A BAYESIAN METHOD FOR THE DETECTION OF UNKNOWN PERIODIC AND NONPERIODIC SIGNALS IN BINNED TIME SERIES

P.C. Gregory
Department of Physics
University of British Columbia
6224 Agricultural Road
Vancouver, British Columbia V6T 1Z1

T.J. Loredo
Department of Astronomy
Space Sciences Building
Cornell University
Ithaca, New York 14853

ABSTRACT. We show that the method we have proposed for the detection and measurement of an unknown periodic signal using arrival times of *individual* events can also be applied to time series consisting of irregularly spaced *binned* samples with gaps. As in our earlier work we again assume a Poisson sampling distribution. The same method can also be used to detect significant *nonperiodic* variation in a signal. We demonstrate this explicitly here with numerical simulations. The approach should be useful to anyone interested in the general problem of inferring the shape of an unknown function from discrete samples.

1. INTRODUCTION

At the MaxEnt 1991 meeting we presented a Bayesian method for the detection and measurement of a periodic signal in an arrival time series when we have no prior knowledge of the existence of such a signal or of its characteristics, including its shape (Gregory and Loredo, 1992a & b, henceforth referred to as GL1 and GL2). Given some data, our goal was to develop Bayesian answers to the questions, "Is there evidence for a periodic signal? If so, what are the best estimates of its period and shape?" Our initial motivation was the detection of pulsars in X-ray astronomy data consisting of a set of arrival times for individual photons, some of which are background events. The lightcurves of X-ray pulsars exhibit a very wide range of shapes, and thus demand use of a very flexible signal model.

To address the detection problem, we used Bayes's theorem to compare a constant rate model for the signal to models with periodic structure. The periodic models describe the signal plus background rate as a stepwise distribution in m bins per period, for various values of m. Such a model is capable of approximating a lightcurve of essentially arbitrary shape. Additionally, many of the calculations required for a Bayesian analysis can be performed analytically with this model, making numerical calculations more efficient than would be possible with, say, Fourier series models.

The Bayesian posterior probability for a periodic model contains a term which quantifies Ockham's razor, penalizing successively more complicated periodic models for their greater complexity even though they are assigned equal prior probabilities. The calculation thus balances model simplicity with goodness-of-fit, allowing us to determine both whether there is evidence for a periodic signal, and the optimum number of bins for describing the structure in the data. Unlike the results of traditional "frequentist" calculations, the outcome of the Bayesian calculation does not depend on the number of periods examined, but only on the range examined. Once a signal is detected, Bayes's theorem can be used to estimate the frequency and shape of the signal.

Our initial presentation of the method assumed that the data consisted of the locations or times of individual discrete events. One of the objectives of this paper, which is discussed in section 2, is to show that the method is also directly applicable to binned data. Consequently the range of applications of the method is much greater than originally conceived. Our second goal, which is presented in section 3, is to show how the same method can be used to reliably detect a nonperiodic signal. The stepwise periodic models referred to above transform to nonperiodic models when the parameter representing the trial period is set equal to the data duration. In this case a signal means any significant variation in the Poisson time series compared to a constant rate model.

2. APPLICATION TO BINNED DATA

All real detectors have finite temporal resolution, reporting the arrival time of an event to within some interval (bin size) Δt, with $\Delta t > 0$. Thus all real discrete time series are binned. But depending on how Δt compares with the timescales of interest, the effect of binning may or may not be significant, calling for different methods of analysis.

In any search for periodicity or time variation using discrete events, there are at least two signal timescales of interest: the typical time between events (*i.e.*, the reciprocal of the average event rate, $\bar{r}$); and the timescale of possible variations of the rate, τ. Clearly, if $\Delta t \gtrsim \tau$, information about interesting temporal variation is being destroyed by the limited detector resolution. Analysis of such data will be severely complicated by the need to "deconvolve." In any case, most experiments are designed to sample on timescales that are small compared with the physical timescales of interest, so we confine our analysis to cases where $\Delta t \ll \tau$.

In our earlier work, we assumed that, in addition, $\Delta t \ll 1/\bar{r}$, so that most time bins are empty, and there is never more than one event per bin. Our analysis held even in the case of infinitesimal Δt, in which case the precise arrival time of each event is known. We refer to time series with $\Delta t \ll 1/\bar{r}$ as *arrival time series.* They are common in high energy astrophysics, where fluxes are small and detectors may have limited effective area, so the expected event rate is small.

But it is also often the case, in astrophysics and elsewhere, that $\Delta t \gtrsim 1/\bar{r}$, so that there may be more than one (perhaps thousands) of events in a bin. As long as $\Delta t \ll \tau$, such *binned time series* can provide useful information about the physical timescales of interest, even though knowledge of the arrival times of individual events is lost. Indeed, if $\tau \gg 1/\bar{r}$, little may be gained by knowing the precise arrival times of individual events, and the expense of recording, storing, and analyzing individual arrival times is thus unjustified.

In this work, we seek a general method for detecting and measuring varying signals of unknown shape using *binned* data, to complement our earlier work on arrival time series.

In fact, our earlier method for analyzing arrival time series can be simply generalized for application to binned time series, as we now show.

Consider data from a counting experiment which is sampled at Z irregular time intervals with sample bin durations, $\{\Delta t_i\}$, for $i = 1$ to Z, centered on the sample times $\{t_i\}$. The number of events counted in each sample bin is represented by $\{n_i\}$. Any number of bins may contain zero counts. The sample bins need not necessarily be contiguous; there may be gaps in the observations. We let L equal the duration of the time series including gaps, and T equal L minus the total duration of all data gaps. A common special case is uniform sampling, where $L = T$, the t_i are evenly spaced, and the sample bin duration is constant.

Note that both the data and the model have bins. To distinguish between them, we always call the bins defining the data the *sample bins*, and those defining the model the *model bins* or *phase bins*. We will use the subscript i to refer to sample bins, and the subscript j to refer to model bins.

A full Bayesian analysis of the problem requires calculation of the following quantities:

1) The odds ratio in favor of the periodic model class over constant alternatives;
2) The probability distribution for the frequency of a hypothetical periodic signal;
3) The probability distribution for the shape of a hypothetical periodic signal, from which a shape estimate and "error bars" can be determined.

To derive equations for these quantities we need to know the likelihood function for each model and the prior probability distributions for the model parameters. We assume the same priors for the model parameters as those used by GL1 & 2. We will now show that for the binned time series, the likelihood functions for the the constant and periodic models are identical in form to those derived previously by GL1 & 2, so that the analysis of that earlier work applies to binned time series as well as to arrival time series.

From the Poisson distribution, the probability for seeing n_i events in a sample bin of duration Δt_i centered at time t_i, given the rate $r(t)$, is,

$$p(n_i \mid r, I) = \frac{[r(t_i)\Delta t_i]^{n_i} e^{-r(t_i)\Delta t_i}}{n_i!}. \tag{1}$$

As mentioned above, we have assumed that the rate does not vary substantially over the duration of a sample bin, so that the integral of the rate over the bin is well approximated by $r(t_i)\Delta t_i$. The likelihood function for r is just the joint probability for all Z samples in the time series, which is simply the product of equation (1) for each sample,

$$p(D \mid r, I) = K \left[\prod_{i=1}^{Z} r(t_i)^{n_i}\right] \exp\left[-\sum_{i=1}^{Z} r(t_i)\Delta t_i\right], \tag{2}$$

where

$$K = \left[\prod_{i=1}^{Z} \Delta t_i{}^{n_i}\right] \Big/ \left[\prod_{i=1}^{Z} n_i!\right]. \tag{3}$$

When the likelihood is used in Bayes's theorem to compute the quantities of interest, the factor K cancels out with the same factor in the global likelihood (the denominator of Bayes's theorem).

For the constant rate model, setting $r(t_i) = A$, where A is a constant, yields,

$$p(D \mid A, M_1) = KA^N e^{-AT}. \tag{4}$$

For the stepwise periodic models the rate is assumed to be a constant in each of the m bins ($m \geq 2$). This is not a single model, but a class of models, one for each choice of m. We denote the information specifying each such model by the symbol M_m. Model M_m has $(m+2)$ parameters: an angular frequency ω (or alternatively a period, P, with $\omega = 2\pi/P$); a phase, ϕ, specifying the location of the bin boundaries; and m values, r_j, specifying the rate in each phase bin, with $j = 1$ to m. The value of the subscript j corresponding to any particular time t is given by,

$$j(t) = \mathrm{int}[1 + m\{(\omega t + \phi) \bmod 2\pi\}/2\pi]. \tag{5}$$

The model assumes that all events in a particular sample bin can be associated with a unique model bin. Because the sample bins have a nonzero width, some fraction of them will straddle model bin boundaries. Provided that the duration of the model bins, P/m, is significantly larger than that of the longest sample bins, this fraction will be very small. This is just the requirement that $\Delta t \ll \tau$, discussed above. This should not be a strong restriction in practice. It merely requires that the duration of sample bins be small compared with the smallest anticipated timescale of variation for the signal. Of course, where possible it is better to use individual event locations because information is irretrievably lost when the event locations are binned.

To facilitate comparison with the constant model, it is convenient to re-express the r_j parameters. We write the rate as the time-averaged rate, A, times a normalized stepwise function that describes the shape of the periodic lightcurve. The average rate is,

$$A = \frac{1}{m}\sum_{j=1}^{m} r_j. \tag{6}$$

The lightcurve shape is completely described by the fraction of the total rate per period that is in each phase bin. These fractions are,

$$\begin{aligned} f_j &= \frac{r_j}{\sum_{k=1}^{m} r_k} \\ &= \frac{r_j}{mA}. \end{aligned} \tag{7}$$

Equations (6) and (7) let us write r_j in terms of A and f_j:

$$r_j = mAf_j. \tag{8}$$

In this way the m values of r_j are replaced by A and the m values of f_j. There are still only m degrees of freedom associated with the rate because by definition the f_j must satisfy the constraint,

$$\sum_{j=1}^{m} f_j = 1, \tag{9}$$

so only $(m-1)$ of them are free. We denote the full set of m values of f_j by the symbol $\vec{f}$.

With this form for the r_j parameters, the likelihood function for model M_m can be calculated using equation (2), giving

$$\begin{aligned} p(D \mid \omega, \phi, A, \vec{f}, M_m) &= K\left[\prod_{i=1}^{Z} mAf_j(t_i)^{n_i}\right]\exp\left[-\sum_{i=1}^{Z} mAf_j(t_i)\Delta t_i\right] \\ &= Km^N A^N\left[\prod_{j=1}^{m} f_j{}^{n_j}\right]\exp\left[-mA\sum_{j=1}^{m} f_j\tau_j(\omega,\phi)\right]. \end{aligned} \tag{10}$$

In this equation $n_j = n_j(\omega,\phi)$ is the total number of events in model bin j (the sum of the n_i which fall in model bin j), and τ_j is the total sample integration time for model bin j (the sum of the durations of all sample bins that fall in bin j).

The $\tau_j(\omega,\phi)$ depend on ω and ϕ (and m), thus the exponential term depends on all of the model parameters. But in many cases this term will be essentially independent of all parameters except A, as we now show.

Unless the "off" times are concentrated in particular bins, we expect the integration time per bin to be approximately T/m. Thus we write the sum in the exponential as follows;

$$\begin{aligned} mA\sum_{j=1}^{m} f_j\tau_j &= AT\sum_{j=1}^{m} f_j\frac{\tau_j}{T/m} \\ &= AT\sum_{j=1}^{m} f_j s_j, \end{aligned} \tag{11}$$

where we have defined the bin time factors $s_j(\omega,\phi)$ by

$$s_j(\omega,\phi) = \frac{\tau_j(\omega,\phi)}{T/m}. \tag{12}$$

Like the τ_j, the s_j are not new parameters; they are numbers that are determined by the data and the model parameters ω, ϕ, and m.

If the observing interval, T, is a contiguous interval containing an integral number of periods, then $\tau_j = T/m$, so $s_j = 1$ for all j. More generally, the observing interval will not be an integral number of periods, and may have gaps, so the s_j will differ from unity. But as long as the number of periods in the observing interval is large, and as long as the gaps are not somehow concentrated in certain bins, the s_j will be very close to unity, and to a good approximation the sum in equation (11) will be equal to $AT\sum_j f_j = AT$. Then the likelihood function is well approximated by,

$$P(D \mid \omega, \phi, A, \vec{f}, M_m) = K(mA)^N e^{-AT}\left[\prod_{j=1}^{m} f_j^{n_j}\right]. \tag{13}$$

We have now derived the likelihood functions for the constant and periodic models (equations (4) & (13)) for a binned time series. Comparing these with the expressions given in GL2 (see equations (4.1) and (4.11)), we find them to be identical except the factor Δt^N has been replaced by the factor K defined in equation (3). Both of these factors cancel out in all subsequent calculations and are thus of no importance. Therefore the final expressions given in GL2 for the odds ratio, the probability for the frequency, and the shape and uncertainty of the light-curve, are all valid for binned time series, provided the bin durations are significantly smaller than the anticipated timescale of variation of the signal.

Note that if the observing duration does not contain a large number of periods, or if there is significant dead time, the s_j may depart significantly from unity. Gregory and Loredo (GL2, Appendix B) considered this case, which leads to the following likelihood function,

$$P(D \mid \omega, \phi, A, \vec{f}, M_m) = \Delta t^N (mA)^N e^{-AT} S(\omega, \phi) \left[\prod_{j=1}^{m} f_j^{n_j} \right], \tag{14}$$

where the binning factor, $S(\omega, \phi)$, is given by

$$S(\omega, \phi) = \prod_{j=1}^{m} s_j^{-n_j}. \tag{15}$$

The s_j are completely determined by the data and the model parameters ω, ϕ and m.

3. NONPERIODIC ANALYSIS

In some situations it is important to determine if a significant *nonperiodic* variation has occurred in the signal producing a time series. The stepwise model, used for the periodic class of models, can just as easily represent a nonperiodic variable lightcurve. Again we assume no prior knowledge about the shape of this variation. Therefore, our calculations can also be used to compare a constant rate model to a variable model, or to compare both constant and variable nonperiodic alternatives to periodic models. All that is necessary is to set the period, $P = 2\pi/\omega$, equal to the total duration of the observations (including any gaps). A discussion of this approach was presented in GL2 (Appendix C). Here we illustrate this capability with two sets of simulated data.

For both simulations the time series consisted of a set of binned arrival times with no data gaps present. The data were simulated for a stepwise lightcurve consisting of 25 bins. In the first case the signal consisted of 26 events occupying bin number 12. To this was added a constant background of 214 events distributed amongst all 25 bins with equal probability. The odds ratio was calculated from,

$$O_{m1} = \frac{1}{2\pi\nu} \binom{N+m-1}{N}^{-1} \int_0^{2\pi} d\phi \; S(\phi) \frac{m^N}{N!} \prod_{j=1}^{m} n_j(\phi)!, \tag{16}$$

as a function of the trial number of bins for the range $m = 2$ to 40. $S(\phi)$ is given by equation (15) and ν is the number of trial choices of m, which in this case equaled 39. This is shown plotted in Figure 1(a). The peak odds ratio, which occurs at $m = 12$, is only 0.18. The odds ratio, $O_{\rm var}$, in favor of the nonperiodic variable class of models over the constant model, is given by,

$$O_{\rm var} = \sum_{m=2}^{m_{\rm max}} O_{m1}. \tag{17}$$

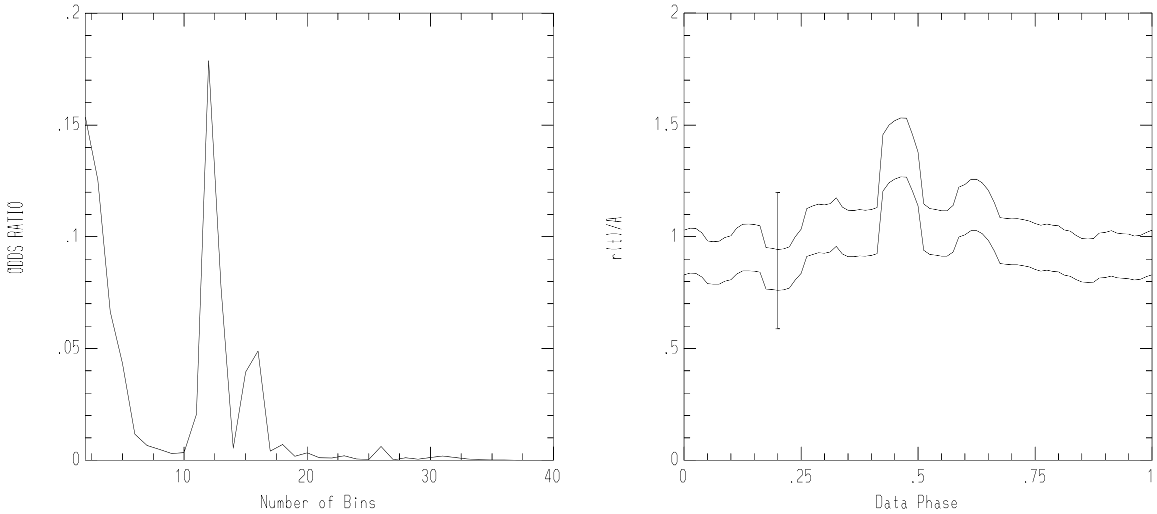

Fig. 1.(a) Odds ratio, O_{m1}, versus m for the first simulation of a stepwise rate function with 25 bins. The ratio of signal event rate to background rate was 3.0. (b) Shape estimate, $r(t)/A$, for the first simulation. Solid curves show ± 1 standard deviation estimates. The error bar shows $\pm\sqrt{N}$ for a typical bin.

Note that $O_{\rm var}$ can exceed unity even if no single variable model is more probable than the constant model. In this particular case $O_{\rm var} = 0.82$, indicating that the constant model is approximately 1.2 times more probable than the variable model. Thus for this level of signal to background no significant signal is indicated in the data set analyzed. Figure 1(b) shows a plot of the Bayesian mean lightcurve marginalized over m according to equations (7.11) and (B.7–9) in GL2. Since $O_{\rm var} < 1$, we marginalized over the constant model, $m = 1$, as well. We calculated the mean and standard deviation for $r(t)/A$ at 80 times, and plotted the two resulting (mean$\pm$standard deviation) curves as solid lines together with a typical $\sqrt{N}$ error bar. Without prior knowledge of the shape and location of the signal it would be very difficult to persuade a seasoned experimenter that a real signal was present, which is consistent with the result from the odds ratio. In this particular simulation the ratio of the signal event rate to background rate was 3.0.

The second simulation was identical to the first, only this time the signal consisted of 30 events and the background 196 events. There is now clear evidence for a signal as the odds ratio, which is plotted in Figure 2(a), reaches a level of 17 for a value of $m = 22$. In this case $O_{\rm var} = 30.8$. This conclusion is amply supported by the estimated lightcurve which is shown in Figure 2(b). The true underlying $r(t)/A$ is shown by the dashed curve. Note that the peak in the odds ratio occurs at $m = 22$, not at the true value of 25 (though the 25 bin model is itself also more probable than a constant model). The data, while sufficiently numerous to indicate a varying rate, are too sparse to fully resolve the signal. In this simulation the ratio of the signal event rate to the background rate was 3.8.

4. SUMMARY

In this paper we have shown that the new Bayesian method for detecting an unknown periodic signal, proposed in GL1 & GL2, is directly applicable to binned time series consisting of samples obtained at irregular intervals in some coordinate, and containing an

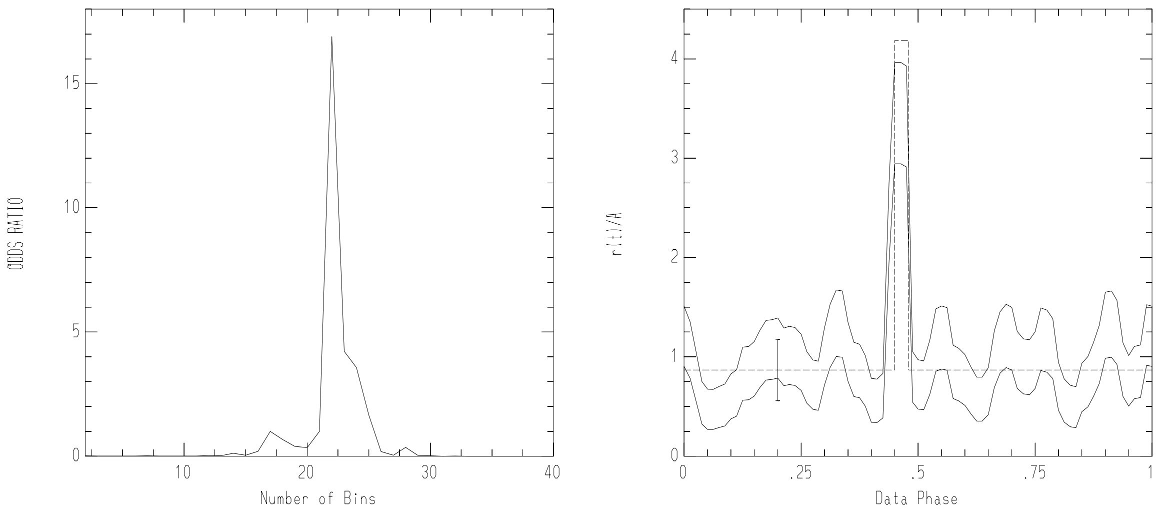

Fig. 2.(a) Odds ratio, O_{m1}, versus m for the second simulation of a stepwise rate function with 25 bins. The ratio of signal event rate to background rate was 3.8. (b) Shape estimate, $r(t)/A$, for the second simulation. Solid curves show ± 1 standard deviation estimates. The error bar shows $\pm\sqrt{N}$ for a typical bin.

arbitrary number of events in each sample bin, as long as the sample bin size is small compared to the anticipated timescale of variation of the signal.

The stepwise model, used for the periodic class of models in our method, can just as easily represent a nonperiodic variable lightcurve for which we have no prior knowledge about its shape. Therefore, our method can also be used to compare a constant rate model to a variable model, or to compare both constant and variable nonperiodic alternatives to periodic models. The method is not restricted to time series, and may prove to be more generally useful for inferring the shape of an unknown function from discrete samples.

ACKNOWLEDGMENTS. This research was supported in part by a grant from the Canadian Natural Sciences and Engineering Research Council at the University of British Columbia, and by a NASA GRO Fellowship (NAG5–1758), NASA grant NAGW–666, and NSF grants AST–87–14475 and AST–89–13112 at Cornell University.

REFERENCES

Gregory, P.C., and Loredo, T. J. 1992a, ‘A Bayesian Method For The Detection Of A Periodic Signal Of Unkown Shape And Period’, in C. R. Smith, G. J. Erickson, & P. O. Neudorfer (ed.), *Maximum Entropy and Bayesian Methods, Seattle, 1991*, Kluwer, Dordrecht, 79.

Gregory, P.C., and T.J. Loredo 1992b, ‘A New Method For The Detection Of A Periodic Signal Of Unknown Shape And Period’, *Ap. J.*, 398, 146.