data set for the blazar 3C 454.3, an X-ray *RXTE* data set for the XRB Cyg X-1.

4.1 The γ-ray blazar 3C 454.3

We use the weekly *Fermi*-LAT light curve of 3C 454.3 consisting of 236 points between 546 84 and 563 34 MJD in the 0.1–300 GeV

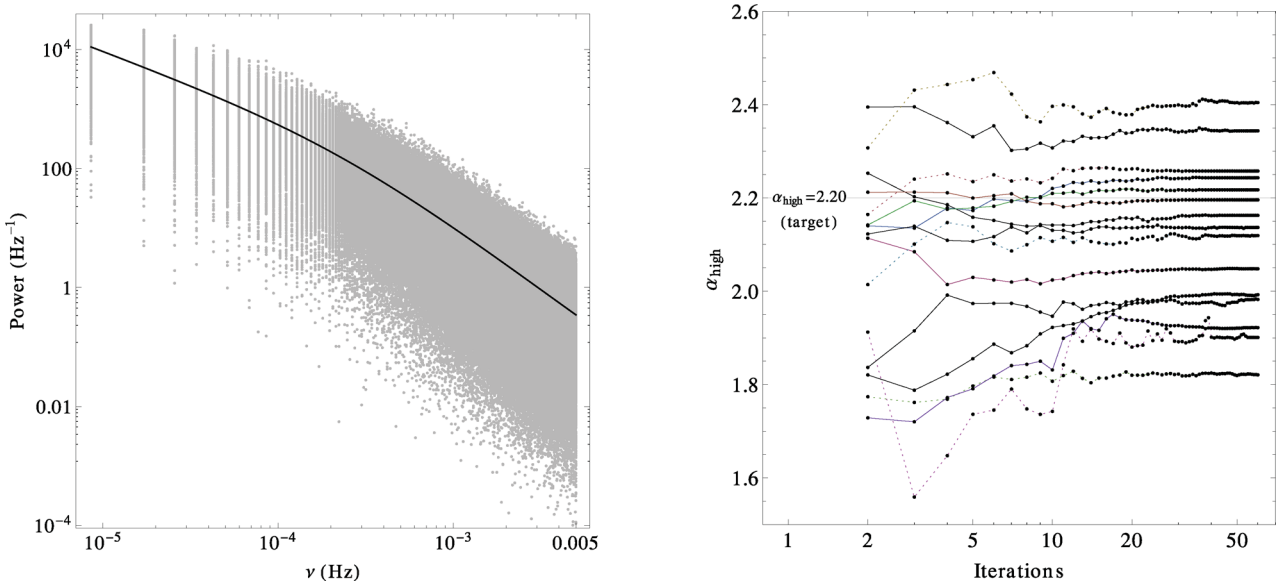

Figure 12. Distribution and convergence of the ensemble periodogram estimates. Left-hand panel: the distribution of the 1000 periodograms (grey points), originating from the 1000 synthetic light curves, around the underlying PSD, $\mathscr{P}(f;\boldsymbol{\gamma}_{\rm bf}, c_{\rm bf})$ (black, solid line). Note that the synthetic light curves contain Poisson noise. Right-hand panel: convergence of the PSD parameter $\alpha_{\rm high}$ as estimated from fitting the periodogram estimates of the 15 synthetic data sets for all the iteration steps.

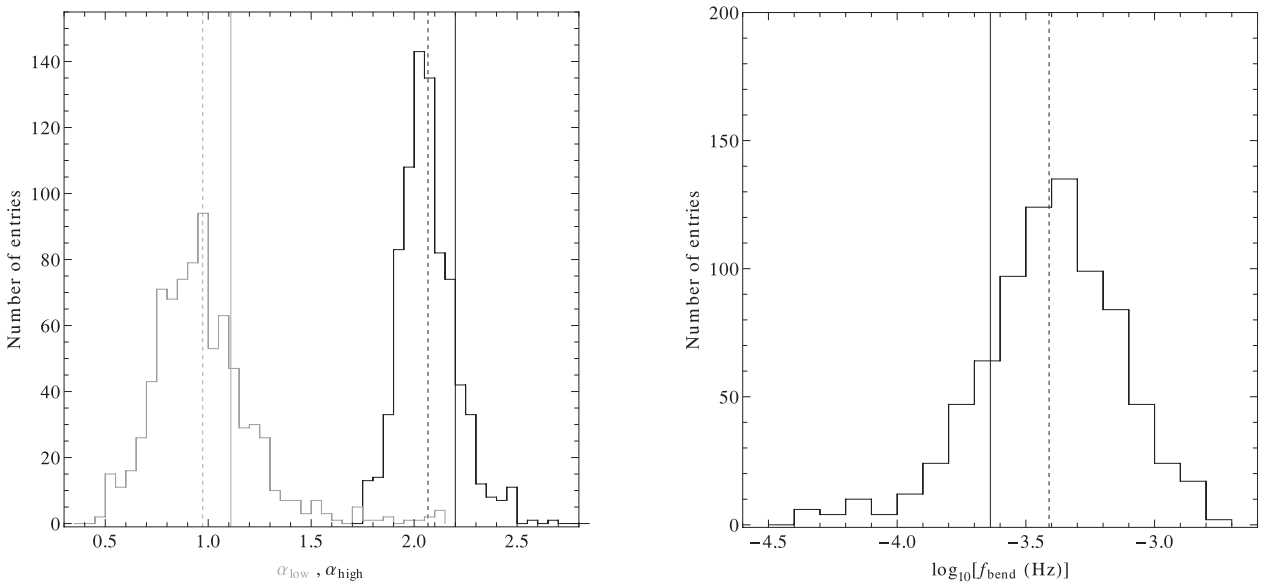

Figure 13. Overall results for 1000 exponential light curves. Left-hand panel: the histogram of the best-fitting $\alpha_{\rm low}$ (grey lines) and $\alpha_{\rm high}$ (black lines). The solid lines correspond to the target values coming directly from the observed data, 1.1 and 2.2, respectively, the dashed line corresponds to the mean estimate of the distribution, 0.97 and 2.06, respectively. Right-hand panel: the histogram of the best-fitting logarithms of $f_{\rm bend}$. The solid line corresponds to the target value coming directly from the observed data $f_{\rm bend} = 2.3 \times 10^{-4}$ Hz (or -3.64 in log scale with base 10) and the dashed line corresponds to the mean estimate of the distribution, 3.9×10^{-4} Hz (or -3.41 in log scale with base 10).

energy range.[8] The light curve is shown in the left-hand panel of Fig. 14 with the black points corresponding to the actual flux measurements and the grey points (around 20 and 90–130 Ms) to the 90 per cent confidence upper limits. For the purposes of this study, we have simply used the upper limits as actual flux measurements but more precise treatment using survival analysis techniques will be presented in a future work.

To remind the readers that the two basic components for the method are the distribution of the data and the corresponding PSD. The PDF histogram of the data is shown in the left-hand panel of Fig. 14 (left inset) and as we can see it is characterized by a long right tail which becomes zero at much higher flux values from those expected by a simple exponential distribution. Note that if we were about to fit an exponential distribution PDF model to this data set, it would yield a best-fitting inverse scale of 4.25 ± 0.06

[8] The *Fermi*-LAT data have been retrieved from http://fermi.gsfc.nasa.gov/ssc/data/access/lat/msl_lc/.

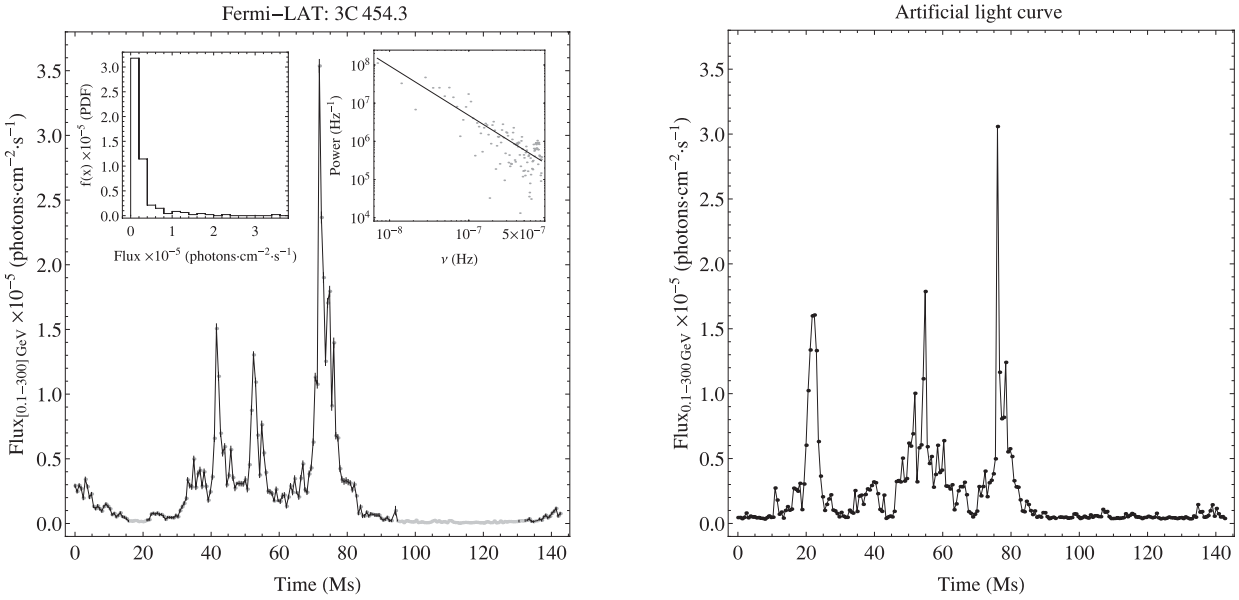

Figure 14. The *Fermi*-LAT data set of the blazar 3C 454.3. Left-hand panel: the weekly averaged γ-ray light curve in the energy range of 0.1–300 GeV with black points corresponding to actual flux measurements and grey points to the 90 per cent confidence upper limits. The left and right insets show the PDF histogram and the periodogram estimates (grey points) together with the best-fitting power-law model (black line), respectively. Right-hand panel: a single artificial light curve coming after 22 iterations (convergence).

having a very poor fit quality, with the Anderson–Darling test statistic value of 30.21 and an H_0 probability (i.e. the data set is drawn from a population with the fitted distribution) of 0. The right inset in the same plot shows the periodogram estimates of the light curve together with the best-fitting bending power-law model (equation 2), $\alpha_{\rm low} = \alpha_{\rm high} = 1.3^{+0.4}_{-0.1}$ and $c = 2.4 \pm 0.3$ Hz^{-1} ($f_{\rm bend} = 7.8^{+2.1}_{-1.9} \times 10^{-8}$ Hz), implying that a simple power-law model is enough to describe the data. Using these two data components, we apply our method and we produce an artificial light curve (convergence occurs after 22 iterations) which conserves all the statistical and variability properties of the original data set (Fig. 14, right-hand panel). Thus, ensembles of such artificial light curves can be used in any sort of statistical analysis that requires establishment of confidence intervals e.g. in cross correlation function (CCF) analysis during multiwavelength campaigns.

4.2 The XRB Cyg X-1

We use the daily averaged *RXTE*-ASM light curve of Cyg X-1 consisting of 5000 points between 501 35.4 and 554 22.6 MJD in the 2–10 keV energy range.[9] The light curve is shown in the left-hand panel of Fig. 15 and contains a small number of gaps (289 in total), which we have filled up with linearly interpolated values. Appropriate treatment, using bootstrapping, should be performed in order to check the effects of the gaps during the following PSD estimation, but for the purposes of this study we ignore this step.[10]

The PDF histogram of the data is shown in the left-hand panel of Fig. 15 (left inset) having a characteristic bimodal shape, depicting the high and the low flux states of the source. Assuming that no artefacts are induced to the periodogram estimates, due to the interpolation, the best-fitting PSD model yields $\alpha_{\rm low} = 0.49^{+0.12}_{-0.21}$, $\alpha_{\rm high} = 1.58^{+0.14}_{-0.16}$, $f_{\rm bend} = 1.32^{+2.6}_{-0.43} \times 10^{-8}$ Hz and $c = 3692^{+18}_{-12}$ Hz^{-1} (Fig. 15, left-hand panel, right inset). After applying our method (convergence occurs after 267 iterations), the resulting artificial light curve (Fig. 15, right-hand panel) resembles remarkably the original data set of Cyg X-1 which was chosen as an example of extreme 'bursticity'.

5 APPLICATION TO CCF ANALYSIS

CCF analysis is one the most common method used for analysing multiwavelength light curves obtained in a simultaneous fashion. There are several different flavours and implementations of the CCF, e.g. discrete correlation function (DCF; Edelson & Krolik 1988), interpolated CCF (Gaskell & Sparke 1986), modified CCF (Li et al. 2004) and z-transform DCF (Alexander 1997) that are used within the astronomical community. Particularly for the case of irregularly sampled light curves, estimation of the confidence levels, in both the CCF values and/or time delays, is usually done by performing Monte Carlo simulations. Application of a given CCF method to an ensemble of paired random artificial light curves, which have the same PSD as the observed data sets, yields the probability of getting a given CCF estimate purely by chance coincidence. At present, the simulated light curves are produced using the TK95 formalism.

Since the TK95 synthetic data sets are distributed normally by construction, the method is appropriate to yield CCF confidence levels only for Gaussian light curves. In order to show that deviations in the CCF levels can occur for the case of non-Gaussian light curves, for illustrative purposes, we create two 'bursty' light curves in the following way. Using the TK95 procedure, we produce two artificial light curves having different bending PSDs, $x_{\rm G}(t)$ and

[9] The *RXTE*-ASM data have been retrieved from http://xte.mit.edu/ASM_lc.html.

[10] The frequency domain bootstrap methodologies are used to estimate the distribution of the re-sampled periodogram estimates. The main difficulty is to select appropriate statistical estimators whose variance fits that of the re-sampled periodogram estimates (at a given frequency). Useful analyses on this topic have been performed by several authors (e.g. Franke & Härdle 1992; Dahlhaus & Janas 1996; Kreiss & Paparoditis 2003).

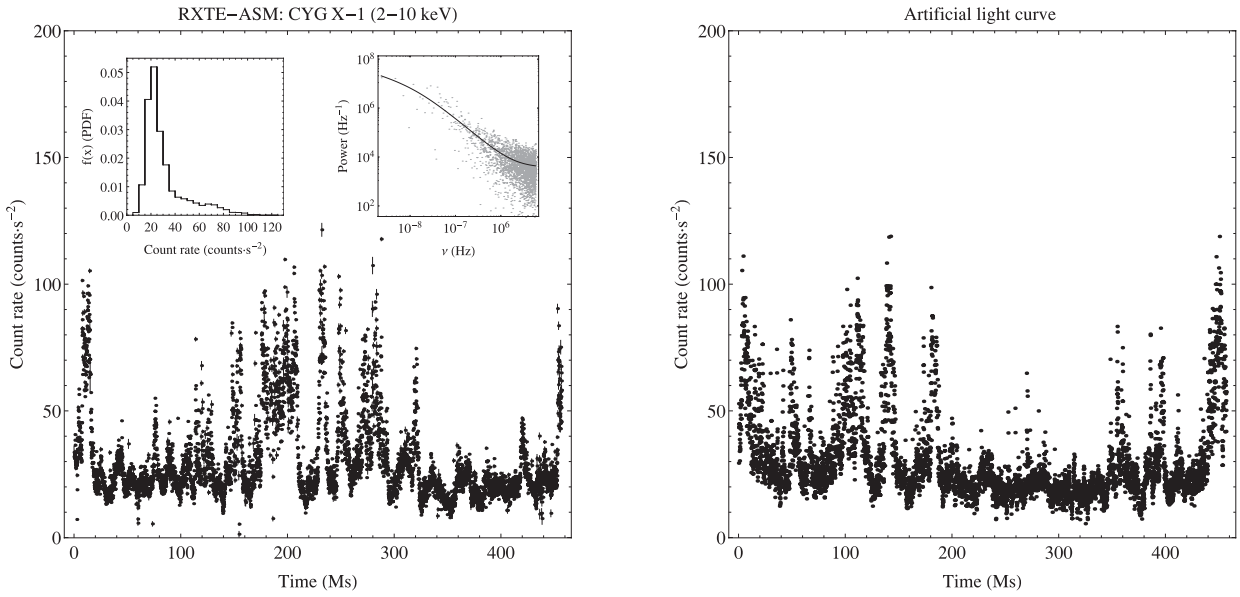

Figure 15. The *RXTE*-ASM data set of the XRB Cyg X-1. Left-hand panel: the daily averaged X-ray light curve in the energy range of 2–10 keV. The left and right insets show the PDF histogram and the periodogram estimates (grey points) together with the best-fitting bending power-law model (black line), respectively. Right-hand panel: a single artificial light curve coming after 267 iterations (convergence).

$y_G(t)$, 200 ks long with a bin size $\Delta t = 100$ s, using the same random seed (in order to be correlated). During the production of $y_G(t)$, we multiply the imaginary part with a random number in the range $[-0.15, 0)$. This will slightly modify the final flare profiles, i.e. amplitudes and phases (equations A2 and A3) yielding an asymmetric CCF profile around the zero time delay. Finally, the resulting normally distributed numbers are used as exponents for the bases 2 and 1.5, respectively, in order to produce two 'bursty-like' light curves, $x(t)$ and $y(t)$ (Fig. 16, left-hand panel), having of course different underlying PSD parameters from the initial ones. The initial PSD parameters, used for the TK95 methodology, together with the final PSD parameters, estimated after fitting the periodogram estimates of the exponentiated light curves (in bases 2 and 1.5, respectively), are given in Table 2. Note that in both light curves we have added Poisson noise following the recipe described in Section 2.2. For the purposes of this study, these two 'bursty' light curves will be used as two simultaneously obtained observations of the same object, in different energy bands, for which we will perform CCF analysis.

Initially, we estimate the DCF for the two 'bursty' light curves (Fig. 16, right-hand panel, black points). Then, in order to assess the confidence level of the correlation, we produce two ensembles of 1000 pairs of artificial light curves: one following the classical

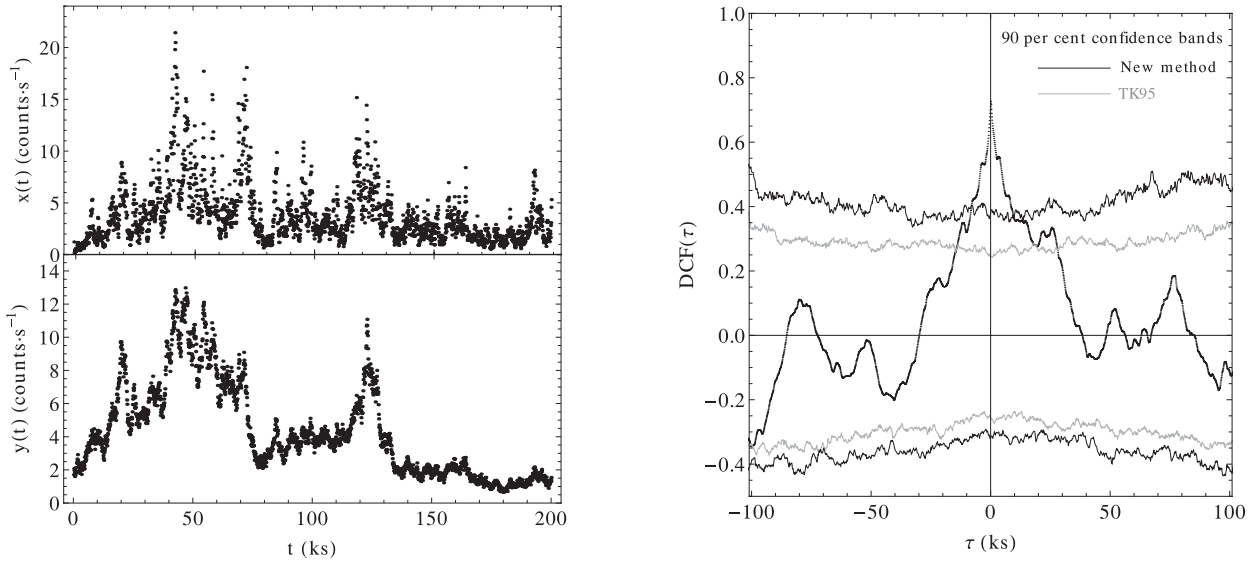

Figure 16. Establishment of the statistical significance on the DCF estimates. Left-hand panel: the two artificially produced 'bursty' light curves $x(t)$ and $y(t)$ having the same random seed. Right-hand panel: the black points correspond to the DCF estimates of $x(t)$ and $y(t)$, separated by 0.1 ks (100 s). The grey and black horizontal lines correspond to the 90 per cent confidence bands as derived from the synthetic data sets of TK95 and the newly proposed method, respectively.

Table 2. PSD model parameters for the CCF simulations.

Model parameter	Initial PSD $x_G(t)$, $y_G(t)$	Final PSD[a] $x(t)$,$y(t)$
α_{low}	0.9, 1.8	$0.91^{+0.09}_{-0.08}$, $0.96^{+0.08}_{-0.07}$
α_{high}	2.3, 2.8	$2.58^{+0.08}_{-0.05}$, 2.67 ± 0.06
f_{bend} ($\times 10^{-4}$ Hz)	2.6, 10	14^{+1}_{-2}, 18^{+3}_{-2}

[a] These are the values used in the simulations.

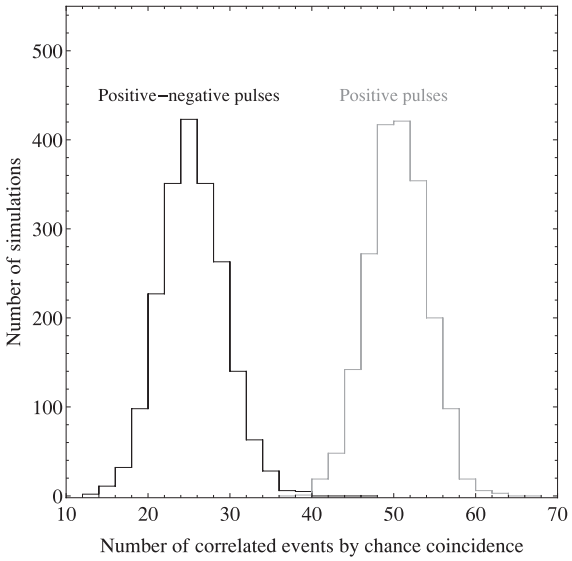

Figure 17. Correlated events for the toy simulation. The black line corresponds to the distribution of the correlated events for the case of positive–negative triangular unit pulses. The grey line corresponds to the same distribution for the case of positive triangular unit pulses.

procedure of TK95 and another one using the proposed methodology described in Section 2. Then, for each method we estimate the DCFs between all the pairs, and for each time delay, τ, we estimate the 0.025 and 0.975 quantiles corresponding to the upper and lower limits of the 90 per cent confidence bands.

As we can see from the right-hand panel of Fig. 16 (TK95: grey lines, new method: black lines), a realistic representation of the non-Gaussian light curves yields in general an increase in the confidence level range of the order of 25 per cent which reduces the detection significance of the DCF peak. The reason is that Gaussian light curves have on average the same number of flares above (positive direction) and below (negative direction) the mean, in contrast to the 'bursty' light curves which exhibit flares only in the positive direction. This means that between two 'bursty' light curves it is much more likely to get a fake correlation by chance coincidence since there is only one possible flare direction.

This can be very well understood with the following toy simulation. We produce a series of 2000 pairs of time series each one consisting of 100 positive and negative triangular positive pulses (the simplified analogue of a Gaussian light curve) occurring at uniformly random non-repetitive integer numbers, and we measure the number of simultaneous pulse occurrences by chance coincidence between all the pairs. Then, we repeat the simulation but now the time series consists only of positive triangular unit pulses (the simplified analogue of a 'bursty' light curves). As we can see from Fig. 17 the mean occurrence of chance coincidence correlated events is almost doubled for the case of the positive triangular pulses due to the flare directionality property. This shows that, for non-Gaussian light curves, the TK95 procedure underestimates the chance coincidence occurrences of correlated events and thus yields erroneous smaller estimates for the confidence intervals, i.e. yielding an overestimation of the CCF's peak significance.

6 THE RMS–FLUX RELATION

In the special case of a parent log-normal distribution, the rms–flux relation (see Section 1) may be sometimes of vital importance for the needs of a statistical study or a theoretical model. The surrogate data sets, following a parent log-normal distribution, have embedded this property in a natural way without the need for further adjustments or tuning.

To show that our method automatically produces the rms–flux relationship, we first create a sample light curve which inherently has the rms–flux relation by following a parent log-normal distribution. Using the TK95 procedure (with initial PSD parameters: $\alpha_{low} = 1.5$, $\alpha_{high} = 2.8$ and $f_{bend} = 1 \times 10^{-3}$ Hz), we produce a synthetic data set, being 200 ks long in bins of 100 s, and then we exponentiate the resultant data set [having final PSD parameters: $\alpha_{low} = 1.11^{+0.05}_{-0.04}$, $\alpha_{high} = 2.48^{+0.09}_{-0.06}$ and $f_{bend} = (2.2 \pm 0.6) \times 10^{-4}$ Hz (Fig. 18, left-hand panel)]. This light curve will be used as the observed light curve that we want to simulate.

We produce 1000 artificial light curves using our proposed method and then for each one of them we estimate the rms–flux relation using the prescription of Uttley et al. (2005). We select three different length segments of 0.5, 1.5 and 5 ks consisting of 5, 15 and 50 bins, respectively. Under a given binning scheme, for each flux value we estimate an average rms and its standard deviation coming from the 1000 surrogate data sets. The results are shown in the right-hand panel of Fig. 18 and as we can readily see the simulated light curves follow remarkably well the linear rms–flux relation for a variety of time-scales below and above the f_{bend}, corresponding approximately to 4.55 ks. This widely observed variability property is embedded in our artificial light curves in a natural way depicting in a vivid way the fact that our artificial light curves are exact replicas of the observed light curves.

This flare directionality, which is actually mapped on the histogram of the 'bursty' light curves in the form of their positive skewness, is taken automatically into account during this newly proposed light-curve simulation method. This means that for the establishment of confidence intervals, it is of great importance to take correctly into consideration the distribution of the measurements since this can affect significantly the level of chance coincidence occurrences. Note, that for the case of Gaussian light curves (i.e. with minuscule skewness), the method automatically is in accordance with the confidence intervals derived by TK95.

7 INVARIANT QUANTITIES AND STATISTICAL DEPENDENCES

During any simulation process, it is very important to understand which light-curve properties are preserved and which are not. The TK95 procedure preserves only the underlying PSD of the observed data set, assuming a Gaussian distribution of measurements for all the cases. Our method preserves both the underlying PSD and the PDF (observed or parent).

The preservation of the PDF means that all the statistical moments of a given data set, i.e. mean (μ), variance (σ^2), skewness (γ_1), kurtosis (γ_2), and so on, are identical between the observed and the surrogate data sets. Since all these quantities are included

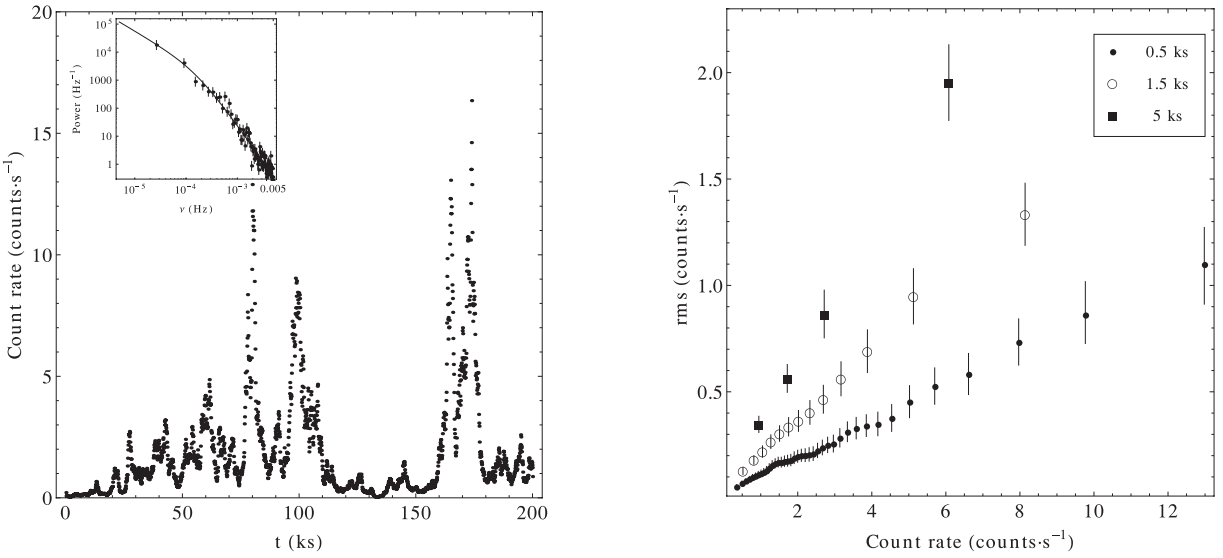

Figure 18. The rms–flux relation. Left-hand panel: the exponentiated light curve together with its binnned logarithmic periodogram and final PSD model (inset). Right-hand panel: the average rms–flux estimates coming from the 1000 simulated light curves using three different binning schemes, i.e. 0.5, 1.5 and 5 ks, respectively.

in the input PDF (they describe its shape e.g. asymmetry, peakedness), they are conserved by construction; during the last iteration step (ranking, step iv) all the measurements are redistributed based on the input PDF (the insets in the right-hand panel of Fig. 4 and the left-hand panel of Fig. 9 are identical). The fact that all the statistical moments are conserved does not mean that all the statistical dependences of the measurements (e.g. non-linear interactions) are preserved. These are two completely different statistical quantities.

The various statistical dependences of the measurements are characterized only by the Fourier transform of the joint cumulant functions, known as polyspectra, e.g. autospectrum, bispectrum, trispectum, and so on (Appendix B). Our method preserves only the second-order joint cumulant (i.e. the covariance) of the input data set, and thus its Fourier transform, autospectrum (and thus its squared amplitude, the PSD), is the only spectral quantity which is preserved in the final converged synthetic light curve. Thus, the only dependences that are preserved are those corresponding to the covariance – all the higher order dependences are ignored.

A very common source of confusion and misunderstandings is that there is a notion that preservation of higher order statistical moments (e.g. γ_1, γ_2, and so on) means preservation of the higher order spectra. The statistical moments characterize only the shape of the PDF, whilst existence of potential dependences between the data points is mapped only on the polyspectra. As we can see in Appendix B, this confusion originates from the fact that σ^2 and γ_1 (depicting the shape of the PDF) are equal to the zero delayed joint cumulants $[C_2(0), C_3(0, 0)]$ (i.e. for $s_1 = s_2 = 0$); however, polyspectra are the Fourier transform of the joint cumulants, i.e. summed over all s_i.

A simple example to demonstrate this is the following. We consider a genuine non-linear process similar to the one given in Provenzale et al. (1992) (also used by Vio et al. 1992):

$$x'(t) = 1 - x(t)^{1/5} + x(t)^3 wn(t) \quad (4)$$

in which $wn(t)$ is a Gaussian white noise process with mean value of 0 and standard deviation of 1. We solve this equation for $x(0.01) = 1$ in $t \in [0.01, 50]$ with a $\Delta t = 0.01$ and then we scale the time axes from 1 to 5000 time units (t.u.) in steps of 1 t.u. One realization of the corresponding process, $x_{\rm nl}(t)$, is shown in the left-hand panel of Fig. 19. Then, we shuffle randomly the data of this process (having an equal probability among all the numbers) yielding a white noise process, $x_{\rm sfl}(t)$, and we plot the data in the right of Fig. 19. In this way, none of the initial dependences between the data points are preserved, but the two data sets still have identical PDFs, since they consist of exactly the same data points (Fig. 20). This PDF has non-zero higher order statistical moments, i.e. skewness and kurtosis of $\gamma_1 = 2.57$ and $\gamma_2 = 12.77$, respectively.

For each data set, we then estimate the normalized squared amplitudes of its bispectrum, known as bicoherence, following Kim & Powers (1979), for the frequencies inside the inner triangle of the principal domain (Hinich & Messer 1995). We have divided each data set into 100 segments, each one consisting of 50 t.u. (i.e. 50 consecutive data points), and for the estimation of the bicoherence, we have averaged the corresponding Fourier transforms and biperiodograms.

As we can see from Fig. 21, the two data sets have genuinely different bicoherences, i.e. genuinely different bispectra. The left-hand panel of Fig. 21 exhibits a great deal of structure for various combinations of (f_1, f_2) depicting the non-linear dependences for the data set $x_{\rm nl}(t)$. In contrast, the right-hand panel of Fig. 21 shows (as expected) a rather quiescent behaviour for the shuffled data set, $x_{\rm sfl}(t)$, despite the fact that it shares exactly the same PDF with the previous data set.

This simple example shows vividly that despite the fact that the two data sets share exactly the same PDF, i.e. have the same high-order statistical moments, they do not share the same bispectra, i.e. the various dependences between the measurements are genuinely different.

8 SUMMARY AND DISCUSSION

We have presented a new algorithm able to produce artificial light curves which are distributed based on a given PDF (parent or observed) and a given underlying PSD. Our publicly available algorithm combines and enhances the methods of TK95 and SS96. The

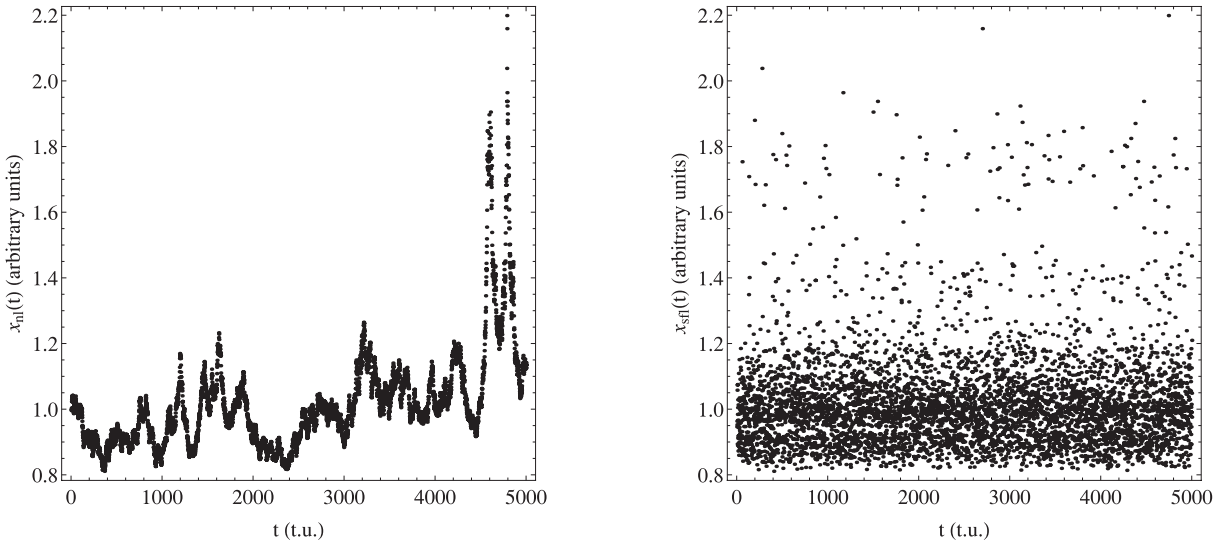

Figure 19. A time series process. Left-hand panel: a realization of the non-linear process (equation 4), $x_{nl}(t)$, with the times rescaled in the range $t \in [1, 5000]$ t.u. and $\Delta t = 1$ t.u. Right-hand panel: a random shuffle of the $x_{nl}(t)$, $x_{sfl}(t)$.

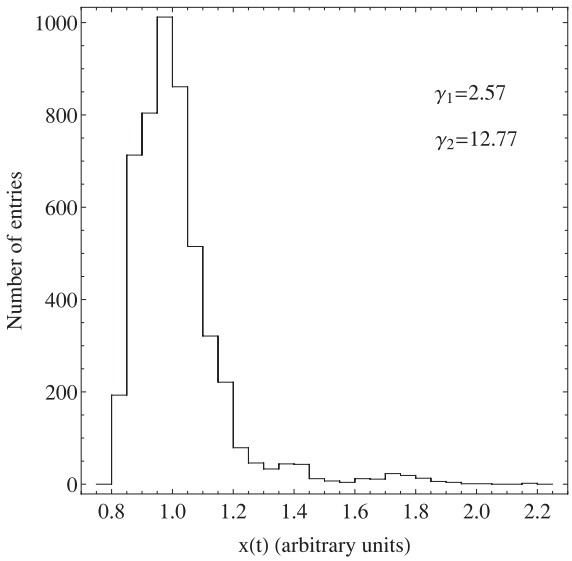

Figure 20. The PDF of the time series process shown in Fig. 19 which has a skewness and a kurtosis of $\gamma_1 = 2.57$ and $\gamma_2 = 12.77$, respectively.

new method improves significantly on the widely used procedure of TK95 which is able to produce artificial light curves which are only normally distributed. Thus, for any sort timing studies, in which simulated data sets are needed, our algorithm preserves all the genuine variability and statistical source properties yielding ensembles of truly random artificial data sets.

The merits of our method can be summarized as follows.

(i) It reproduces the exact variability properties of the observed data, since the synthetic light curves follow the input PSD. The input PSD can originate either from an actual observation or a theoretical model.

(ii) It reproduces the exact statistical properties of the observed data/theoretical model since it uses their/its PDF. Thus, the surrogate light curves carry all the statistical moments and depending on the nature of the statistical study, the PDF corresponds either to the observed or the parent PDF.

(iii) Introduction of higher statistical moments (other than mean value and variance that characterize completely only the normal distribution) and definition of genuinely positively probability distributions allow the construction of realistic 'bursty' light curves which cannot be created by TK95.

(iv) For the special case of Gaussian light curves, the method yields synthetic data sets which are by construction equivalent to those of TK95.

(v) For the special case of a parent log-normal distribution, the simulated light curves exhibit the rms–flux relation.

Particularly for the case of 'bursty' light curves, having by definition non-Gaussian positively defined PDFs which can be even sometimes described by right *heavy-tailed* PDFs, representative for extreme flaring states, this new method is the most appropriate for the correct establishment of confidence intervals of a given method, e.g. CCF analysis.

Due to its generality, the method can be employed to a vast variety of statistical analysis purposes involving light curves obtained across the electromagnetic spectrum for any object. The Monte Carlo simulation studies, which are currently performed using the TK95 products, can now be extended to statistically much more accurate synthetic light curves, thus providing us with robust results with respect to e.g. cross-correlation analysis, establishment of detection significance for future missions (e.g. *LOFT*, *CTA*), detection and characterization of variability, understudying of the effects of irregular sampling.

ACKNOWLEDGEMENTS

DE and IMM acknowledge the Science and Technology Facilities Council (STFC) for support under grant ST/G003084/1. This research has made use of NASA's Astrophysics Data System Bibliographic Services. Finally, we are grateful to the anonymous referee for the very useful comments and suggestions that helped improved the quality of the manuscript.

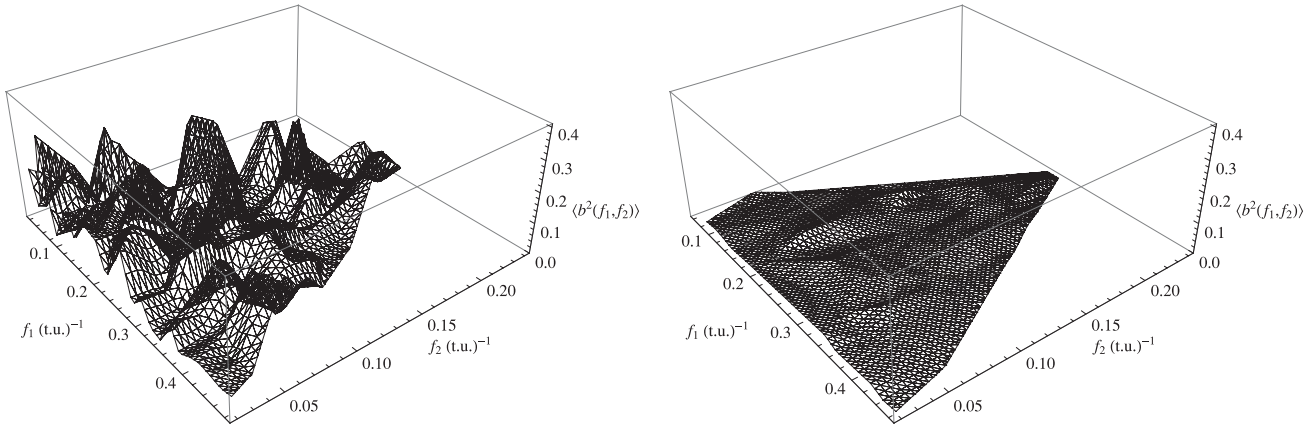

Figure 21. Bispectrum analysis. Left-hand panel: the bicoherence of the non-linear data set, $x_{\rm nl}(t)$ (Fig. 19, left-hand panel). Right-hand panel: the bicoherence of the shuffled data set, $x_{\rm sfl}(t)$ (Fig. 19, right-hand panel).

REFERENCES

Abramowitz M., Stegun I. A., 1972, Handbook of Mathematical Functions. Dover Press, New York
Agudo I. et al., 2011, ApJ, 726, L13
Aharonian F. et al., 2007, ApJ, 664, L71
Aharonian F. et al., 2008, Phys. Rev. Lett., 101, 170402
Alexander T., 1997, in Maoz D., Sternberg A., Leibowitz E. M., eds, Astrophysics and Space Science Library, Vol. 218, Astronomical Time Series. Kluwer, Dordrecht, p. 163
Anderson T. W., Darling D. A., 1952, Ann. Math. Stat., 23, 193
Anderson E. R., Duvall T. L., Jr, Jefferies S. M., 1990, ApJ, 364, 699
Barret D., Vaughan S., 2012, ApJ, 746, 131
Bartlett E. S., Clark J. S., Coe M. J., Garcia M. R., Uttley P., 2013, MNRAS, 429, 1213
Bauer A., Baltay C., Coppi P., Ellman N., Jerke J., Rabinowitz D., Scalzo R., 2009, ApJ, 696, 1241
Benlloch S., Wilms J., Edelson R., Yaqoob T., Staubert R., 2001, ApJ, 562, L121
Bialas A., Peschanski R., 1986, Nucl. Phys. B, 273, 703
Blomme R. et al., 2011, A&A, 533, A4
Cash W., 1979, ApJ, 228, 939
Chatterjee R. et al., 2009, ApJ, 704, 1689
Chaudhry M. A., Zubair S. M., 1994, J. Comput. Appl. Math., 55, 99
Chitnis V. R., Pendharkar J. K., Bose D., Agrawal V. K., Rao A. R., Misra R., 2009, ApJ, 698, 1207
Claerbout J. F., 1990, in SEP-67, Stanford Exploration Project, p. 139
Dahlhaus R., Janas D., 1996, Ann. Stat., 24, 1934
Davies R. B., Harte D. S., 1987, Biometrika, 74, 95
de Wolf A. E., Dremin I. M., Kittel W., 1996, Phys. Rep., 270, 1
Deeter J. E., 1984, ApJ, 281, 482
Deeter J. E., Boynton P. E., 1982, ApJ, 261, 337
Do T., Ghez A. M., Morris M. R., Yelda S., Meyer L., Lu J. R., Hornstein S. D., Matthews K., 2009, ApJ, 691, 1021
Done C., Madejski G. M., Mushotzky R. F., Turner T. J., Koyama K., Kunieda H., 1992, ApJ, 400, 138
Doro M. et al., 2013, Astropart. Phys., 43, 189
Edelson R. A., Krolik J. H., 1988, ApJ, 333, 646
Emmanoulopoulos D., McHardy I. M., Uttley P., 2010, MNRAS, 404, 931
Franke J., Härdle W., 1992, Ann. Stat., 20, 121
Frisken B. J., 2001, Appl. Opt., 40, 4087
Gandhi P., 2009, ApJ, 697, L167
Gaskell C. M., Sparke L. S., 1986, ApJ, 305, 175
Gierliński M., Middleton M., Ward M., Done C., 2008, Nat, 455, 369
Góra D., Bernardini E., Cruz Silva A. H., 2011, Astropart. Phys., 35, 201
Greene J. E. et al., 2010, ApJ, 723, 409
Grosso N., Hamaguchi K., Kastner J. H., Richmond M. W., Weintraub D. A., 2010, A&A, 522, A56
Gurley K. R., Kareem A., Tognarelli M. A., 1996, Int. J. Non-Linear Mech., 31, 601
Hinich M., Messer H., 1995, IEEE Trans. Signal Process., 43, 2130
Hunter I. W., Kearney R. E., 1983, Biol. Cybern., 47, 141
Hwang S., Satchell S. E., 1999, Int. J. Financ. Econ., 4, 271
Johnson G. E., 1994, Proc. IEEE, 82, 270
Kelly B. C., Sobolewska M., Siemiginowska A., 2011, ApJ, 730, 52
Khabibullin I., Sazonov S., Sunyaev R., 2012, MNRAS, 426, 1819
Kim Y. C., Powers E. J., 1979, IEEE Trans. Plasma Sci., 7, 120
Kirchner J. W., 2005, Phys. Rev. E, 71, 066110
Kirkpatrick S., Gelatt C. D., Vecchi M. P., 1983, Sci, 220, 671
Knuth K. H., 2006, Preprint (astro-ph/0605197)
Koptelova E., Oknyanskij V. L., Artamonov B. P., Burkhonov O., 2010, MNRAS, 401, 2805
Kreiss J. P., Paparoditis E., 2003, Ann. Stat., 31, 1923
Li T., Qu J., Feng H., Song L., Ding G., Chen L., 2004, Chin. J. Astron. Astrophys., 4, 583
Liu B., Munson D. C. J., 1982, IEEE Trans. Acoust. Speech Signal Process., 30, 973
MacLeod C. L. et al., 2010, ApJ, 721, 1014
McHardy I., 2010, in Belloni T., ed., Lecture Notes in Physics, Vol. 794, X-Ray Variability of AGN and Relationship to Galactic Black Hole Binary Systems. Springer-Verlag, Berlin, p. 203
McHardy I. M., Papadakis I. E., Uttley P., Page M. J., Mason K. O., 2004, MNRAS, 348, 783
Nelder J. A., Mead R., 1965, Comput. J., 7, 308
Ofek E. O., Maoz D., 2003, ApJ, 594, 101
Owen A. B., 2001, Empirical Likelihood. Chapman and Hall/CRC, Boca Raton
Papadakis I. E., Lawrence A., 1993, MNRAS, 261, 612
Peterson B. M., Wanders I., Horne K., Collier S., Alexander T., Kaspi S., Maoz D., 1998, PASP, 110, 660
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd edn. Cambridge Univ. Press, Cambridge
Priestley M. B., 1981, Spectral Analysis and Time Series: Probability and Mathematical Statistics, Vol. 1 and 2. Academic Press, London
Primini F. A. et al., 2011, ApJS, 194, 37
Provenzale A., Smith L. A., Vio R., Murante G., 1992, Physica D: Nonlinear Phenomena, 58, 31
Rajaguru S. P., Hughes S. J., Thompson M. J., 2004, Sol. Phys., 220, 381
Schreiber T., Schmitz A., 1996, Phys. Rev. Lett., 77, 635 (SS96)
Sowey E. R., 1986, J. R. Stat. Soc. A, 149, 83
Stanishev V., Kraicheva Z., Boffin H. M. J., Genkov V., 2002, A&A, 394, 625
Tewes M., Courbin F., Meylan G., 2012, preprint (astro-ph/1208.5598)
Theiler J., Eubank S., Longtin A., Galdrikian B., Doyne Farmer J., 1992, Physica D: Nonlinear Phenomena, 58, 77

Timmer J., Koenig M., 1995, A&A, 300, 707 (TK95)
Uttley P., McHardy I. M., 2001, MNRAS, 323, L26
Uttley P., McHardy I. M., Papadakis I. E., 2002, MNRAS, 332, 231
Uttley P., McHardy I. M., Vaughan S., 2005, MNRAS, 359, 345
van der Klis M., 1988, in Ogelman H., van den Heuvel E. P. J., eds, NATO ASI Series C, Vol. 262, Timing Neutron Stars. Kluwer, Dordrecht, p. 27
Vaughan S., 2005, A&A, 431, 391
Vaughan S., Edelson R., Warwick R. S., Uttley P., 2003, MNRAS, 345, 1271
Vaughan S., Uttley P., Pounds K. A., Nandra K., Strohmayer T. E., 2011, MNRAS, 413, 2489
Venema V., Bachner S., Rust H. W., Simmer C., 2006, Nonlinear Process. Geophys., 13, 449
Villforth C., Koekemoer A. M., Grogin N. A., 2010, ApJ, 723, 737
Vio R., Cristiani S., Lessi O., Provenzale A., 1992, ApJ, 391, 518
Wilks S. S., 1938, Ann. Math. Stat., 9, 60
Wooldridge J. M., 2001, J. Econ. Perspect., 15, 87
Yamazaki F., Shinozuka M., 1988, J. Eng. Mech., 114, 1183
Zhang Y. H., Cagnoni I., Treves A., Celotti A., Maraschi L., 2004, ApJ, 605, 98

APPENDIX A: DEFINITIONS AND NOMENCLATURE

Below we briefly describe the DFT, the calculation of the periodogram, the PSD estimation and the derivation of the PDF.

A1 The periodogram

Consider a light curve $x(t)$ consisting of N equidistant observations: $\{t_k, x(t_k)\}$ for $k = 1, 2, \ldots, N$ with a sampling period $t_{\rm bin}$, a mean value of μ and a standard deviation of σ. The DFT of the data set is defined following Press et al. (1992):[11]

$$DFT(j) = \sum_{k=1}^{N} x(t_k) \mathrm{e}^{2\pi \mathrm{i}(k-1)j/N} \tag{A1}$$

yielding N estimates for $j = 0, \ldots, N-1$, each one corresponding to a Fourier frequency f_j depending on the parity of N (i.e. even or odd).

At $f_0 = 0$ ($j = 0$) the zero Fourier frequency component, DFT(0), corresponds always to the sum of the light-curve estimates.

For even N:

(i) positive: $f_j^+ = j/(Nt_{\rm bin})$ for $j = 1, \ldots, N/2 - 1$;
(ii) negative: $f_j^- = -(N-j)/(Nt_{\rm bin})$ for $j = N/2 + 1, \ldots, N-1$;
(iii) Nyquist: $f_{N/2} = f_{\rm Nyq} = 1/(2t_{\rm bin})$ for $j = N/2$.

Note that the negative frequencies are mirrored versions of the positive frequencies with opposite signs (around $f_{\rm Nyq}$), e.g. $-f^-_{N/2-1} = f^+_{N/2+1}, \ldots, -f_1^- = f^+_{N-1}$.

For odd N:

(i) positive: $f_j^+ = j/(Nt_{\rm bin})$ for $j = 1, \ldots, (N-1)/2$;
(ii) negative: $f_j^- = -(N-j)/(Nt_{\rm bin})$ for $j = (N+1)/2, \ldots, N-1$;
(iii) Nyquist: there is no Nyquist frequency estimate.

Note again that the negative frequencies are mirrored versions of the positive frequencies with opposite signs, e.g. $-f^-_{(N-1)/2} = f^+_{(N+1)/2}, \ldots, -f_1^- = f^+_{N-1}$.

[11] In this case, the exponential function contains as a running index $k-1$ instead of k since the data start for $k = 1$ and not for $k = 0$.

At a given frequency f_j, $DFT(j)$ is a complex number of the form[12] $q + wi$ and which carries information about the amplitude and the phase of the corresponding sinusoidal component. The amplitude of the sinusoid at a frequency, f_j, is given by

$$\mathscr{A}_j = \frac{1}{N}\sqrt{Re[DFT(j)]^2 + Im[DFT(j)]^2} \tag{A2}$$

and its phase is given by

$$\phi_j = arg[DFT(j)] = \arctan\{Im[DFT(j)], Re[DFT(j)]\} \tag{A3}$$

taking values in the closed-open interval $(-\pi, \pi]$. For the complex number 0, one may use $\phi = 0$ but formally its phase angle is indeterminate.

The periodogram of $x(t)$ at a given Fourier frequency f_j, $P(f_j)$, is defined as the squared amplitude (equation A2) of the corresponding sinusoid component

$$P(f_j) = \mathscr{A}_j^2 = \frac{1}{N^2}\left\{Re[DFT(j)]^2 + Im[DFT(j)]^2\right\} \quad \text{for } j = 0, \ldots, N-1. \tag{A4}$$

Since the light curve consists only of real measurements, $x(t_k) \in \mathbb{R}$, there is a symmetry between the positive and the negative DFT estimates: $DFT(j^-) = [DFT(j^+)]^*$, where j^- and j^+ represent the indices for the negative and positive frequencies, respectively, and the asterisk denotes complex conjugation. Thus, the amplitudes of the corresponding positive and negative components are equal and the periodogram is estimated as

$$P(f_j) = \frac{2}{N^2}\left\{Re[DFT(j)]^2 + Im[DFT(j)]^2\right\} \quad \begin{array}{l}\text{even } N\text{: } j = 0, \ldots, N/2 \\ \text{odd } N\text{: } j = 0, \ldots, (N-1)/2\end{array} \tag{A5}$$

with $f_j = j/(Nt_{\rm bin})$. There is a plethora of normalization factors that can be applied to the periodogram (e.g. Vaughan et al. 2003). In this work, we employ the fractional rms normalization $Nt_{\rm bin}/\mu^2$, and the periodogram (equation A5) becomes

$$P(f_j) = \frac{2t_{\rm bin}}{\mu^2 N}\left\{Re[DFT(j)]^2 + Im[DFT(j)]^2\right\}. \tag{A6}$$

With this normalization, the square root of the integral of the underlying PSD between two frequencies f_1 and f_2 yields the contribution to the fractional rms squared variability (i.e. σ^2/μ^2). Thus, integration between f_1 and $f_{\rm Nyq}$ (even) or $f_{(n-1)/2}$ (odd) yields the total rms squared variability.

A2 PSD estimation

The 'statistical natural' estimator of the underlying PSD, $\mathscr{P}(f)$, is the periodogram, $P(f)$. In the manner of Priestley (1981), assume that the light curve, x_t, originates from a linear process of the form

$$x_t = \sum_0^\infty g_u \epsilon_{t-u}, \tag{A7}$$

where ϵ_t is a purely random Gaussian process and g_u is a given sequence of constants satisfying $\sum_{u=0}^{\infty} g_u^2 < \infty$. At a given frequency,

[12] For the case of even N, the DFT($N/2$) (i.e. at Nyquist frequency) is a real number since, from equation (A1), the exponential function for $j = N/2$ is equal to 1 (for odd k) or -1 (for even k).

f_j, $P(f_j)$ is then asymptotically distributed around the $\mathscr{P}(f_j)$ as

$$P(f_j) = \begin{cases} \frac{1}{2}\chi^2_2 \mathscr{P}(f_j) & j = 1, \ldots, \begin{matrix} N/2-1 \text{ (even } N) \\ (N-1)/2 \text{ (odd } N) \end{matrix} \\ \frac{1}{2}\chi^2_1 \mathscr{P}(f_{\rm Nyq}) & j = N/2 \text{ (even } N), \end{cases}$$

where χ^2_ν represents the χ^2 distribution with ν d.o.f. This means that, for a given frequency, the standard deviation of the periodogram estimates is 100 per cent, automatically making the ensemble of periodogram estimates an inconsistent estimator of the underlying PSD.

In order to retrieve the $\mathscr{P}(f_j)$, one can use either binning or maximum likelihood methodologies. For the former, the binned logarithmic periodogram has been proposed by Papadakis & Lawrence (1993) ensuring that the logarithmic periodogram estimates are normally distributed within each geometric mean frequency bin. Thus, PSD models can be fitted to the logarithmic periodogram estimates using a simple least-squares method, requiring Gaussianity within each bin. The latter should include at least 10 periodogram estimates, a fact which partially limits the usefulness of the method for small data sets. Another approach is to fit a PSD model directly to the ensemble of periodogram estimates by performing maximum likelihood estimation which makes direct use of the underlying distribution at a given Fourier frequency (equation A8). This is the approach used in this work, and detailed references can be found in e.g. Anderson et al. (1990), Vaughan (2005) and Barret & Vaughan (2012).

Consider an underlying PSD model, $\mathscr{P}(f_j;\boldsymbol{\gamma})$, in which $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ is a vector consisting of the unknown model parameters such as normalization, break/bend frequency, low/high-frequency slopes, etc. The probability of obtaining a given single periodogram estimate $P(f_j)$ for the given PSD model $\mathscr{P}(f_j;\boldsymbol{\gamma})$ is

$$\lambda_j\left[P(f_j)|\mathscr{P}(f_j;\boldsymbol{\gamma})\right] = \begin{cases} \frac{\mathrm{e}^{-P(f_j)/\mathscr{P}(f_j;\boldsymbol{\gamma})}}{\mathscr{P}(f_j;\boldsymbol{\gamma})} & j = 1, \ldots, \begin{matrix} N/2-1 \text{ (even } N) \\ (N-1)/2 \text{ (odd } N) \end{matrix} \\ \frac{\mathrm{e}^{-P(f_{\rm Nyq})/\mathscr{P}(f_{\rm Nyq};\boldsymbol{\gamma})}}{\left[\pi P(f_{\rm Nyq})\mathscr{P}(f_{\rm Nyq};\boldsymbol{\gamma})\right]^{1/2}} & j = N/2 \text{ (even } N), \end{cases}$$

The constituent functions of the above piecewise expression are usually referred to as 'scaled χ^2 distributions' with 2 and 1 d.o.f. for the top and lower branch, respectively. More precisely, these functions are special forms of the gamma distribution $\Gamma[\nu/2, \mathscr{P}(f_j;\boldsymbol{\gamma})]$, where ν corresponds to the d.o.f., i.e. $\nu = 1$ corresponds only to the Nyquist frequency, $f_{\rm Nyq}$ ($j = N/2$, even N), and $\nu = 2$ to all other frequencies (for either even or odd N).[13]

The joint probability of obtaining the ensemble of periodogram estimates for the given PSD model is

$$\mathscr{L} = \prod_{i=1}^{\begin{matrix} N/2-1 \text{ (even } N) \\ (N-1)/2 \text{ (odd } N) \end{matrix}} \lambda_j\left[P(f_j)|\mathscr{P}\left(f_j;\boldsymbol{\gamma}\right)\right] \tag{A10}$$

since asymptotically (i.e. $N \to \infty$) the various periodogram estimates are strictly independent at the Fourier frequencies f_j (Priestley 1981) (this is the reason why the periodogram is estimated only for these frequencies and not for intermediate values). The maximum likelihood estimate of the model function parameters, $\boldsymbol{a}$, is obtained by maximizing the above probability, or equivalently by minimizing the log-likelihood function $\mathcal{C} = -2\ln\mathscr{L}$ which is equal to

$$\mathcal{C} = \begin{cases} \left\{2\sum_{j=1}^{j=N/2-1}\left\{\ln\left[\mathscr{P}(f_j;\boldsymbol{\gamma})\right] + \frac{P(f_j)}{\mathscr{P}(f_j;\boldsymbol{\gamma})}\right\}\right\} + \\ \ln[\pi P(f_{\rm Nyq})\mathscr{P}(f_{\rm Nyq};\boldsymbol{\gamma})] + 2\frac{P(f_{\rm Nyq})}{\mathscr{P}(f_{\rm Nyq};\boldsymbol{\gamma})} \quad \text{(even } N) \\ 2\sum_{j=1}^{j=(N-1)/2}\left\{\ln\left[\mathscr{P}(f_j;\boldsymbol{\gamma})\right] + \frac{P(f_j)}{\mathscr{P}(f_j;\boldsymbol{\gamma})}\right\} \text{ (odd } N). \end{cases} \tag{A11}$$

In this work, we have employed two minimization routines (see for details Press et al. 1992): a direct search method, Nelder–Mead (Nelder & Mead 1965) and a stochastic function minimizer, simulated annealing (Kirkpatrick, Gelatt & Vecchi 1983). The PSD models, $\mathscr{P}(f_j;\boldsymbol{\gamma})$, which are usually fitted to the data have a power-law form (e.g. broken power law, continuous bending power law, etc.) and, for these type of minimization problems, both methods have identical results.

The joint confidence intervals for q model parameters from a total of n components of $\boldsymbol{a}$, $\{\alpha_1, \alpha_2, \ldots, \alpha_q, \alpha_{q+1}, \alpha_{q+2}, \ldots, \alpha_n\}$ can be estimated using the method of Cash (1979), based on the theorem of Wilks (1938). Initially, a global minimum is found by varying all the n model parameters yielding $(\mathcal{C}_{\rm min})_n$. The q parameters of interest are fixed to their best-fitting values and the rest, $q+1, q+2, \ldots, n$, then are varied until a global minimum is reached corresponding to $(\mathcal{C}_{\rm min})_{n-q_{\rm bf}}$. The quantity $\Delta\mathcal{C} = (\mathcal{C}_{\rm min})_{n-q_{\rm bf}} - (\mathcal{C}_{\rm min})_n$ is then distributed as a χ^2 distribution with q d.o.f. Thus, the 68.3 and 90 per cent single confidence intervals for one parameter ($q = 1$) correspond to $\Delta\mathcal{C}$ of 1 and 2.71, respectively. Similarly, the 68.3 and 90 per cent joint confidence intervals for one parameter ($q = 2$) correspond to $\Delta\mathcal{C}$ of 2.30 and 4.61, respectively. In general, for a given confidence interval p and a given value of q, the corresponding value of $\Delta\mathcal{C}$ is given by $2Q^{-1}(\nu/2, 0, p)$, where Q^{-1} corresponds to the inverse of the generalized regularized incomplete gamma function (definitions for Q^{-1} can be found in Chaudhry & Zubair 1994).

A3 Probability density function estimation

The probability density function (PDF) of the observations should always be represented by a positive real-valued distribution and, depending on the purpose of the statistical study, should correspond either to the parent or the observed distribution. The PDF is used in the proposed method (Section 2) to produce a sample of independent and identically distributed random variates.

In general, there are three ways to derive the probability density function (parent or observed) from a given data set. Note that for the case of the parent distribution, we need very large data sets to be able to match the overall variability profile of the source under study. One approach is to fit a probability density function model, $\mathfrak{f}(x_i;\boldsymbol{\eta})$, to the histogrammed data (where $\boldsymbol{\eta}$ is a vector consisting of the unknown distribution's model parameters), using the maximum likelihood method in a similar fashion to that described above in Appendix A2, i.e. maximizing the log-likelihood function $\sum_i \ln \mathfrak{f}(x_i;\boldsymbol{\eta})$. For pathological cases of histogrammed observations exhibiting e.g. highly skewed, non-zero kurtosis in conjunction with extreme long-tailed distributions, one can use appropriate methodologies developed for such purposes, such as the method of generalized moments (Wooldridge 2001), the method of cumulants (Frisken 2001) and the method of factorial moments (Bialas & Peschanski 1986), the latter being particularly useful in the presence of low count rates, i.e. high Poisson noise (de Wolf, Dremin & Kittel 1996).

[13] An even more general representation can be obtained through the Pearson's Type III distribution (p. 930 in Abramowitz & Stegun 1972) for $\alpha = 0$, $\beta = \mathscr{P}(f_j;\boldsymbol{\gamma})$ and $p = \nu/2$.

Another approach is to use directly a piecewise constant representation of the unknown PDF (Knuth 2006), using directly the data set consisting of x_i observations ($i = 1, \ldots, N$):

$$\mathfrak{h}(x) = \sum_{k=1}^{M} \frac{N_k}{N \upsilon_k} \Pi(\xi_{k-1}, x, \xi_k), \tag{A12}$$

where N_k is the number of data points in the kth bin, υ_k is the width of the kth bin, ξ_{k-1} and ξ_k are the edges of the kth bin, and $\Pi(\xi_\alpha, x, \xi_\beta)$ is the boxcar function being equal to 1 for $\xi_\alpha \leq x \leq \xi_\beta$ and 0 otherwise.

Finally, instead of using the PDF representation of the data set, one can use a cumulative distribution function by employing the empirical distribution function of the data set consisting of x_i observations ($i = 1, \ldots, N$). This can be done by estimating the following quantity (after Owen 2001):

$$\mathfrak{H}(y) = \frac{1}{N} \sum_{i=1}^{N} \Pi(-\infty, x_i, y). \tag{A13}$$

This can then be used directly in order to produce random numbers, which is the primary reason why we need to estimate the distribution of the data points.

APPENDIX B: STATISTICAL MOMENTS, CUMULANTS AND POLYSPECTRA

In the manner of Priestley (1981), let X be a random variable with *moment generating function* $M(t)$, then the *cumulant generating function*, $K(t)$ is defined as

$$K(t) = \ln [M(t)]. \tag{B1}$$

By expanding the above expression in a power series we get

$$K(t) = k_1 t + k_2 \frac{t^2}{2!} + \cdots + k_r \frac{t^r}{r!}. \tag{B2}$$

The coefficient of $t^r/(r!)$ is called the *rth cumulant*. Only, the first three cumulants coincide with the first three statistical moments (mean, variance, skewness) and all the other are given by more complicated polynomial expressions, i.e $k_1 = \mu$, $k_2 = \sigma^2$, $k_3 = \gamma_1$, $k_4 = \gamma_2 - 3\sigma^4$, etc.

Generalizing the above to more random variables, let X_t be a process stationary up to order k and let $C(s_1, s_2, \ldots, s_{k-1})$ denote the *joint cumulant* of order k of the set of random variables, $\{X_t, X_{t+s_1}, \ldots, X_{t+s_{k-1}}\}$ is the coefficient of $(z_1, z_2, \ldots, z_k)$ in the expansion of the joint cumulant generating function

$$K(z_1, z_2, \ldots, z_n) = \ln [M(z_1, z_2, \ldots, z_n)] \tag{B3}$$

in which $M(z_1, z_2, \ldots, z_n)$ is the *joint moment generating function*.

The second-order joint cumulant, $C_2(s_1)$, is simply the covariance, $cov(X_t, X_{t+s_1})$ and the third-order joint cumulant, $C_3(s_1, s_2)$, is identical to the third-order joint moment, $\gamma_1(s_1, s_2)$ (sometimes in economics this is called co-skewness and the next joint cumulant co-kurtosis; e.g. Hwang & Satchell 1999),

$$C_2(s_1) = \left\langle (X_t - \mu)(X_{t+s_1} - \mu) \right\rangle \tag{B4}$$

$$C_3(s_1, s_2) = \left\langle (X_t - \mu)(X_{t+s_1} - \mu)(X_{t+s_2} - \mu) \right\rangle; \tag{B5}$$

for $s_1 = s_2 = 0$ these two quantities are directly related to the variance and the skewness, $C_2(0) = \sigma^2$ and $C_3(0) = \gamma_1 \sigma^3$ and these are two properties that are mapped on the PDF.

The Fourier transforms of the corresponding higher order cumulants are called polyspectra,

$$h_k(f_1, f_2, \ldots, f_n) = \sum_{s_1=-\infty}^{\infty} \cdots \sum_{s_{k-1}=-\infty}^{\infty} C(s_1, \ldots, s_{k-1}) \mathrm{e}^{-2\pi(f_1 s_1 + \cdots + f_{k-1} s_{k-1})}. \tag{B6}$$

The second-order polyspectrum is the autospectrum and its squared amplitude is the PSD, $|h_2(f)|^2 \equiv \mathscr{P}(f)$. The third- and the fourth-order polyspectra are known as bispectrum and trispectrum, respectively. These are the quantities that characterize the various dependences between the various measurements. The fact that two data sets have e.g. the same variance and skewness [i.e. $C_2(0)$, $C_3(0, 0)$] does not mean that they have the same covariance and third-order joint cumulant, $C_2(s_1)$ and $C_3(s_1, s_2)$, respectively. Thus, data sets which have the same statistical moments (i.e. same PDFs) does not mean that they have the same polyspectra.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Wolfram Notebook file (http://mnras.oxfordjournals.org/lookup/suppl/doi:10.1093/mnras/stt764/-/DC1).

Please note: Oxford University Press are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

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