Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. **402,** 307–320 (2010) doi:10.1111/j.1365-2966.2009.15868.x

A Bayesian test for periodic signals in red noise

S. Vaughan★

X-Ray and Observational Astronomy Group, University of Leicester, Leicester LE1 7RH

Accepted 2009 October 14. Received 2009 October 14; in original form 2009 September 11

ABSTRACT

Many astrophysical sources, especially compact accreting sources, show strong, random brightness fluctuations with broad power spectra in addition to periodic or quasi-periodic oscillations (QPOs) that have narrower spectra. The random nature of the dominant source of variance greatly complicates the process of searching for possible weak periodic signals. We have addressed this problem using the tools of Bayesian statistics; in particular, using Markov Chain Monte Carlo techniques to approximate the posterior distribution of model parameters, and posterior predictive model checking to assess model fits and search for periodogram outliers that may represent periodic signals. The methods developed are applied to two example data sets, both long *XMM–Newton* observations of highly variable Seyfert 1 galaxies: RE J1034 + 396 and Mrk 766. In both cases, a bend (or break) in the power spectrum is evident. In the case of RE J1034 + 396, the previously reported QPO is found but with somewhat weaker statistical significance than reported in previous analyses. The difference is due partly to the improved continuum modelling, better treatment of nuisance parameters and partly to different data selection methods.

Key words: methods: data analysis – methods: statistical – galaxies: Seyfert – X-rays: general.

1 INTRODUCTION

A perennial problem in observational astrophysics is detecting periodic or almost-periodic signals in noisy time series. The standard analysis tool is the periodogram (see e.g. Jenkins & Watts 1969; Priestley 1981; Press et al. 1992; Bloomfield 2000; Chatfield 2003), and the problem of period detection amounts to assessing whether or not some particular peak in the periodogram is due to a periodic component or a random fluctuation in the noise spectrum (see Fisher 1929; Priestley 1981; Leahy et al. 1983; van der Klis 1989; Percival & Walden 1993; Bloomfield 2000).

If the time series is the sum of a random (stochastic) component and a periodic one, we may write $y(t) = y_R(t) + y_P(t)$ and, due to the independence of $y_R(t)$ and $y_P(t)$, the power spectrum of $y(t)$ is the sum of the two power spectra of the random and stochastic processes: $S_Y(f) = S_R(f) + S_P(f)$. This is a *mixed* spectrum (Percival & Walden 1993, section 4.4) formed from the sum of $S_P(f)$, which comprises only narrow features, and $S_R(f)$, which is a continuous, broad spectral function. Likewise, we may consider an evenly sampled, finite time series $y(t_i)$ ($i = 1, 2, \ldots, N$) as the sum of two finite time series: one is a realization of the periodic process, the other a random realization of the stochastic process. We may compute the periodogram (which is an estimator of the true power spectrum) from the squared modulus of the discrete Fourier transform (DFT) of the time series, and, as with the power spectra, the periodograms of the two processes add linearly: $I(f_j) = I_R(f_j) + I_P(f_j)$. The periodogram of the periodic time series will contain only narrow 'lines' with all the power concentrated in only a few frequencies, whereas the periodogram of the stochastic time series will show power spread over many frequencies. Unfortunately, the periodogram of stochastic processes fluctuates wildly around the true power spectrum, making it difficult to distinguish random fluctuations in the noise spectrum from truly spectral periodic components. See van der Klis (1989) for a thorough review of these issues in the context of X-ray astronomy.

Particular attention has been given to the special case that the spectrum of the stochastic process is flat (a *white noise* spectrum $S(f) = \text{const}$), which is the case when the time series data $y_R(t_i)$ are independently and identically distributed (IID) random variables. Reasonably well-established statistical procedures have been developed to help identify spurious spectral peaks and reduce the chance of false detections (e.g. Fisher 1929; Priestley 1981; Leahy et al. 1983; van der Klis 1989; Percival & Walden 1993). In contrast, there is no comparably well-established procedure in the general case that the spectrum of the stochastic process is not flat.

In a previous paper (Vaughan 2005, henceforth V05), we proposed what is essentially a generalization of Fisher's method to the case where the noise spectrum is a power law: $S_R(f) = \beta f^{-\alpha}$ (where α and β are the power-law index and normalization parameters). Processes with power spectra that show a power-law dependence on frequency with $\alpha > 0$ (i.e. increasing power to lower frequencies) are called *red noise* and are extremely common in astronomy and elsewhere (see Press 1978). In this paper, we expand upon the ideas

★E-mail: sav2@star.le.ac.uk

in V05 and, in particular, address the problem from a Bayesian perspective that allows further generalization of the spectral model of the noise.

The rest of this paper is organized as follows. In Section 2, we introduce some of the basic concepts of the Bayesian approach to statistical inference; readers familiar with this topic may prefer to skip this section. Section 3 gives a brief overview of classical significance testing using p-values (tail area probabilities) and test statistics, and Section 4 discusses the posterior predictive p-value, a Bayesian counterpart to the classical p-value. Section 5 reviews the conventional (classical) approaches to testing for periodogram peaks. Section 6 outlines the theory of maximum likelihood estimation from periodogram data, which is developed into the basis of a fully Bayesian analysis in Sections 7 and 8. The Bayesian method is then applied to two real observations if AGN in Section 9. Section 10 discusses the limitations of the method, and alternative approaches to practical data analysis. A few conclusions are given in Section 11, and two appendices describe details of the simulation algorithms used in the analysis.

2 BAYESIAN BASICS, BRIEFLY

There are two main tasks in statistical inference: parameter estimation and model checking (or comparison). Bayesian parameter estimation is concerned with finding the probability of the parameters given the model $p(\boldsymbol{\theta}|\boldsymbol{x}, H)$, where $\boldsymbol{x}$ $(= \{x_1, \ldots, x_N\})$ are data values , $\boldsymbol{\theta}$ $(= \{\theta_1, \ldots, \theta_M\})$ are parameter values and H represents the model. In contrast, *frequentist* (or *classical*) statistics restricts attention to the sampling distribution of the data given the model and parameters $p(\boldsymbol{x}|\boldsymbol{\theta}, H)$. These two probability functions are related by Bayes' Theorem

$$p(\boldsymbol{\theta}|\boldsymbol{x}, H) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H)}{p(\boldsymbol{x}|H)}. \quad (1)$$

Each of the terms in Bayes' theorem has a name when used in Bayesian data analysis: $p(\boldsymbol{\theta}|\boldsymbol{x}, H)$ is the *posterior* distribution of the parameters; $p(\boldsymbol{x}|\boldsymbol{\theta}, H)$ is the *likelihood* function of the parameters;[1] $p(\boldsymbol{\theta}|H)$ is the *prior* distribution of the parameters, and $p(\boldsymbol{x}|H)$ is a normalizing constant sometimes referred to as the *marginal likelihood* (of the data) or the *prior predictive* distribution.[2] General introductions to Bayesian analysis for the non-specialist include Jeffreys & Berger (1992), Berger & Berry (1988) and Howson & Urbach (1991); more thorough treatments include Berry (1996), Carlin & Louis (2000), Gelman et al. (2004) and Lee (2004); and discussions more focused on physics and astrophysics problems include Sivia (1996), Gregory (2005) and Loredo (1990, 1992).

In Bayesian analysis, the posterior distribution is a complete summary of our inference about the parameters given the data $\boldsymbol{x}$, model H and any prior information. But this can be further summarized using a point estimate for the parameters such as the mean, median or mode of the posterior distribution. For one parameter the posterior mean is

$$\mathrm{E}[\theta|\boldsymbol{x}, H] = \int \theta\, p(\theta|\boldsymbol{x}, H)\mathrm{d}\theta. \quad (2)$$

[1] Note that when considered as a function of the data for known parameters, $p(\boldsymbol{x}|\boldsymbol{\theta}, H)$ is the sampling distribution of the data, but when considered as a function of the parameters for fixed data, $p(\boldsymbol{x}|\boldsymbol{\theta}, H)$ is known as the likelihood, sometimes denoted $L(\boldsymbol{\theta})$.

[2] Some physicists use the *evidence* for this term, e.g. Sivia (1996), Trotta (2008).

A slightly more informative summary is a *credible interval* (or credible region for multiple parameters). This is an interval in parameter space that contain a specified probability mass (e.g. 90 per cent) of the posterior distribution. These intervals give an indication of the uncertainty on the point inferences:

$$\int_R p(\theta|\boldsymbol{x}, H)\mathrm{d}\theta = C, \quad (3)$$

where C is the probability content (e.g. $C = 0.9$) and R is the interval in parameter space. One common approach is to select the interval satisfying equation (3) that contains the highest (posterior) density (i.e. the posterior density at any point inside is higher than at any point outside). This will give the smallest interval that contains a probability C, usually called the highest posterior density region (abbreviated to HDR or HPD interval by different authors). An alternative is the equal tail posterior interval, which is defined by the two values above and below which is $(1 - C)/2$ of the posterior probability. These two types of interval are illustrated in Park, van Dyk & Siemiginowska (2008, see their fig. 1).

If we have multiple parameters but are interested in only one parameter, we may *marginalize* over the other parameters. For example, if $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$, then the posterior distribution for θ_1 is

$$\begin{aligned} p(\theta_1|\boldsymbol{x}, H) &= \int p(\theta_1, \theta_2|\boldsymbol{x}, H)\mathrm{d}\theta_2 \\ &= \int p(\theta_1|\theta_2, \boldsymbol{x}, H)p(\theta_2|\boldsymbol{x}, H)\mathrm{d}\theta_2. \end{aligned} \quad (4)$$

This is the average of the *joint* posterior $p(\theta_1, \theta_2|\boldsymbol{x}, H)$ over θ_2. In the second formulation, the joint posterior has been factored into two distributions, the first is the conditional posterior of θ_1 given θ_2 and the second is the posterior density for θ_2.

Most present day Bayesian analysis is carried out with the aid of Monte Carlo methods for evaluating the necessary integrals. In particular, if we have a method for simulating a random sample of size N from the posterior distribution $p(\theta|\boldsymbol{x}, H)$, then the posterior density may be approximated by a histogram of the random draws. This gives essentially complete information about the posterior (for a sufficiently large N). The posterior mean may be approximated by the sample mean

$$\mathrm{E}[\theta|\boldsymbol{x}, H] \approx \frac{1}{N}\sum_{i=1}^{N}\theta^i, \quad (5)$$

where θ^i are the individual simulations from the posterior. If the parameter is a vector $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$, the mth component of each vector is a sample from the marginal distribution of the mth parameter. This means the posterior mean of the each parameter is approximated by the sample mean of each component of the vector. Intervals may be calculated from the sample quantiles, e.g. the 90 per cent equal tail area interval on a parameter may be approximated by the interval between the 0.05 and 0.95 quantiles of the sample. In this manner the difficult (sometimes insoluble) integrals of equations (2)–(4) may be replaced by trivial operations on the random sample. The accuracy of these approximations is governed by the accuracy with which the distribution of the simulations matches the posterior density, and the size of the random sample N. Much of the work on practical Bayesian data analysis methods has been devoted to the generation and assessment of accurate Monte Carlo methods, particularly the use of Markov Chain Monte Carlo (MCMC) methods, which will be discussed and used later in this paper.

For model comparison, we may again use Bayes theorem to give the posterior probability for model H_i

$$p(H_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|H_i)p(H_i)}{p(\boldsymbol{x})}, \tag{6}$$

and then compare the posterior probabilities for two (or more) competing models, say H_0 and H_1 (with parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, respectively). (In effect, we are treating the choice of model, H_i, as a discrete parameter.) The ratio of these two eliminates the term in the denominator (which has no dependence on model selection):

$$O = \frac{p(H_1|\boldsymbol{x})}{p(H_0|\boldsymbol{x})} = \frac{p(\boldsymbol{x}|H_1)}{p(\boldsymbol{x}|H_0)}\frac{p(H_1)}{p(H_0)}. \tag{7}$$

The first term on the right-hand side of equation (7) is the ratio of likelihoods and is often called the *Bayes factor* (see Kass & Raftery 1995) and the second term is the ratio of the priors. However, in order to obtain $p(H_i|\boldsymbol{x})$, we must first remove the dependence of the posterior distributions on their parameters, often called *nuisance* parameters in this context (we are not interested in making inferences about $\boldsymbol{\theta}_i$, but they are necessary in order to compute the model). In order to do this, the full likelihood function must be integrated or *marginalized* over the joint prior probability density function (PDF) of the parameters:

$$p(\boldsymbol{x}|H_i) = \int p(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i)p(\boldsymbol{\theta}_i|H_i)\mathrm{d}\boldsymbol{\theta}_i. \tag{8}$$

Here, $p(\boldsymbol{x}|\boldsymbol{\theta}_i, H_i)$ is the likelihood and $p(\boldsymbol{\theta}_i|H_i)$ the prior for the parameters of model H_i. Table 1 summarises the notation used throughout the rest of the paper.

3 TEST STATISTICS AND SIGNIFICANCE TESTING

We return briefly to the realm of frequentist statistics and consider the idea of significance testing using a test statistic. A test statistic $T(\boldsymbol{x})$ is a real-valued function of the data chosen such that extreme values are unlikely when the *null hypothesis* H_0 is true. If the sampling distribution of T is $p(T|H_0)$, under the null hypothesis, and the observed value is $T^{\rm obs} = T(\boldsymbol{x}^{\rm obs})$, then the classical p-value is

$$p_{\rm C}(\boldsymbol{x}^{\rm obs}) = \int_{T^{\rm obs}}^{+\infty} p(T|H_0)\mathrm{d}T = \Pr\{T(\boldsymbol{x}^{\rm rep}) \geq T(\boldsymbol{x}^{\rm obs})|H_0\}, \tag{9}$$

where $\Pr x|y$ is the probability of event x given that event y occurred. The second formulation is in terms of *replicated* data that could have been observed, or could be observed in repeat experiments (Meng 1994; Gelman, Meng & Stern 1996; Gelman et al. 2004). The p-value gives the fraction of $p(T|H_0)$ lying above the observed value $T^{\rm obs}$. As such, p-values are *tail area* probabilities, and one usually uses small $p_{\rm C}$ as evidence against the null hypothesis. If the null hypothesis is *simple*, i.e. has no free parameters, or the sampling distribution of T is independent of any free parameters, then the test statistic is said to be *pivotal*. If the distribution of the test statistic does depend on the parameters of the model, i.e. $p(T|\boldsymbol{\theta}, H_0)$, as is often the case, then we have a *conditional* p-value

$$p_{\rm C}(\boldsymbol{x}^{\rm obs}, \boldsymbol{\theta}) = \int_{T^{\rm obs}}^{+\infty} p(T|\boldsymbol{\theta})dT = \Pr\{T(\boldsymbol{x}^{\rm rep}) \geq T(\boldsymbol{x}^{\rm obs})|\boldsymbol{\theta}\}. \tag{10}$$

(For clarity, we have omitted the explicit conditioning on H_0.) In order to compute this, we must have an estimate for the nuisance parameters $\boldsymbol{\theta}$.

4 POSTERIOR PREDICTIVE *P*-VALUES

In Bayesian analysis, the *posterior predictive distribution* is the distribution of $\boldsymbol{x}^{\rm rep}$ given the available information which includes $\boldsymbol{x}^{\rm obs}$ and any prior information (e.g. section 6.3 of Gelman et al. 2004):

$$p(\boldsymbol{x}^{\rm rep}|\boldsymbol{x}^{\rm obs}) = \int p(\boldsymbol{x}^{\rm rep}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x}^{\rm obs})\mathrm{d}\boldsymbol{\theta}. \tag{11}$$

Here, $p(\boldsymbol{\theta}|\boldsymbol{x}^{\rm obs})$ is the posterior distribution of the parameters (see equation 1) and $p(\boldsymbol{x}^{\rm rep}|\boldsymbol{\theta})$ is the sampling distribution of the data given the parameters. The Bayesian p-value is the (tail area) probability that the replicated data could give a test statistic at least as extreme as that observed:

$$\begin{aligned} p_{\rm B}(\boldsymbol{x}) &= \int p_{\rm C}(\boldsymbol{x}^{\rm obs}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x}^{\rm obs})\mathrm{d}\boldsymbol{\theta} \\ &= \Pr\{T(\boldsymbol{x}^{\rm rep}) \geq T(\boldsymbol{x}^{\rm obs})|\boldsymbol{x}^{\rm obs}, H_0\}. \end{aligned} \tag{12}$$

This is just the classical p-value (equation 10) averaged over the posterior distribution of $\boldsymbol{\theta}$ (equation 1), i.e. the posterior mean $\mathrm{E}[p_{\rm C}|\boldsymbol{x}^{\rm obs}, H_0]$ which may be calculated using simulations. In other words, it gives the average of the conditional p-values evaluated at

Table 1. Definitions used throughout the paper.

Term	Definition
f_j	The jth Fourier frequency $f_j = j/N\ \Delta T$ $(j = 1, \ldots, N/2)$
I_j	Periodogram at frequency f_j
$\boldsymbol{I}$	vector of periodogram values $\boldsymbol{I} = \{I_1, \ldots, I_{N/2}\}$
$\boldsymbol{\theta}$	Model parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$
$\hat{\boldsymbol{\theta}}_{\rm MLE}$	Maximum likelihood estimates of parameters (equation 18)
$\boldsymbol{x}$	Data (e.g. time series) $\boldsymbol{x} = \{x_1, \ldots, x_N\}$
$p_{\rm C}$	Frequentist/classical (conditional) p-value (equation 10)
$p_{\rm B}$	Bayesian (posterior predictive) p-value (equation 12)
S_j	Model spectral density at frequency f_j, i.e. $S(f_j; \boldsymbol{\theta})$
$\hat{S}_j$	The model computed at the estimate $\hat{\boldsymbol{\theta}}$ (equation 15)
$D(\boldsymbol{x}, \boldsymbol{\theta})$	Deviance $(-2 \log p(\boldsymbol{x}|\boldsymbol{\theta}))$ given model H (equation 17)
$p(\boldsymbol{x}|\boldsymbol{\theta}, H)$	Likelihood for parameters $\boldsymbol{\theta}$ of model H given data $\boldsymbol{x}$ (equation 1)
$p(\boldsymbol{\theta}|H)$	Prior probability density for parameters $\boldsymbol{\theta}$ (equation 1)
$p(\boldsymbol{\theta}|\boldsymbol{x}, H)$	Posterior probability density for parameters $\boldsymbol{\theta}$ given data $\boldsymbol{x}$ (equation 1)
$p(\boldsymbol{x}|H)$	Prior predictive density (aka marginal likelihood) of the data $\boldsymbol{x}$ (equation 1)
$\boldsymbol{x}^{\rm rep}$	Replicated data (from repeat observations or simulations) (equation 11)
$p(\boldsymbol{x}^{\rm rep}|\boldsymbol{x}^{\rm obs})$	Posterior predictive distribution given data $\boldsymbol{x}^{\rm obs}$ (equation 11)
$T(\boldsymbol{x})$	A test statistic

over the range of parameter values, weighted by the (posterior) probability of the parameter values. The aim of the posterior predictive p-value (or, more generally, comparing the observed value of a test statistic to its posterior predictive distribution) is to provide a simple assessment of whether the data are similar (in important ways) to the data expected under a particular model.

This tail area probability does not depend on the unknown value of parameters $\boldsymbol{\theta}$, and is often called the *posterior predictive* p-value (see Rubin 1984; Meng 1994; Gelman et al. 1996, 2004; Protassov et al. 2002). The (classical) conditional p-value and the (Bayesian) posterior predictive p-value are in general different but are equivalent in two special cases. If the null hypothesis is simple or the test statistic T is pivotal, then the sampling and posterior predictive distributions of T are the same, $p_{\rm C} = p_{\rm B}$.

Like the classical (conditional) p-value, $p_{\rm B}$ is used for model checking but has the advantage of having no dependence on unknown parameters. The posterior predictive distribution of T includes the uncertainty in the classical p-value $p_{\rm C}$ due to the unknown nuisance parameters (Meng 1994).

The posterior predictive p-value is a single summary of the agreement between data and model, and may be used to assess whether the data are consistent with being drawn from the model: a p-value that is not extreme (i.e. not very close to 0 or 1) shows the observed value $T^{\rm obs}$ is not an outlier in the population $T^{\rm rep}$. Gelman et al. (2004) and Protassov et al. (2002) argue that model checking based on the posterior predictive distribution is less sensitive to the choice of priors (on the parameters), and more useful in identifying deficiencies in the model, compared to Bayes factors or posterior odds (equation 7).

5 CONDITIONAL SIGNIFICANCE OF PERIODOGRAM PEAKS

We now return to the problem of assessing the significance of peaks in periodograms of noisy time series. The null hypothesis, H_0, in this case is that the time series was the product of a stochastic process. It is well known that the periodogram of any stochastic time series of length N, denoted $I_j = I(f_j)$ at Fourier frequency $f_j = j/N\Delta T$ (with $j = 1, \ldots, N/2$), is exponentially distributed[3] about the true spectral density $S_j = S(f_j)$

$$p(I_j|S_j) = \frac{1}{S_j}\exp(-I_j/S_j) \tag{13}$$

(see Jenkins & Watts 1969; Groth 1975; Priestley 1981; Leahy et al. 1983; van der Klis 1989; Press et al. 1992; Percival & Walden 1993; Timmer & König 1995; Bloomfield 2000; Chatfield 2003). Strictly speaking, this is valid for the Fourier frequencies other than the zero and Nyqist frequency ($j = 1$ and $N/2$), which follow a different distribution, although in the limit of large N this difference is almost always inconsequential. This distribution means that the ratio of the periodogram ordinates I_j to the true spectrum S_j will be identically distributed. If we have a parametric spectral model with known parameters, $S_j(\boldsymbol{\theta})$, the ratio

$$R_j^{\rm obs} = 2I_j^{\rm obs}/S_j(\boldsymbol{\theta}) \tag{14}$$

will be distributed as χ^2_ν with $\nu = 2$ degrees of freedom (see V05) and it is trivial to integrate this density function to find the classical tail area p-value corresponding to a given observed datum $I_j^{\rm obs}$. This simple fact is the basis of many 'textbook' frequentist tests for periodicity. However, $p_{\rm C}$ depends the parameters $\boldsymbol{\theta}$ (and, more generally, the model H), which in general we do not know.

The standard solution is to estimate the parameters, e.g. by fitting the periodogram data, and thereby estimate the spectral density S_j under the null hypothesis, call this $\hat{S}_j$, and use this estimate in the test statistic:

$$\hat{R}_j^{\rm obs} = 2I_j^{\rm obs}/\hat{S}_j. \tag{15}$$

The problem is that the distribution of $\hat{R}_j$ will not be simply χ^2_2 since that does not account for the uncertainty in the spectral estimate $\hat{S}_j$. V05 presented a partial solution to this, by treating the statistic $\hat{R}_j$ as the ratio of two random variables under certain simplifying assumptions. In what follows, we use Bayesian methods to develop a much more general method for estimating the parameters of a power spectral model, and posterior predictive model checking to check the quality of a model fit and to map out the distribution of $\hat{R}_j$ conditional on the observed data.

6 PERIODOGRAM ANALYSIS VIA THE LIKELIHOOD FUNCTION

As discussed in V05, and based on the results of Geweke & Porter-Hudak (1983), a very simple way to obtain a reasonable estimate of the index and normalization of a power-law power spectrum, $S(f) = \beta f^{-\alpha}$, is by linear regression of $\log I_j^{\rm obs}$ on $\log f_j$ (see also Pilgram & Kaplan 1998). This provides approximately unbiased and normally distributed estimates of the power-law index (α) and normalization (actually $\log\beta$) even for relatively few periodogram points (i.e. short time series).

The log periodogram regression method has the advantage of being extremely simple computationally, so that estimates of the power-law parameters (and their uncertainties) can be found with minimal effort. However, the method does not easily generalize to other model forms and does not give the same results as direct maximum likelihood analysis[4] even in the special case of a power-law model.

As discussed in Anderson, Duvall & Jefferies (1990), and also appendix A of V05, maximum likelihood estimates (MLEs) of the parameters of a model $S(\boldsymbol{\theta})$ may be found by maximizing the joint likelihood function

$$p(\boldsymbol{I}|\boldsymbol{\theta}, H) = \prod_{j=1}^{N/2} p(I_j|S_j) \tag{16}$$

(cf. equation 13), or equivalently minimizing the following function

$$D(\boldsymbol{I}, \boldsymbol{\theta}, H) = -2\log p(\boldsymbol{I}|\boldsymbol{\theta}, H) = 2\sum_{j=1}^{N/2}\left\{\frac{I_j}{S_j} + \log S_j\right\}, \tag{17}$$

[3] The exponential distribution $p(x|\lambda) = \lambda e^{-\lambda x}$ is a special case of the chi square distribution χ^2_ν with $\nu = 2$ degrees of freedom, and a special case of the gamma distribution, $\Gamma(1, 1/\lambda)$. See e.g. Eadie et al. (1971), Carlin & Louis (2000), Gelman et al. (2004) or Lee (2004) for more on specific distribution functions.

[4] Andersson (2002) provided a modification of the Geweke & Porter-Hudak (1983) fitting method based on the fact that the logarithm of the periodogram ordinates follow a Gumbel distribution. He gives the log likelihood function for the logarithm of the periodogram fitted with a linear function. Maximizing this function should give the maximum likelihood estimates of the power-law parameters.

which is twice the minus log likelihood.[5] This is sometimes known as the Whittle likelihood method, after Whittle (1953) and Whittle (1957), and has been discussed in detail elsewhere (e.g. Hannan 1973; Pawitan & O'Sullivan 1994; Fan & Zhang 2004; Contreras-Cristán, Gutiérrez-Peña & Walker 2006). Here, we use the notation $D(\boldsymbol{I}, \boldsymbol{\theta})$ for consistency with Gelman et al. (2004, section 6.7) where it is used as the *deviance*, a generalization of the common weighted square deviation (or chi square) statistic. Finding the MLEs of the parameters is the same as finding[6]

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\rm MLE} &= \arg\min_{\boldsymbol{\theta}} D(\boldsymbol{I}^{\rm obs}, \boldsymbol{\theta}, H) \\ &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{I} = \boldsymbol{I}^{\rm obs}|\boldsymbol{\theta}, H).\end{aligned} \tag{18}$$

7 BAYESIAN PERIODOGRAM ANALYSIS THROUGH MCMC

We have now laid the groundwork for a fully Bayesian periodogram analysis. Equation (16) gives the likelihood function for the data given the model $S(\boldsymbol{\theta})$, or equivalently, equation (17) gives the minus log likelihood function, which is often easier to work with. Once we assign a prior distribution on the model parameters, we can obtain their joint posterior distribution using Bayes theorem (equation 1):

$$p(\boldsymbol{\theta}|\boldsymbol{I}, H) \propto p(\boldsymbol{I}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H) = q(\boldsymbol{\theta}|\boldsymbol{I}, H), \tag{19}$$

where $q(\boldsymbol{\theta}|\boldsymbol{I}, H)$ is the unnormalized (joint posterior) density function (the normalization does not depend on $\boldsymbol{\theta}$). This can be summarized by the posterior mean (or mode) and credible intervals (equations 2 and 3). We may also assess the overall fit using a posterior predictive p-value for some useful test quantity.

We may now write an expression for the joint posterior density (up to a normalization term), or its negative logarithm (up to an additive constant)

$$-\log q(\boldsymbol{\theta}|\boldsymbol{I}, H) = D(\boldsymbol{I}, \boldsymbol{\theta}, H)/2 - \log p(\boldsymbol{\theta}|H). \tag{20}$$

The posterior mode may then be found by minimizing this function (e.g. using a good numerical non-linear minimization algorithm, or Monte Carlo methods in the case of complex, multiparameter models).

In the limit of large sample size ($N \to \infty$), the posterior density will tend to a multivariate normal under quite general conditions (see chapter 4 of Gelman et al. 2004). For finite N, we may make a first approximation to the posterior using a multivariate normal distribution centred on the mode and with a covariance matrix Σ equal to the curvature of the log posterior at the mode (see section 12.2 of Gelman et al. 2004 and section 5.5 of Albert 2007). (Approximating the posterior as a normal in this way is often called the *Laplace approximation*.) This can be used as the basis of a *proposal distribution* in a MCMC algorithm that can efficiently generate draws from the posterior distribution $q(\boldsymbol{\theta}|\boldsymbol{I}, H)$, given some data $\boldsymbol{I} = \boldsymbol{I}^{\rm obs}$.

[5] The periodogram is χ^2_2 distributed (equation 13) for Fourier frequencies $j = 1, 2, \ldots, N/2 - 1$. At the Nyquist frequency ($j = N/2$), it has a χ^2_1 distribution. One could choose to ignore the Nyquist frequency (sum over $j = 1, \ldots, N/2 - 1$ only), or modify the likelihood function to account for this. But in the limit of large N, the effect on the overall likelihood should be negligible, and so we ignore it here and sum over all non-zero Fourier frequencies.

[6] For a function $f(x)$, $\arg\min f(x)$ gives the set of points of x for which $f(x)$ attains its minimum value.

The MCMC was generated by a random-walk Metropolis–Hastings algorithm using a multivariate normal (with the covariance matrix as above, but centred on the most recent iteration) as the proposal distribution. More details on posterior simulation using MCMC is given in Appendix A. For each set of simulated parameters, we may generate the corresponding spectral model $S(\boldsymbol{\theta})$ and use this to generate a periodogram $\boldsymbol{I}^{\rm rep}$ from the posterior predictive distribution (which in turn may be used to generate a time series if needed, see Appendix B).

8 POSTERIOR PREDICTIVE PERIODOGRAM CHECKS

With the data simulated from the posterior predictive distribution, $\boldsymbol{I}^{\rm rep}$, we may calculate the distribution of any test statistic. Of course, we wish to use statistics that are sensitive to the kinds of model deficiency we are interested in detecting, such as breaks/bends in the smooth continuum and narrow peaks due to quasi-periodic oscillation (QPOs). Given the arguments of Sections 5, a sensible choice of statistic for investigating QPOs is $T_{\rm R} = \max_j \hat{R}_j$ (see equation 15). Note that there is no need to perform a multiple trial (Bonferroni) correction to account for the fact that many frequencies are tested before the strongest candidate is selected, as long as this exact procedure is also applied to the simulated data as the real data.

Another useful statistic is based on the traditional χ^2 statistic, i.e. the sum of the squared standard errors

$$\chi^2(\boldsymbol{I}, \boldsymbol{\theta}) = \sum_{j=1}^{N/2} \frac{(I_j - {\rm E}[I_j|\boldsymbol{\theta}])^2}{{\rm V}[I_j|\boldsymbol{\theta}]} = \sum_{j=1}^{N/2} \left(\frac{I_j - S_j(\boldsymbol{\theta})}{S_j(\boldsymbol{\theta})}\right)^2, \tag{21}$$

where ${\rm E}[\cdot]$ and ${\rm V}[\cdot]$ indicate expectation and variance, respectively. We use $T_{\rm SSE} = \chi^2(\boldsymbol{I}, \hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is the mode of the posterior distribution. This is an 'omnibus' test of the overall data model match ('goodness-of-fit') and will be more sensitive to inadequacies in the continuum modelling since all data points are included (not just the largest outlier as in $T_{\rm R}$). This is the same as the merit function used by Anderson et al. (1990, equation 16), which we call $T_{\rm SSE}$ (for summed square error).

The above two statistics are useful for assessing different aspects of model fitness. By contrast, the likelihood ratio test (LRT) statistic (Eadie et al. 1971; Cowan 1998; Protassov et al. 2002) is a standard tool for comparing nested models. As such it may be used to select a continuum model prior to investigating the residuals for possible QPOs. The LRT statistic is equal to twice the logarithm of the ratio of the likelihood maxima for the two models, equivalent to the difference between the deviance (which is twice the minimum log likelihood) of the two models:

$$\begin{aligned}T_{\rm LRT} &= -2\log\frac{p\left(\boldsymbol{I}|\hat{\boldsymbol{\theta}}^0_{\rm MLE}, H_0\right)}{p\left(\boldsymbol{I}|\hat{\boldsymbol{\theta}}^1_{\rm MLE}, H_1\right)} \\ &= D_{\rm min}(H_0) - D_{\rm min}(H_1).\end{aligned} \tag{22}$$

Asymptotic theory shows that, given certain regularity conditions are met, this statistic should be distributed as a chi square variable, $T_{\rm LRT} \sim \chi^2_\nu$, where the number of degrees of freedom ν is the difference between the number of free parameters in H_1 and H_0. When the regularity conditions are not met (see Freeman et al. 1999; Protassov et al. 2002; Park et al. 2008), we do not expect the distribution to be that of the asymptotic theory. Nevertheless, the LRT is a powerful statistic for comparing models and can be calibrated by posterior predictive simulation, as shown by Protassov et al. (2002) and Rubin & Stern (1994).

9 APPLICATION TO AGN DATA

In this section we apply the method detailed above to two example data sets, both long observations of nearby, variable Seyfert 1 galaxies, obtained from the *XMM–Newton* Science Archive.[7]

9.1 The power spectrum model

We shall restrict ourselves to two simple models for the high-frequency power spectrum of the Seyferts. The first (H_0) is a power law plus a constant (to account for the Poisson noise in the detection process)

$$S(f) = \beta f^{-\alpha} + \gamma \tag{23}$$

with three parameters $\boldsymbol{\theta} = \{\alpha, \beta, \gamma\}$, where β (the power-law normalization) and γ (the additive constant) are constrained to be non-negative. The second model (H_1) is a bending power law as advocated by McHardy et al. (2004):

$$S(f) = \beta f^{-1} \left(1 + \left\{ \frac{f}{\delta} \right\}^{\alpha - 1} \right)^{-1} + \gamma \tag{24}$$

with four parameters $\boldsymbol{\theta} = \{\alpha, \beta, \gamma, \delta\}$. For this model, β, γ and δ(the bending frequency) are all non-negative. The parameter α gives the slope at high frequencies ($f \gg \delta$) in model H_1, and the low-frequency slope is assumed to be – 1. (This assumption simplifies the model fitting process, and seems reasonable given the results of Uttley, McHardy & Papadakis 2002; Markowitz et al. 2003; McHardy et al. 2004, 2006, but could be relaxed if the model checking process indicated a significant model misfit.) In the limit of $\delta \to 0$, the form of H_1 tends to that of the simple power law H_0.

Following the advice given in Gelman et al. (2004), we apply a logarithmic transformation to the non-negative parameters. The motivation for this is that the posterior should be more symmetric (closer to normal), and so easier to summarize and handle in computations, if expressed in terms of the transformed parameters. We assign a uniform (uninformative) prior density[8] to the transformed parameters, e.g. $p(\alpha, \log\beta, \log\gamma) = \text{const}$ for model H_0. This corresponds to a uniform prior density on the slope α and a Jeffreys prior on the parameters restricted to be non-negative [e.g. $p(\beta) = 1/\beta$], which is the conventional prior for a scale factor (Sivia 1996; Gelman et al. 2004; Lee 2004; Gregory 2005; Albert 2007).

9.2 Application to *XMM–Newton* data of RE J1034+396

The first test case we discuss is the interesting *XMM–Newton* observation of the ultrasoft Seyfert 1 galaxy RE J1034 + 396. Gierliński et al. (2008) analysed these data and reported the detection of a significant QPO which, if confirmed in repeat observations and by independent analyses, would be the first robust detection of its kind. For the present analysis a 0.2–10 keV time series was extracted from the archival data using standard methods (e.g. Vaughan et al. 2003) and binned to 100 s, to match that used by Gierliński et al. (2008).

[7] See `http://xmm.esac.esa.int/`

[8] Although strictly speaking these prior densities are improper, meaning they do not integrate to unity, we may easily define the prior density to be positive only within some large but reasonable range of parameter values, and zero elsewhere, and thereby arrive at a proper prior density. In the limit of large N, the likelihood will dominate over the uninformative prior and hence the exact form of the prior density will become irrelevant to the posterior inferences.

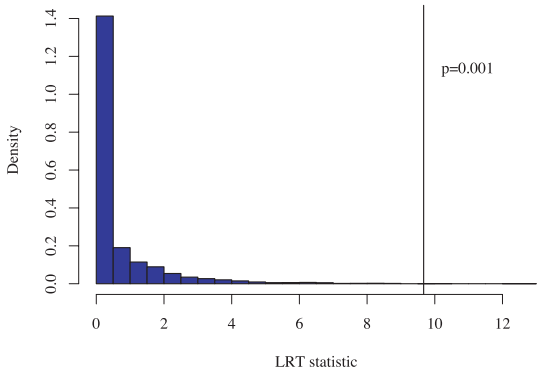

Figure 1. Posterior predictive distribution of the LRT statistic under H_0 for the RE J1034+396 data (computed using 5000 data sets simulated under H_0). The observed value $T_{\rm LRT}^{\rm obs} = 9.67$ is shown with the vertical line, alongside the corresponding p-value. The distribution is not χ^2_1 as would be predicted by the standard theory but instead resembles a mixture of distributions with half the probability in a χ^2_1 distribution and half concentrated around zero. This might be expected given the arguments of Titterington, Smith & Makov (1985, section 5.4) and Protassov et al. (2002, Appendix B).

Table 2. Posterior summaries of parameters for model H_1 for the RE J1034+396 data.

Parameter	Mean	5 per cent	95 per cent
α	3.4	2.2	5.2
β	2.3×10^{-3}	1.4×10^{-3}	3.7×10^{-3}
γ	0.40	0.34	0.45
δ	4.3×10^{-4}	2.0×10^{-4}	6.5×10^{-4}

Note. The four parameters are as follows: α = power-law index, β = normalization (in power density units at 1 Hz, i.e. [rms/mean]^2 Hz^{-1}), γ (Poisson noise level in power density units, [rms/mean]^2 Hz^{-1}), δ(bend frequency in Hz). The columns give the parameter name, the posterior mean and the lower and upper bounds of the 90 per cent credible intervals.

The two candidate continuum models discussed above, H_0 and H_1, were compared to the data, which gave $D_{\rm min}^{\rm obs}\ (H_0) = 504.89$ and $D_{\rm min}^{\rm obs}\ (H_1) = 495.22$, therefore $T_{\rm LRT}^{\rm obs} = 9.67$. The MCMC was used to draw from the posterior of model H_0, and these draws were used to generate posterior predictive periodogram data, which were also fitted with the two models and the results used to map out the posterior predictive distribution of $T_{\rm LRT}$, which is shown in Fig. 1. The corresponding tail area probability for the observed value is $p = 0.001$, small enough that the observed reduction in $D_{\rm min}$ between H_0 and H_1 is larger than might be expected by chance if H_0 were true. We therefore favour H_1 and use this as the continuum model. In the absence of complicating factors (see below), this amounts to a significant detection of a power spectral break.

Using H_1 as the continuum model, we then map out the posterior distribution of the parameters using another MCMC sample. Table 2 presents the posterior means and intervals for the parameters of model H_1, and Fig. 2 shows the pairwise marginal posterior distributions for the parameters of the model. Fig. 3 shows the data and model evaluated at the posterior mode.

Clearly there is a large outlier at $\sim 2.5 \times 10^{-3}$ Hz in the residuals after dividing out the model (H_1, computed at the posterior mode)

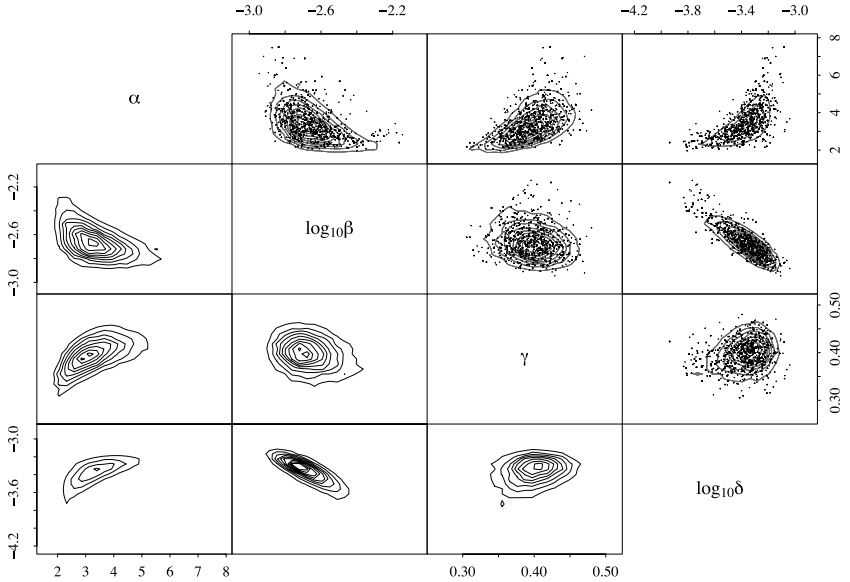

Figure 2. Pairwise marginal posterior distributions for the parameters of H_1: α = power-law index, β = normalization (in power density units at 1 Hz, i.e. [rms/mean]^2 Hz^{-1}), γ (Poisson noise level in power density units, [rms/mean]^2 Hz^{-1}), δ(bend frequency in Hz). The parameters β and δ are shown on a logarithmic scale. The lower left panels show the contours evaluated using all 75 000 posterior simulations, and the upper left panels show some of the simulated posterior data (for clarity only 1000 points are shown).

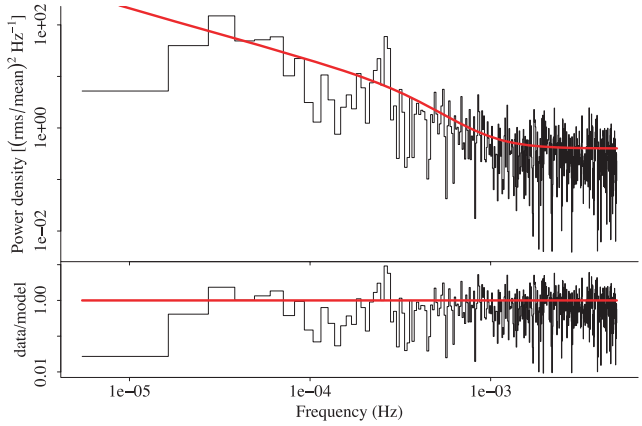

Figure 3. RE J1034+396 data and model (H_1) computed at the posterior mode. The data are shown as the histogram and the model is shown with the smooth curve. The lower panel shows the data/model residuals on a logarithmic scale. See Gierliński et al. (2008) for details of the observation.

which may be due to additional power from a QPO. We therefore calculate the posterior predictive distributions of the two test statistics $T_{\rm R}$ and $T_{\rm SSE}$ and compared these to the observed values ($T_{\rm R}^{\rm obs} = 18.41$ and $T_{\rm SSE}^{\rm obs} = 542.3$). The posterior predictive distributions of these two statistics, derived from 5000 simulations, are shown in Fig. 4. Both these statistics give moderately low p-values ($p_{\rm R} = 0.035$ and $p_{\rm SSE} = 0.025$), indicating there is room for improvement in the model and that the largest outlier is indeed rather unusual under H_1. This may indicate the presence of power from a QPO or some other deficiency in the continuum model. Very similar results were obtained after repeating the posterior predictive p-value calculations with a variant of H_1 in which the low-frequency index (at $f \ll \delta$) is fixed at 0 rather than -1, indicating that the p-values are not very sensitive to this aspect of the continuum model.

Gierliński et al. (2008) split the time series into two segments and focused their analysis on the second of these, for which the periodogram residual was largest and concentrated in one frequency bin only. The division of the data into segments is based on a

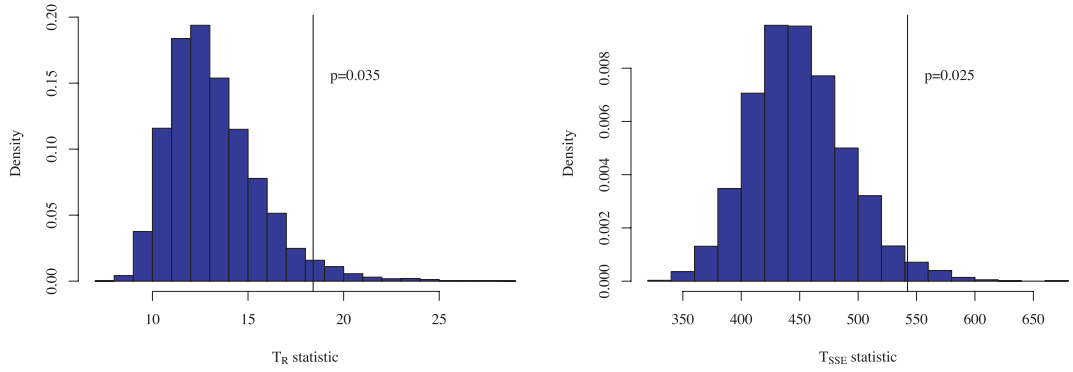

Figure 4. Posterior predictive distributions of the $T_{\rm R}$ and $T_{\rm SSE}$ statistics under H_1 for the RE J1034+396 data. The observed value of each is shown with a vertical line.

Table 3. Posterior summaries of parameters for model H_1 for the Mrk 766 data. The columns are as in Table 2.

Parameter	Mean	5 per cent	95 per cent
α	2.7	2.4	3.1
β	1.6×10^{-2}	0.95×10^{-2}	2.7×10^{-2}
γ	0.10	0.084	0.12
δ	2.1×10^{-4}	0.97×10^{-4}	3.4×10^{-4}

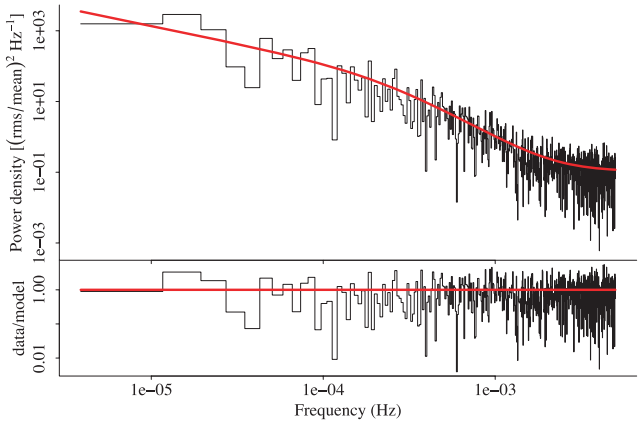

Figure 5. Mrk 766 data and model (H_1) computed at the posterior mode. The panels are the same as in Fig. 3.

partial analysis of the data – it is in effect the application of a data-dependent 'stopping rule' – and it is extremely difficult to see how such a procedure could be included in the generation of replicated data $\boldsymbol{I}^{\rm rep}$ used to calibrate the posterior predictive p-values. We therefore consider p-values only for the analysis of the entire time series and do not try to replicate exactly the analysis of Gierliński et al. (2008).

9.3 Application to *XMM–Newton* data of Mrk 766

A similar analysis was performed on the *XMM–Newton* observation of Mrk 766 discussed previously by Vaughan & Fabian (2003), who claimed to have detected a power spectral break using frequentist (classical) statistical tools such as χ^2 fitting. The LRT statistic for the data was $T_{\rm LRT}^{\rm obs} = 18.56$, and the posterior predictive distribution for this statistic had the same shape as in the case of RE J1034 + 396 (Fig. 1). The p-value for the LRT comparison between H_0 and H_1 was $p < 2 \times 10^{-4}$ (i.e. not one of the 5000 simulations gave a larger value of $T_{\rm LRT}$). This amounts to a very strong preference for H_1 over H_0, i.e. a solid detection of a spectral break.

Table 3 summarizes the posterior inferences for the parameters of H_1 and Fig. 5 shows the data, model and residuals. The residuals show no extreme outliers, and indeed the observed values of the test statistics $T_{\rm R}$ and $T_{\rm SSE}$ were not outliers in their posterior predictive distributions ($p_{\rm R} = 0.93$ and $p_{\rm SSE} = 0.89$). These suggest that H_1 provides an adequate description of the data (i.e. without any additional components).

9.4 Sensitivity to choice of priors

It is important to check the sensitivity of the conclusions to the choice of the prior densities, by studying, for example, the effect of a different or modified choice of prior on the posterior inferences. We have therefore repeated the analysis of the RE J1034 + 396 data using a different choice of priors. In particular, we used independent normal densities on the four transformed parameters of H_1, this is equivalent to a normal density on the index α and lognormal densities on the non-negative valued parameters β, γ and δ. In other words, for each of the transformed parameters $p(\theta_i|H_1) = N(\mu_i, \sigma_i^2)$, where the *hyperparameters* μ_i and σ_i control the mean and width of the prior density functions. After choosing values for the hyperparameters based on knowledge gained from previous studies of nearby, luminous Seyfert galaxies (e.g. Uttley et al. 2002; Markowitz et al. 2003; Papadakis 2004; McHardy et al. 2006), as outlined below, the posterior summaries (parameter means and intervals, pairwise marginal posterior contours, and posterior predictive p-values) were essentially unchanged, indicating that the inferences are relatively stable to the choice of prior.

Previous studies usually gave a high-frequency index parameter in the range $\alpha \sim 1$–3, and so we assigned $p(\alpha|H_1) = N(2, 4)$, i.e. a prior centred on the typical index of 2 but with a large dispersion (standard deviation of 2). The normalization of the f^{-1} part of the power spectrum is thought to be similar between different sources, with $\beta \sim 0.005$–0.03 (see Papadakis 2004), we assigned $p(\log\beta|H_1) = N(-2, 1)$, i.e. a decade dispersion around the mean of $\beta \sim 10^{-2}$. The Poisson noise level is dependent on the count rate, which can be predicted very crudely based on previous X-ray observations; we assign a prior $p(\log\gamma|H_1) = N(0, 1)$. The bend/break frequency δ is thought to correlate with other system parameters such as $M_{\rm BH}$, bolometric luminosity $L_{\rm Bol}$ and optical line width (e.g. FWHMHβ). Using the estimated luminosity, and assuming RE J1034 + 396 is radiating close to the Eddington limit (Middleton et al. 2009) gave a prediction for the bend time-scale of $T_{\rm b} \sim 1.6 \times 10^{-3}$ s, and using the optical line width of Véron-Cetty, Véron & Gonçalves (2001) gave $T_{\rm b} \sim 1.2 \times 10^{-3}$ s, using the relations of McHardy et al. (2006). Both these (independent) predictions suggest $\delta = 1/T_{\rm b} \sim 10^{-3}$ Hz, and we therefore assigned a prior density $p(\log\delta|H_1) = N(-3, 1)$. All of these priors are reasonably non-informative – they have quite large dispersion around the mean values, to account for the fact that the empirical relations used make these predictions are rather uncertain themselves and also contain intrinsic scatter (i.e. there are significant source to source differences) – yet they do include salient information about the model obtained from other sources.

10 DISCUSSION

We have described, in Sections 6–8, a Bayesian analysis of periodogram data that can be used to estimate the parameters of a power spectral model of a stochastic process, compare two competing continuum models and test for the presence of a narrow QPO (or strict periodicity).

10.1 Limitations of the method

The Whittle likelihood function (equation 16) is only an approximation to the true sampling distribution of a periodogram. In the absence of distortions due to the sampling window (more on this below), the ordinates of the periodogram of all stationary, linear (and many non-linear) stochastic processes become independently distributed following equation (13) as $N \to \infty$. With finite N (i.e. for real data) this is only approximately true, although with reasonable sample sizes (e.g. $N > 100$) it is a very good approximation.

More serious worries about the distribution of the periodogram, and hence the validity of the Whittle likelihood, come from

distortions due to the sampling effects known as aliasing and leakage (e.g. Uttley et al. 2002). It is fairly well established that X-ray light curves from Seyfert galaxies are stationary once allowance has been made for their red noise character and the linear 'rms-flux' relation (see Vaughan et al. 2003; Uttley, McHardy & Vaughan 2005). Distortions in the expectation of the periodogram can be modelled by simulating many time series for a given power spectral model, resampling these in time as for the original data, and then calculating the average of their periodograms (Uttley et al. 2002, and Appendix B). This does not account for distortions in the distribution of the periodogram ordinates (away from equation 13 predicted by asymptotic theory), which is a more challenging problem with (as yet) no accepted solution. However, these effects will be minor or negligible for the data analysed in Section 9 which are contiguously binned, as the effect of aliasing will be lost in the Poisson noise spectrum which dominates at high frequencies (van der Klis 1989; Uttley et al. 2002), and the leakage of power from lower to higher frequencies is very low in cases where the power spectrum index is $\alpha \lesssim 1$ at the lowest observed frequencies. The task of fully accounting for sampling distortions in both the expectation and distribution of the periodogram, and hence having a more general likelihood function, is left for future work.

We should also point out that the usual limitations on the use and interpretation of the periodogram apply. These include the (approximate) validity of the Whittle likelihood only when the time series data are evenly sampled. It may be possible to adjust the likelihood function to account for the non-independence of ordinates in the modified periodogram usually used with unevenly sampled time series (e.g. Scargle 1982), but here we consider only evenly sampled data. It is also the case that the periodogram, based on a decomposition of the time series into sinusoidal components, is most sensitive to sinusoidal oscillations, especially when they lie close to a Fourier frequency (i.e. the time series spans an integer number of cycles; see van der Klis 1989). In situations where the time series is large and spans many cycles of any possible periods (the large N regime), there is no reason to go beyond the standard tools of time series processing such as the (time and/or frequency) binned periodogram with approximately normal error bars (van der Klis 1989). The current method uses the raw periodogram of a single time series (with the Whittle likelihood) in order to preserve the frequency resolution and bandpass of the data, which is more important in the low N regime (e.g. when only a few cycles of a suspected period are observed).

The time series data analysed in Section 9 were binned up to 100 s prior to computing the periodogram; this in effect ignores frequencies above 5×10^{-3} Hz which are sampled by the raw data from the detectors (recorded in counts per CCD frame at a much higher rate). The choice of bin size does affect the sensitivity to periodic signals of the method described in Sections 6–8. Obviously, one looses sensitivity to periodic components at frequencies higher than the Nyquist frequency. But also as more frequencies are included in the analysis, there are more chances to find high $T_{\rm R}$ values from each simulation, which means the posterior predictive distribution of the test statistic does depend on the choice of binning.

One could mitigate against this by imposing a priori restrictions on the frequencies of any allowed periods, for example by altering the test statistic to be $T_{\rm R} = \max_{j<J_0} R_j$ where J_0 is some upper limit. (The lower frequency of the periodogram is restricted by the duration of the time series, which is often dictated by observational constraints.) But these must be specified independently of the data, otherwise this is in effect another data-dependent stopping rule (the effect of limiting the frequency range of the search is illustrated below in the case of the RE J1034 + 396). This sensitivity to choice of binning could be handled more effectively by considering the full frequency range of the periodogram (i.e. no rebinning of the raw data) and explicitly modelling the periodic component of the spectrum with an appropriate prior on the frequency range (or an equivalent modelling procedure in the time domain). But this suffers from the practical drawbacks discussed below.

10.2 Alternative approaches to model selection

In many settings, the LRT (or the closely related F-test) is used to choose between two competing models: the observed value of the LRT statistic is compared to its theoretical sampling (or reference) distribution, and this is usually summarized with a tail area probability, or p-value. As discussed above, this procedure is not valid unless certain specific conditions are satisfied by the data and models. In the case of comparing a single power law (H_0 of section 9) to a bending power law (H_1), the simpler model is reproduced by setting the extra parameter $\delta \rightarrow 0$ in the more complex model, which violates one of the conditions required by the LRT (namely that null values of the extra parameters should not lie at the boundaries of the parameter space). In order to use the LRT, we must find the distribution of the statistic appropriate for the given data and models, which can be done using posterior predictive simulations. This method has the benefit of naturally accounting for nuisance parameters by giving the expectation of the classical p-value over the posterior distribution of the (unknown) nuisance parameters.

One could, in principle, use the posterior predictive checks to compare a continuum only model (e.g. H_0 or H_1) to a continuum plus line (QPO) model (H_2) and thereby test for the presence of an additional QPO. Protassov et al. (2002) and Park et al. (2008) tackled just this problem in the context of X-ray energy spectra with few counts. However, we deliberately do not define and use a model with an additional line for the following reasons. First, this would require a specific line model and a prior density on the line parameters, and it is hard to imagine these being generally accepted. Unless the line signal is very strong the resulting posterior inferences may be more sensitive to the (difficult) choice of priors than we would generally wish. Secondly, as shown by Park et al. (2008), there are considerable computational difficulties when using models with additional, narrow features and data with high variance (as periodograms invariably do), due to the highly multimodal structure of the likelihood function. Our pragmatic alternative is to leave the continuum plus line model unspecified, but instead choose a test statistic that is particularly sensitive to narrow excesses in power such as might be produced under such a model (see Gelman et al. 1996, and associated discussions, for more on the choice of test statistic in identifying model deficiency). This has the advantages of not requiring us to specify priors on the line parameters and simplifying the computations, but means the test is only sensitive to specific types of additional features that have a large effect on the chosen test statistic. (It is also worth pointing out that the periodogram ordinates are randomly distributed about the spectrum of the stocastic process $S_R(f)$. If a deterministic process is also present, e.g. producing a strictly periodic component to the signal, this will not in general follow the same χ^2 distribution and the Whittle likelihood function would need to be modified in order to explicitly model such processes in the spectral domain.)

One of the most popular Bayesian methods for choosing between competing models is the Bayes factor (Kass & Raftery 1995;

Carlin & Louis 2000; Gelman et al. 2004; Lee 2004). These provide a direct comparison of the weight of evidence in favour of one model compared to its competitor, in terms of the ratios of the marginal likelihoods for the two models (equation 7). This may be more philosophically attractive than the posterior predictive model checking approach but in practice suffers from the same problems outlined above, namely the computational challenge of handling a multimodal likelihood, and the sensitivity to priors on the line parameters, which may be even greater for Bayes factors than other methods (see arguments in Protassov et al. 2002; Gelman et al. 2004).

10.3 Comparison with V05

V05 tackled the same problem – the assessment of red noise spectra and detection of additional periodic components from short time series – using frequentist methods. The method developed in the present paper is superior in a number of ways. The new method is more general in the sense that the model for the continuum power spectrum (i.e. the 'null hypothesis' model that contains no periodicities) may in principle take any parametric form but was previously restricted to a power law. It also provides a natural framework for assessing the validity of the continuum model, which should be a crucial step in assessing the evidence for additional spectral features (see below). Also, by using the Whittle likelihood rather than the Geweke & Porter-Hudak (1983) fit function, the new method actually gives smaller mean square errors on the model parameters (see Andersson 2002).

10.4 Comparison with other time series methods

Previous work on Bayesian methods for period detection (e.g. Bretthorst 1988; Gregory & Loredo 1992; Gregory 1999) has focused on cases where the stochastic process is assumed to be white (uncorrelated) noise on which a strictly periodic signal is superposed. They do not explicitly tackle the more general situation of a non-white continuum spectrum that is crucial to analysing data from compact accreting X-ray sources.

The only non-Bayesian (i.e. frequentist) methods we are aware of for assessing evidence for periodicities in data with a non-white spectrum involve applying some kind of smoothing to the raw periodogram data. This gives a non-parametric estimate of the underlying spectrum, with some associated uncertainty on the estimate, which can then be compared to the unsmoothed periodogram data and used to search for outlying periodogram points. The multitaper method (MTM) of Thomson (1982) (also Thompson 1990) achieves the smoothing by averaging the multiple periodograms, each computed using one member of a set of orthogonal data tapers. See Percival & Walden (1993, chapter 7) for a good discussion of this method. The data tapers are designed to reduce spectral leakage and so reduced bias in the resulting spectrum estimate. The method proposed by Israel & Stella (1996) involves a more straightforward running mean of the peridogram data. Both of these are non-parametric methods, meaning that they do not involve a specific parametric model for the underlying spectrum. This lack of model dependence might appear to be an advantage, but in fact may be a disadvantage in cases where we do have good reasons for invoking a particular type of parametric model (e.g. the bending power laws seen in the Seyfert galaxy data). The continuum model's few parameters may be well constrained by the data, where the non-parametric (smoothed) estimate at each frequency is not. The non-parametric methods also leave a somewhat arbitrary choice of how to perform the smoothing, i.e. the type and number of data tapers in the MTM, or the size/shape of the smoothing kernel in the Israel & Stella (1996) method. Also, it is less obvious how to combine the sampling distribution of the periodogram ordinate (line component) and the spectrum estimate (continuum), and how to account for the number of 'independent' frequencies searched. These are all automatically included in the posterior predictive p-value method as outlined above.

In the present paper, we have deliberately concentrated on the periodogram since this is the standard tool for time series analysis in astronomy. But the periodogram is by no means the best or only tool for the characterization of stochastic processes or the identification of periodicities. Methods that explicitly model the original time series data in the time domain (see e.g. Priestley 1981; Chatfield 2003) may yet prove to be valuable additions to the astronomers toolkit. Indeed the raw form of the *XMM–Newton* data used in the AGN examples is counts per CCD frame, for the source (and possibly background region if this is a non-negligible contribution). The most direct data analysis would therefore model this process explicitly as a Poisson process with a rate parameter that varies with time (i.e. the 'true' X-ray flux) that is itself a realization of some stochastic process with specific properties (e.g. power spectrum or, equivalently, autocorrelation function, and stationary distribution).

10.5 The importance of model assessment

The posterior predictive approach provides an attractive scheme for model checking. In particular, it allows us to select a continuum model that is consistent with the observed data[9] before testing for the presence of additional features. This is crucial since any simple test statistic, whether used in a frequentist significance test or a posterior predictive test, will be sensitive to certain kinds of deficiencies in the model without itself providing any additional information about the specific nature of any deficiency detected (a p-values is after all just a single number summary). A low p-value (i.e. a 'significant' result) may be due to the presence of interesting additional features or just an overall poor match between the data and the continuum model (for more on this in the context of QPO detection see Vaughan & Uttley 2006). The use of more than one test statistic, properly calibrated using the posterior predictive simulations, as well as other model diagnostics (such as data/model residual plots) are useful in identifying the cause of the data/model mismatch.

10.6 Analysis of two Seyfert galaxies

Section 9 presents an analysis of *XMM–Newton* data for the Seyfert galaxies RE J1034+396 and Mrk 766. The former has produced the best evidence to date for a QPO in a Seyfert galaxy (Gierliński et al. 2008), while the latter showed no indication of QPO behaviour (Vaughan & Fabian 2003; Vaughan & Uttley 2005). Gierliński et al. (2008) used the method presented in V05 to show that the observed peak in the periodogram was highly unlikely under the assumption than the underlying power spectrum continuum is a power law, but the present analysis gave somewhat less impressive evidence to suggest a QPO.

[9] Strictly, we compare the observed data to simulations drawn from the posterior predictive distribution under the chosen model H using test statistics. If the observed data do not stand out from the simulations, by having extreme values of the statistics when compared to the simulations, we may assume that the data are consistent with the model (as far as the particular test statistics are concerned).

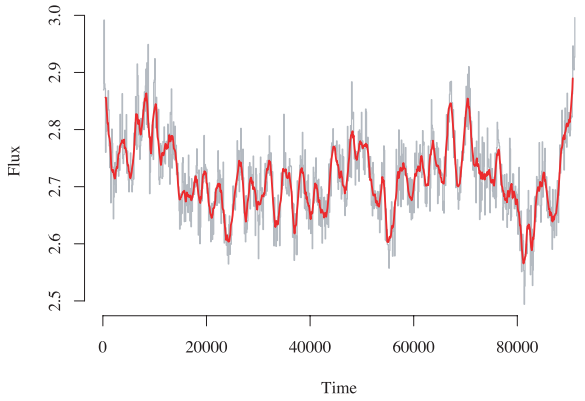

Figure 6. Simulated time series generated from the posterior predictive distribution of the RE J1034+396 periodogram data. The (grey) histogram shows the simulated data in 100 s bins and the smooth (red) curve shows the 6 bin moving average of these data. Compare with fig. 1 of Gierliński et al. (2008). The power spectrum used to generate these data is a smoothly bending power law (plus white 'measurement' noise) with no periodic or quasi-periodic components, and there the time series appears to show oscillatory structure.

The posterior predictive $p \approx 0.03$ comes from the fact that ∼150 out of the 5000 posterior predictive simulations of the RE J1034+396 periodogram data showed $T_{\rm R} \geq T_{\rm R}^{\rm obs}$ (and approximately the same figure was obtained using $T_{\rm SSE}$). This might at first seem doubtful given how periodic the observed time series appears (see fig. 1 of Gierliński et al. 2008). But to demonstrate that such apparently periodic time series may indeed be generated from non-periodic processes, we simulated time series from the posterior predictive periodogram data (for model H_1) that showed $T_{\rm R} \geq T_{\rm R}^{\rm obs}$. (The time series simulation method is given in Appendix B.) One example of these time series, chosen at random from the subset that had the largest residual R_j occurring at a frequency of the same order as that seen in RE J1034+396 (in this case $\approx 1.3 \times 10^{-4}$ Hz), is shown in Fig. 6.

There are several reasons for the very different p-values between the analyses. One of these factors is that we based our calculation on a more general form of the continuum model. In the absence of a QPO (spectral line component), the power spectrum continuum is well modelled using a power law with a steep slope ($\alpha \sim 3$) that smoothly changes to a flatter slope (assumed index of -1) below a frequency $\delta \sim 4 \times 10^{-4}$ Hz, than a single power law. The bend frequency is close to that of the candidate QPO, which does have a large effect on the 'significance' of the QPO as summarized in the p-value (see Vaughan & Uttley 2005, for previous examples of this effect). Indeed, the posterior predictive p-value was 2×10^{-3} when recalculated assuming a simple power-law continuum (H_0). A second factor is that Gierliński et al. (2008) gave special consideration to a particular subset of the times series chosen because of its apparently coherent oscillations, which in effect enhanced the apparent significance of the claimed periodicity, while the entire time series is treated uniformly in the present analysis (for reasons discussed in Section 9). A third factor is that we made no restriction on the allowed frequency of a period component, and so openly searched 457 frequencies, where Gierliński et al. (2008) concentrated on the ≈60 frequencies in their periodogram below 10^{-3} Hz. This will result in a factor ∼8 change in the p-value (since the probability of finding a $T_{\rm R}$ value in a simulation that is larger than the observed in the real data scales approximately linearly with the number of frequencies examined). If we take $T_{\rm R}$ to be the largest residual at frequencies below 10^{-3} Hz (but including all the data in the rest of the modelling process), we find 21/5000 of the RE J1034+396 simulations showed $T_{\rm R} \geq T_{\rm R}^{\rm obs}$ under these restricted conditions, corresponding to $p = 0.004$, which is smaller by about the expected factor. A relatively minor difference is the more complete treatment of parameter uncertainties using the posterior distribution (which is treated in an approximation fashion in the method of V05). One is therefore left with a choice between two models that could plausibly explain the data, a power-law spectrum with a strong QPO or a bending power-law spectrum (with weaker evidence for a QPO). The most powerful and least ambiguous confirmation of the reality of the QPO feature would come from an independent observation capable of both constraining the continuum more precisely and allowing a sensitive search for the candidate QPO.

The results of the present analysis of the Mrk 766 data agree reasonably well with those previously reported by Vaughan et al. (2003) which were obtained using standard frequentist methods (e.g. binning the data and estimating parameters by minimizing the χ^2 statistic). The high-frequency slopes are essentially the same, but the frequency of the bend differs by a factor of ∼2.5. This is most likely due to the slightly different models used, i.e. bending versus sharply broken power laws. [Repeating the frequentist analysis of Vaughan et al. (2003) using the bending power-law model gave a lower characteristic frequency, more consistent with that of the present analysis].

10.7 Other applications of this method

The techniques discussed in this paper may find application well beyond the specific field for which they were devised (namely, analysis of X-ray light curves from Seyfert galaxies), since the problems of estimating a noisy continuum spectrum and assessing the evidence for additional narrow features over and above that continuum are common to many fields. Other examples from X-ray astronomy include analysis of long time-scale light curves from Galactic X-ray binaries and ultraluminous X-ray sources (ULXs) in order to characterize the low-frequency power spectrum and search for periodicities (e.g. due to orbital modulation).

But the applications are by no means restricted to astronomy. For example, in geology, there is considerable interest in detecting and characterizing periodicities in stratigraphic records of environmental change, which may be connected to periodicities in external forcing such as might be expected from Milankovich cycles (see e.g. Weedon 2003). However, there is controversy over the statistical and physical significance of the periodicities in these data, which are often dominated by stochastic red noise variations (Bailey 2009).

11 CONCLUSIONS

We have presented Bayesian methods for the modelling of periodogram data that can be used for both parameter estimation and model checking, and may be used to test for narrow spectral features embedded in noisy data. The model assessment is performed using simulations of posterior predictive data to calibrate (sensibly chosen) test statistics. This does however leave some arbitrariness in the method, particularly in the choice of test statistic[10] (and in some situations the choice of what constitutes a simulation

[10] In situations where two competing models can be modelled explicitly, the LRT provides a natural choice of statistic.

of the data). Such issues were always present, if usually ignored, in the standard frequentist tests. The posterior predictive approach has the significant advantage of properly treating nuisance parameters, and provides a clear framework for checking the different aspects of the reasonableness of a model fit. The issue of choosing a test statistic does not arise in more 'purist' Bayesian methods such as Bayes factors, which concentrate on the posterior distributions and marginal likelihoods, but such methods of model selection carry their own burden in terms of the computational complexity and the difficulty of selecting (and the sensitivity of inferences to) priors on the model parameters. The method presented in this paper, making use of the posterior predictive checking, is an improvement over the currently popular methods that use classical p-value; but Bayesian model selection is an area of active research and it is not unreasonable to expect that new, powerful and practical computational tools will be developed or adapted to help bridge the gap between the pragmatic and the purist Bayesian approaches.

The routines used to perform the analysis of the real data presented in Section 9 will be made available as an R[11] script from the author upon request.

ACKNOWLEDGMENTS

The author wishes to thank David van Dyk and Phil Uttley for valuable discussions during the final stages of writing this paper, and an anonymous referee for a helpful report.

REFERENCES

Albert J., 2007, Bayesian Computation with R. Springer, New York
Anderson E. R., Duvall T. L. Jr., Jefferies S. M., 1990, ApJ, 364, 699
Andersson J., 2002, Econ. Lett., 77, 137
Bailey R. J., 2009, Terra Nova, 21, 340
Berger J. O., Berry D. A., 1988, Am. Sci., 76, 159
Berry D. A., 1996, Statistics: A Bayesian Perspective. Duxbury, London
Bloomfield P., 2000, Fourier Analysis of Time Series: An Introduction, 2nd edn. Wiley, New York
Bretthorst G. L., 1988, Bayesian Spectrum Analysis and Parameter Estimation, Lecture Notes in Statistics. Springer-Verlag, Heidelberg
Carlin B. P., Louis T. A., 2000, Bayes and Empirical Bayes Methods for Data Analysis, 2nd edn. Chapman & Hall/CRC, London
Chatfield C., 2003, The Analysis of Time Series: An Introduction. Chapman & Hall/CRC, London
Chib S., Greenberg E., 1995, Am. Stat., 49, 327
Contreras-Cristán E., Gutiérrez-Peña E., Walker S. G., 2006, Commun. Stat. - Simul. Comput., 35, 857
Cowan G., 1998, Statistical Data Analysis. Clarendon Press, Oxford
Davies R. B., Harte D. S., 1987, Biometrica, 74, 95
Eadie W. T., Drijard D., James F. E., Roos M., Sadoulet B., 1971, Statistical Methods in Experimental Physics. North-Holland, Amsterdam
Fan J., Zhang W., 2004, Biometrika, 91, 195
Fisher R. A., 1929, Proc. R. Soc. A, 125, 54
Freeman P. E., Graziani C., Lamb D. Q., Loredo T. J., Fenimore E. E., Murakami T., Yoshida A., 1999, ApJ, 524, 753
Gamerman D., 1997, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman and Hall/CRC, London
Gelman A., Rubin D. B., 1992, Stat. Sci., 7, 457
Gelman A., Meng X.-L., Stern H. S., 1996, Stat. Sinica, 6, 733
Gelman A., Carlin J. B., Stern H. S., B. R. D., 2004, Bayesian Data Analysis, 2nd edn. Chapman & Hall, London
Geweke J., Porter-Hudak S., 1983, J. Time Ser. Anal., 4, 221
Gierliński M., Middleton M., Ward M., Done C., 2008, Nat, 455, 369
Gilks W. R., Richardson S., Spiegelhalter D., 1995, Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC, London
Gregory P. C., 1999, ApJ, 520, 361
Gregory P. C., 2005, Bayesian Logical Data Analysis for the Physical Sciences. Cambridge Univ. Press, Cambridge, UK
Gregory P. C., Loredo T. J., 1992, ApJ, 398, 146
Groth E. J., 1975, ApJS, 29, 285
Hannan E. J., 1973, J. Appl. Prob., 10, 130
Howson C., Urbach P., 1991, Nat, 350, 371
Israel G. L., Stella L., 1996, ApJ, 468, 369
Jeffreys W. H., Berger J. O., 1992, Am. Sci., 80, 64
Jenkins G. M., Watts D. G., 1969, Spectral Analysis and Its Applications. Holden-Day, London
Kass R. E., Raftery A. E., 1995, J. Am. Stat. Assoc., 90, 773
Kass R. E., Carlin B. P., Gelman A., Neal R. M., 1998, Am. Stat., 52, 93
Leahy D. A., Darbro W., Elsner R. F., Weisskopf M. C., Kahn S., Sutherland P. G., Grindlay J. E., 1983, ApJ, 266, 160
Lee P. M., 2004, Bayesian Statistics: An Introduction, 3rd edn. Wiley, New York
Loredo T. J., 1990, in Fougere P., ed., Maximum-Entropy and Bayesian Methods, Dartmouth. Kluwer Academic Publishers, Dordrecht, p. 81
Loredo T. J., 1992, in Feigelson D., Babu G., eds, Statistical Challenges in Modern Astronomy, Springer-Verlag, New York, p. 275
Markowitz A. et al., 2003, ApJ, 593, 96
McHardy I. M., Papadakis I. E., Uttley P., Page M. J., Mason K. O., 2004, MNRAS, 348, 783
McHardy I. M., Koerding E., Knigge C., Uttley P., Fender R. P., 2006, Nat, 444, 730
Meng X.-L., 1994, Ann. Stat., 22, 1142
Middleton M., Done C., Ward M., Gierliński M., Schurch N., 2009, MNRAS, 394, 250
Papadakis I. E., 2004, MNRAS, 348, 207
Park T., van Dyk D. A., Siemiginowska A., 2008, ApJ, 688, 807
Pawitan Y., O'Sullivan F., 1994, J. Am. Stat. Assoc., 89, 600
Percival D. B., Walden A. T., 1993, Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques. Cambridge Univ. Press, Cambridge
Pilgram B., Kaplan D. T., 1998, Phys. D, 114, 108
Press W. H., 1978, Comments Astrophys., 7, 103
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. Academic Press, London
Protassov R., van Dyk D. A., Connors A., Kashyap V. L., Siemiginowska A., 2002, ApJ, 571, 545
R Development Core Team 2009, R: A Language and Environment for Statistical Computing. http://www.R-project.org, Vienna, Austria
Rubin D. B., 1984, Ann. Stat., 12, 1151
Rubin D. B., Stern H. S., 1994, in von Eye A., Clogg C., eds, Latent Variables Analysis: Applications for Developmental Research Testing in Latent Class Models using a Posterior Predictive Check Distribution. Sage Publications, Thousand Oaks, CA, p. 420
Scargle J. D., 1982, ApJ, 263, 835
Sivia D. S., 1996, Data Analysis: A Bayesian Tutorial. Oxford Univ. Press, Oxford
Thomson D. J., 1982, Proc. IEEE, 70, 1055
Thompson D. J., 1990, Phil. Trans. R. Soc. Lond. A, 332, 539
Tierney L., 1994, Ann. Stat., 22, 1701
Timmer J., König M., 1995, A&A, 300, 707
Titterington D. M., Smith A. F. M., Makov U. E., 1985, Statistical Analysis of Finite Mixture Distributions. Wiley, New York
Trotta R., 2008, Contem. Phys., 49, 71
Uttley P., McHardy I. M., Papadakis I. E., 2002, MNRAS, 332, 231

[11] R is a powerful, open-source computing environment for data analysis and statistics that may be downloaded for free from http://www.r-project.org/ (R Development Core Team 2009; Venables & Smith 2009).

Uttley P., McHardy I. M., Vaughan S., 2005, MNRAS, 359, 345
van der Klis M., 1989, in Ögelman H., van den Heuvel E. P. J., eds, Timing Neutron Stars Fourier Techniques in X-ray Timing. Cambridge Univ. Press, Cambridge, p. 27
Vaughan S., 2005, A&A, 431, 391 (V05)
Vaughan S., Fabian A. C., 2003, MNRAS, 341, 496
Vaughan S., Uttley P., 2005, MNRAS, 362, 235
Vaughan S., Uttley P., 2006, Adv. Space Res., 38, 1405
Vaughan S., Edelson R., Warwick R. S., Uttley P., 2003, MNRAS, 345, 1271
Venables W. N., Smith D. M., 2009, An Introduction to R. http://cran.r-project.org/manuals.html
Véron-Cetty M.-P., Véron P., Gonçalves A. C., 2001, A&A, 372, 730
Weedon G. P., 2003, Time-Series Analysis and Cyclostratigraphy. Cambridge Univ. Press, Cambridge
Whittle P., 1953, Ark. Mat., 2, 423
Whittle P., 1957, J R. Statistical Soc. B, 19, 38

APPENDIX A: SIMULATING FROM THE POSTERIOR

Here, we briefly discuss a method for simulating data from the posterior density, which is useful for two main reasons. For simple models with few parameters, it may be possible to make inferences from the posterior without the need for Monte Carlo simulations, e.g. by directly evaluating the posterior density on a fine grid of parameter values. However, even in this case simulations from the posterior are needed in order to form the posterior predictive distribution, and hence the distribution of a test statistic and its posterior predictive p-value. For more complicated models or a greater number of parameters Monte Carlo methods may be necessary simply in order to calculate summaries of the posterior (such as means and intervals).

Markov Chain Monte Carlo (MCMC) methods provide a powerful and popular method for drawing random values from the posterior density. General introductions to MCMC computations for Bayesian posterior calculations are given by Gelman et al. (2004), Gregory (2005), Albert (2007), and more thorough treatments may be found in Tierney (1994), Chib & Greenberg (1995), Gilks, Richardson & Spiegelhalter (1995), Gamerman (1997).

The output of an MCMC calculation is a series of parameter values (or vectors) $\boldsymbol{\theta}^t$ for $t = 0, \ldots, L$ (where L is the number of simulations performed, i.e. the length of the chain). The Metropolis–Hastings MCMC algorithm generates a sequence of random draws as follows:

(i) Draw a starting point in parameter space $\boldsymbol{\theta}^0$ for which $p(\boldsymbol{\theta}|\boldsymbol{I}, H) > 0$.

(ii) Repeat for $t = 1, \ldots, L$:

(a) Draw a proposed new parameter point $\boldsymbol{\theta}^*$ from a *proposal* distribution $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})$ that is conditional only on the previous point $\boldsymbol{\theta}^{t-1}$.

(b) Evaluate the ratio

$$r = \frac{p(\boldsymbol{\theta}^*|\boldsymbol{I}, H)g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})}{p(\boldsymbol{\theta}^{t-1}|\boldsymbol{I}, H)g(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)}. \quad \text{(A1)}$$

(c) Set the new value of $\boldsymbol{\theta}^t$

$$\boldsymbol{\theta}^t = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}^{t-1} & \text{otherwise.} \end{cases} \quad \text{(A2)}$$

In order to use this algorithm, we need to have defined a proposal density function $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})$, from which we can compute densities and draw random values, and be able to evaluate the posterior density at any valid point in parameter space. Note that only the ratio of posterior densities need be calculated (to give r), meaning that we can use the unnormalized posterior density $q(\boldsymbol{\theta}|\boldsymbol{I}, H)$ in the computation. The remarkable property of the MCMC algorithm is that the distribution of the output $\boldsymbol{\theta}^t$ converges on the target distribution $p(\boldsymbol{\theta}|\boldsymbol{I}, H)$ for any form of proposal distribution (see Tierney 1994; Gilks et al. 1995, for regularity conditions).

The choice of proposal density does however affect the speed of convergence to the target distribution (i.e. the efficiency of the MCMC calculation) – the algorithm will be most efficient when the choice of proposal density closely matches the posterior density. We use as the proposal density a normal random walk, specifically a multivariate normal distribution centred on $\boldsymbol{\theta}^{t-1}$ with the covariance Σ from the normal approximation to the posterior (see Section 7). This is a popular and well understood choice of proposal and has been discussed extensively in the MCMC literature. In fact, it is usually better to use $g(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1}) = N(\boldsymbol{\theta}^{t-1}, c\Sigma)$, where c is a constant scale factor ($\gtrsim 1$) tuned to improve the efficiency of the calculation (see section 11.9 of Gelman et al. 2004). In the present analysis, we found $c = 1.2$ to work well. As the normal distribution is symmetric, i.e. $g(\boldsymbol{\theta}^i|\boldsymbol{\theta}^j) = g(\boldsymbol{\theta}^j|\boldsymbol{\theta}^i)$, the ratio r simplifies to the ratio to the posterior densities $r = q(\boldsymbol{\theta}^*|\boldsymbol{I}, H)/q(\boldsymbol{\theta}^{t-1}|\boldsymbol{I}, H)$. (We also found that a multivariate Student's t-distribution worked comparably well, with a covariance matrix $\Sigma\nu/(\nu-2)$, and $\nu = 3$ degrees of freedom.)

One must take some care to ensure the output of the MCMC has reached its *stationary* distribution and is efficiently generating draws from the complete posterior density. Gelman & Rubin (1992), Gilks et al. (1995), Kass et al. (1998) and Gelman et al. (2004) offer advice for checking the quality of the output. We calculate J separate chains, each starting from a different initial positions $\boldsymbol{\theta}^0$ spread over the parameter space, and check for convergence before merging the results. In order to remove any dependence on the initial position, we retain only the second half of each chain. We then compute the $\hat{R}$ statistic (Gelman & Rubin 1992; Gelman et al. 2004),[12] which compares the variance within each chain to the variance between the chains. With a sufficiently large number of iterations $\hat{R}$ will be close to unity, which indicates that the chains have reached the desired stationary distribution.

In practice, we discarded the first half of each chain (the 'burn in' phase) to remove any dependence on the starting position, and required $\hat{R} < 1.1$ from the remaining half of the chain before assuming the stationary distribution has been reached. We also inspect, for each chain, the acceptance rate, and the time series and histograms of the parameters, which may also reveal problems with convergence or efficiency of the chains. Once convergence has been reached, we combine the remaining $L/2$ points from each chain to yield $JL/2$ sets of parameter values sampled from the posterior distribution.

Given a sufficient number of simulations, $\boldsymbol{\theta}^t$, we can estimate the posterior distribution of any quantity of interest, such as the untransformed parameters, and from these estimate the posterior means (or modes or medians) and credible intervals by calculating the sample means and quantiles. If necessary, we can also simulate data $\boldsymbol{I}^{\rm rep}$ from the posterior predictive distribution by sampling parameters from the MCMC output, $\boldsymbol{\theta}^{\rm rep}$, and for each point calculating $S(\boldsymbol{\theta}^{\rm rep})$ and then randomly drawing periodogram points according to equation (13) (i.e. a scaled χ^2_2 distribution). We can then calculate the posterior distribution of any test statistic T using these simulated data, and hence calculate a p-value.

[12] In Gelman & Rubin (1992) and the first edition of the book Gelman et al. (2004), this statistic was called $\hat{R}^{1/2}$.

For the analysis of Section 9, we performed an initial fit to the data, using a non-linear optimization algorithm to find the posterior mode and covariance matrix, and used this to form the proposal distribution for the MCMC. Five chains of length 30 000 were generated from different locations randomly dispersed around the posterior mode. The first half of each chain (the 'burn-in phase') was discarded and, after checking convergence was achieved (as measured using the diagnostics discussed above), the remaining 75 000 values were merged into a single sample.

APPENDIX B: SIMULATION OF TIME SERIES

The calculation of a sample from the posterior, $p(\boldsymbol{\theta}|\boldsymbol{I}^{\rm obs}, H)$, requires only the output from an MCMC such as outlined above. Posterior predictive checks require the generation of simulated (replicated) periodogram data $\boldsymbol{I}^{\rm rep}$. The simplest approach is to draw a random parameter vector from the posterior, $\boldsymbol{\theta}^t$, generate the corresponding power spectral density function $S_j(\boldsymbol{\theta}^t)$ at frequencies $j = 1, \ldots, N/2 - 1$ and multiply each of these by a random draw from the χ^2_2 (exponential) distribution (equation 13):

$$I_j^{\rm rep} = S_j(\boldsymbol{\theta}^t) X_j/2, \tag{B1}$$

where X_j are independent random variables drawn from the χ^2_2 distribution. The periodogram at the Nyquist frequency $j = N/2$ should be multiplied by a random draw from a χ^2_1 distribution instead. (The zero frequency component can be safely given zero power.)

Time series may be generated by inverse Fourier transforming the randomized periodogram into the time domain (with time steps ΔT), with appropriate phase randomization. However, the resulting Fourier transformed data are strictly periodic with a period N, and so there is a wrap-around effect where the start and end of the time series are forced to converge. Also, this procedure does not include any effects due to transfer of power from frequencies just below or above the observed range. More realistic data may be generated from a posterior draw $\boldsymbol{\theta}^t$ by calculating a power spectrum over a wider range of frequencies than are included in the data, e.g. over a frequency grid $f_k = k/VN\Delta T$ with $k = 1, \ldots, K$, where $V \geq 1$ and $W \geq 1$ are the factors by which the lowest and highest frequencies are extended (respectively), and $K = VWN/2$. The power spectral densities $S_k(\boldsymbol{\theta}^t)$ may then be used in the algorithm of Davies & Harte (1987)[13] to produce a time series.

[13] An equivalent algorithm for generating time series from a power spectrum was introduced to astronomy by Timmer & König (1995).

An alternative, but mathematically equivalent method, is as follows:

(i) Generate $K - 1$ random periodogram ordinates $I_k^{\rm rep}$ by multiplying the spectrum $S_k(\boldsymbol{\theta}^t)$ with random χ^2_2 variables as in equation (B1). (At the Nyquist frequency, $k = K$, use a χ^2_1 variable instead of $X_k/2$.)

(ii) Generate $K - 1$ independent, random phases ϕ_k over the range $[-\pi/2, \pi/2)$ from a uniform distribution. (At the zero and Nyquist frequency use $\phi_k = 0$.)

(iii) Produce a complex vector $F_k = A_k \exp(-i\ \phi_k)$ with arguments $A_k = \sqrt{I_k/2}$ and phases ϕ_k.

(iv) Extend the vector (of Fourier amplitudes and phases) to negative frequencies, setting the Fourier components for the negative frequencies $F_{-k} = F_k^*$ where the asterisk denotes complex conjugation. (Note that the Fourier components are real valued at the zero and Nyquist frequencies.)

(v) Inverse Fourier transform the $\{F_k\}$ from the frequency domain to the time domain.

The resulting series will be $2K$ $(= VWN)$ points in length, with a sampling rate of $\Delta T/W$ and a duration of $VN\Delta T$, and will have a mean of approximately zero. (Time series that more closely resemble those of accreting compact objects can be obtained using the exponential transformation of Uttley et al. 2005.) One may then resample a segment of this to match the sampling pattern of the observation, to give a time series of N points, as required (and with no wrap-around effect). For most processes it should make little difference whether the noise due to 'measurement error' is included in the power spectrum, or excluded from the power spectrum and added at the resampling stage (e.g. by drawing the counts per bin from the normal or Poisson distribution after appropriate normalization of the series). See Uttley et al. (2002) for more on time series simulation.

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