A NEW METHOD FOR THE DETECTION OF A PERIODIC SIGNAL OF UNKNOWN SHAPE AND PERIOD

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

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

Submitted to *The Astrophysical Journal*
31 January 1992

ABSTRACT

We present a new method for the detection and measurement of a periodic signal in a data set when we have no prior knowledge of the existence of such a signal or of its characteristics. It is applicable to data consisting of the locations or times of discrete events. We use Bayes' theorem to address both the signal detection problem, and the estimation problem of measuring the characteristics of a detected signal. To address the detection problem, we use Bayes' 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. 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, we again use Bayes' theorem to estimate various parameters of the signal, such as its frequency or the shape of the lightcurve. The probability density for the frequency is inversely proportional to the multiplicity of the binned events, which is simply related both to the combinatorial entropy of the binned distribution and to the χ^2 measure of its misfit to a uniform distribution used in the "epoch folding" method for period detection. The probability density for the lightcurve shape produces lightcurve estimates that are superpositions of stepwise distributions with various phases and number of bins, and which are thus smoother than a simple histogram. Error bars for the lightcurve shape are also easily calculated. The method also handles gaps in the data due to intermittent observing or dead time. We apply the method to simulated data generated with both stepwise and sinusoidal lightcurves and demonstrate that it can sensitively detect such signals, and accurately estimate both the signal frequency and its shape, even when the lightcurve does not have a stepwise shape. We also describe a test for nonperiodic source variability that is a simple modification of our period detection procedure.

Subject Headings: analytical methods — numerical methods — pulsars — X-rays: general — gamma-rays: general

1. INTRODUCTION

A problem that frequently arises in astronomy is to determine if there is a periodic signal present in data. In some cases it is known that a periodic signal of known shape and/or period is present but difficult to find because of additive noise. For an excellent Bayesian treatment of this problem in the case of additive Gaussian noise see Bretthorst (1988). In this paper we are concerned with the problem of detecting a periodic signal for which we have no prior knowledge of its period or shape. Usually the experimenter has some reason to suspect the presence of a periodic signal but in other cases the search for a periodicity represents an unexplored avenue with important consequences if successful. One example of the latter is the recently reported regularity in the distribution of galaxy redshifts (Broadhurst *et al.* 1990).

There are two special cases of general interest, depending on whether the appropriate sampling distribution is Gaussian or Poisson. This particular study is motivated by astronomical data in which the objective was the detection of X-ray pulsars. The data are the arrival times of individual X-ray photons, some or all of which are background events, and the appropriate sampling distribution is the Poisson distribution. Our method is directly applicable to data consisting of any event locations (*e.g.* spatial locations or redshifts), not just time series.

Our method detects a signal by using Bayesian probability theory to compare a constant model for the signal to members of a class of models with periodic structure. The periodic models describe the signal plus background with a stepwise function, resembling a histogram, with m phase bins per period (Figure 1). Such a model is capable of approximating a lightcurve of essentially arbitrary shape. If the result of these calculations implies the presence of a periodic signal, we use Bayesian methods to estimate the signal frequency and the shape of the lightcurve.

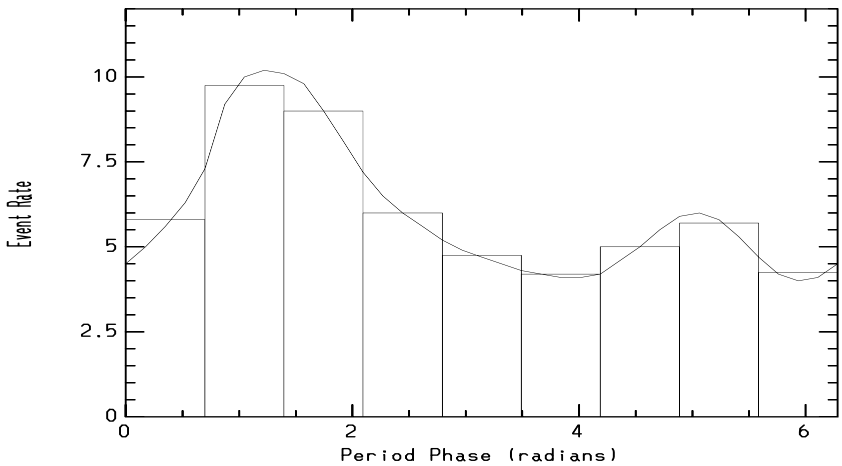

Fig.1 Periodic model, M_m, assumes the periodic signal plus background can be modeled by a stepwise distribution in m bins as illustrated here.

The stepwise model is useful for several reasons. It is a usefully realistic representation of the lightcurves observed in X-ray and gamma-ray astronomy, where time series consisting of the arrival times of events are common. The pulses observed from periodic sources in the X-ray and gamma-ray bands often have very small duty cycles, occupying one or two isolated bins in a histogram of folded event times. 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. Finally, a Bayesian analysis with this model bears a close relationship to the well known "epoch folding" method for detecting periodic signals (Leahy *et al.* 1983), and illuminates some of the strengths and weaknesses of that method.

The calculations required to detect and measure signals with stepwise models are straightforward and could be presented very briefly to an audience familiar with Bayesian probability theory. Many of the results follow from basic Bayesian analysis with the multinomial distribution (such as arises in analyzing tosses of coins or a die). But Bayesian methods are not in wide use in the physical sciences. Accordingly, we describe our calculation in some detail, intending it to be self-contained and pedagogical as well as offering a solution to an important research problem. Our main results are simple and intuitively appealing, and we briefly summarize them here as an outline of this paper before proceeding to the detailed calculations themselves.

We begin with a brief review of Bayesian methods and terminology in § 2. We show how to use Bayes' theorem to estimate the values of parameters in a parametrized model: the Bayesian posterior probability density for a parameter, θ, given some data, D, and a model, M, is proportional to the product of a prior density for the parameter, $p(\theta \mid M)$, and the likelihood function familiar from traditional "frequentist" statistics, $\mathcal{L}(\theta)$:

$$p(\theta \mid D, M) \propto p(\theta \mid M)\ \mathcal{L}(\theta). \tag{1.1}$$

If there are many parameters, Bayes' theorem yields a joint density. We note that the implications of a joint density for a subset of the parameters are simply obtained by "marginalizing": integrating out the unwanted parameters. No satisfactory procedure for accomplishing this is available in frequentist statistics. We then show how to use Bayes' theorem to compare competing parametrized models by calculating the probability of a model as a whole. The probability for model M, given competing models specified by I, is of the form,

$$p(M \mid D, I) \propto p(M \mid I)\ \mathcal{L}_{\mathrm{global}}, \tag{1.2}$$

the product of a prior probability and the global likelihood for the model. The global likelihood is calculated from the familiar likelihood function and the prior density for the model parameters by marginalizing. We show that the resulting global likelihood can be written in the form,

$$\mathcal{L}_{\mathrm{global}} = \mathcal{L}_{\mathrm{max}}\ \Omega_\theta; \tag{1.3}$$

the product of the maximum likelihood of the model and a factor called the Ockham factor associated with the model parameters. We show how the Ockham factor "corrects" the maximum likelihood, used in many frequentist model assessment methods, in a manner which quantifies Ockham's Razor, automatically and objectively penalizing complicated models.

Likelihood functions play a key role in Bayesian methods. Accordingly, we derive a general likelihood function for arrival time data in § 3. This likelihood function has been used in frequentist analyses of astrophysical data by others (*e.g.*, Cleveland 1983). We note how its use in Bayesian calculations allows simple treatment of gaps in data, such as might arise from dead time or intermittent observing.

In § 4 we specialize this likelihood function to the constant and stepwise models used in this work, and we assign the priors needed to use Bayes' theorem. We show that the likelihood function for the stepwise model leads one to count data falling in various phase bins in the same manner as is done in the commonly used "epoch folding" method.

In § 5 we perform the calculations needed to compare periodic and constant models using Bayes' theorem. The key function of the data arising in the analysis is the multiplicity of the binned event times,

$$W = \frac{N!}{n_1!\, n_2!\, \ldots\, n_m!}, \tag{1.4}$$

where $n_j(\omega, \phi)$ is the number of events falling into the j'th of m phase bins, given the frequency, ω, and phase, ϕ, and N is the total number of events. The multiplicity is the number of ways the binned distribution could have arisen "by chance," giving our results intuitively appealing interpretations. We show that the probability of a periodic model when the period and phase of the signal are known beforehand is inversely proportional to the multiplicity of the binned distribution it leads to, and that this probability has Ockham factors penalizing it in a way that increases with the number of bins, m. When the frequency and phase are unknown, the probability of a periodic model is proportional to the integral of the inverse multiplicity over frequency and phase. It thus depends on the range of frequency searched, but not on the number or location of the frequencies at which the multiplicity statistic is calculated. The frequency range enters the Ockham factor penalizing the model for its unknown frequency.

In § 6 we describe how to use Bayes' theorem to estimate the frequency of a signal if the model comparison calculations show a signal is present. We eliminate all model parameters except the frequency by marginalizing. Again the multiplicity plays a key role: those frequencies which lead to a binned distribution with small multiplicity (small probability of arising "by chance") are assigned a high probability.

In § 7 we perform similar calculations to estimate the shape of the lightcurve of a detected signal. Here marginalization of all parameters except those describing the shape leads to shape estimates that are weighted superpositions of stepwise models with varying frequencies, phases, amplitudes, and numbers of bins, leading to an estimate that can be substantially smoother than the basic stepwise functions comprising our model space. We also show how to calculate error bars for the estimated lightcurve.

We illustrate our method in § 9 by applying it to simulated data. We first apply it to data simulated from a piecewise constant rate, and compare with the results of an epoch folding (EF) method analysis. Our method clearly detects the signal and correctly identifies the number of bins needed to model the signal. In contrast the EF method leads to a 2.5 σ detection which is not usually considered significant. Then we apply it to data simulated from a sinusoidal signal, demonstrating the ability of the method to correctly detect and estimate smooth signal shapes.

In § 8 we elaborate on the relationship between our method and other better-known statistics. We show that the logarithm of the multiplicity is well-approximated by the combinatorial entropy of the binned distribution (a measure of its complexity), so that our method chooses the frequencies near those leading to binned distributions with *minimum* entropy. A further approximation shows that when no periodic signal is present, the logarithm of the multiplicity is approximately proportional to the χ^2 statistic used in the epoch folding (EF) method. However, the Bayesian calculation uses the statistic in a very different manner. To test for the presence of a periodic signal, the EF method calculates χ^2 at a number of frequencies, assigns a significance to the maximum χ^2 value found, and adjusts this significance to account for the number of frequencies searched. Bayesian methods instead *average* the multiplicity over frequency, so that only the range of frequencies examined, and not the number, enters the calculation. To account for the unknown phase, Collura *et al.* (1987) have advocated averaging χ^2 over phase for each examined frequency. Bayesian methods instead average the multiplicity, approximately proportional to the *exponential* of $\chi^2/2$.

A concluding section summarizes our results, and a "recipe" for analyzing arrival time data with our method is given in Appendix A. Appendix B outlines a modified version of the method to deal with significant data gaps. Additionally, Appendix C describes a simple modification of our method that leads to a test for nonperiodic source variability, somewhat similar in spirit to the test proposed by Collura *et al.* (1987).

We close the introduction by emphasizing that this work only scratches the surface of what Bayesian methods have to offer astrophysicists for analyzing complicated data. In regard to analyzing event locations or arrival times, other useful models could be considered, either simpler or more complicated than the class examined here (some simple models with direct relationships to frequentist tests with the Rayleigh and Z_n^2 statistics are discussed in Loredo 1992b). More generally, we hope our work demonstrates some of the practical and conceptual advantages of Bayesian methods over more traditional methods in a manner that encourages their application to other, very different problems.

2. BAYESIAN INFERENCE

We adopt the Bayesian approach to statistical inference in this work. In this section we briefly describe the methods and terminology employed in the remainder of this paper. We will not enter the debate over the conceptual superiority of Bayesian or more standard "frequentist" methods here. Interested readers can find further details and references regarding conceptual and methodological aspects of Bayesian inference elsewhere (Loredo 1990; 1992a,b).

In Bayesian inference, the viability of each member of a set of rival hypotheses, $\{H_i\}$, is assessed in the light of some observed data, D, by calculating the probability of each hypothesis, given the data and any background information, I, we may have regarding the hypotheses and data. Following a notation introduced by Jeffreys (1961), we write such a probability as $p(H_i \mid D, I)$, explicitly denoting the background information by the proposition, I, to the right of the bar. At the very least, the background information must specify the class of alternative hypotheses being considered, and the relationship between the hypotheses and the data (the statistical model).

The basic rules for manipulating Bayesian probabilities are the sum rule,

$$p(H_i \mid I) + p(\overline{H}_i \mid I) = 1, \tag{2.1}$$

and the product rule,

$$\begin{aligned} p(H_i, D \mid I) &= p(H_i \mid I) p(D \mid H_i, I) \\ &= p(D \mid I) p(H_i \mid D, I). \end{aligned} \tag{2.2}$$

The various symbols appearing as arguments should be understood as propositions; for example, D might be the proposition, "N photons were counted in a time T." The symbol $\overline{H}_i$ signifies the negation of H_i (a proposition that is true if H_i is false, so that one of the alternatives to H_i is true), and (H_i, D) signifies the logical conjunction of H_i and D (a proposition that is true only if H_i *and* D are both true). The rules hold for any propositions, not just those indicated above. In particular, we will have occasion to use equation (2.1) with I replaced by the proposition (D, I).

Throughout this work, we will be concerned with exclusive hypotheses, so that if one particular hypothesis is true, all others are false. For such hypotheses, the sum and product rules imply the generalized sum rule,

$$p(H_i + H_j \mid I) = p(H_i \mid I) + p(H_j \mid I), \tag{2.3}$$

and the normalization rule,

$$\sum_i p(H_i \mid I) = 1. \tag{2.4}$$

Here a "+" within a probability symbol stands for logical disjunction, so that $H_i + H_j$ is a proposition that is true if either H_i *or* H_j is true.

One of the most important calculating rules in Bayesian inference is Bayes' theorem, found by equating the two right hand sides of equation (2.2) and solving for $p(H_i \mid D, I)$:

$$p(H_i \mid D, I) = p(H_i \mid I)\, \frac{p(D \mid H_i, I)}{p(D \mid I)}. \tag{2.5}$$

Bayes' theorem describes a type of learning: how the probability for each member of a class of hypotheses should be modified on obtaining new information, D. The probabilities $p(H_i \mid I)$ for the hypotheses in the absence of D are called their prior probabilities, and the probabilities $p(H_i \mid D, I)$ including the information D are called their posterior probabilities. The quantity $p(D \mid H_i, I)$ is called the sampling probability for D, or the likelihood for H_i, and the quantity $p(D \mid I)$ is called the prior predictive probability for D, or the global likelihood for the entire class of hypotheses.

All of the rules we have written down so far show how to manipulate known probabilities to find the values of other probabilities. But to be useful in applications, we additionally need rules that assign numerical values or functions to the initial *direct probabilities* that will be manipulated. For example, to use Bayes' theorem, we need to know the values of the three probabilities on the right side of equation (2.5). These three probabilities are not independent; the sum and product rules imply that,

$$p(D \mid I) = \sum_i p(H_i \mid I)\, p(D \mid H_i, I). \tag{2.6}$$

That is, the denominator of Bayes' theorem, which does not depend on H_i, must be equal to the sum of the numerator over H_i. It thus plays the role of a normalization constant.

Usually, we will treat priors and likelihoods as direct probabilities: we assign priors either from invariance and consistency requirements (which in problems more complicated than those we treat here can lead to use of the maximum entropy principle), or simply from intuition, and we use likelihoods derived from the Poisson distribution. Then we calculate the value of the global likelihood with equation (2.6), and use all three quantities to evaluate Bayes' theorem. But other approaches are possible. We may, for example, assign the likelihood and the global likelihood directly, and solve equation (2.6) for the prior.

2.1. *Parameter Estimation*

In astrophysics we frequently consider sets of hypotheses ("parameterized models") that are labeled, not by discrete numbers, but by the possible values of a continuous parameter, θ. In such cases the quantities of interest become probability *densities*. For example, given some background information, M, specifying a parametrized model with one parameter, θ, $p(\theta \mid M)$ is the prior density for θ, which means that $p(\theta \mid M) d\theta$ is the prior probability that the true value of the parameter is in the interval $[\theta, \theta + d\theta]$. We use the same symbol, $p(\ldots)$, for densities and probabilities; the nature of the argument will identify which use is intended.

Bayes' theorem, and all the other rules just discussed, hold for densities, with all sums replaced by integrals. For example, the global likelihood for model M can be calculated with the continuous counterpart of equation (2.6),

$$p(D \mid M) = \int d\theta\, p(\theta \mid M)\, p(D \mid \theta, M). \tag{2.7}$$

If there is more than one parameter, multiple integrals are used. If the prior density and the likelihood are assigned directly, the global likelihood is just a normalization constant, and the posterior density for the parameters is simply proportional to the product of the prior and the likelihood.

The use of Bayes' theorem to determine what one can learn about the values of parameters from data is called *parameter estimation*, though strictly speaking, Bayesian inference does not provide estimates for parameters. Rather, the Bayesian solution to the parameter estimation problem is the full posterior *distribution*, $p(\theta \mid D, M)$, and not just a single point in parameter space. Of course, it is often useful to summarize this distribution for textual, graphical, or tabular display in terms of a "best-fit" value and "error bars." Possible summarizing best-fit values are the posterior mode (most probable value of θ) or the posterior mean,

$$\langle\theta\rangle = \int d\theta\, \theta\, p(\theta \mid D, M). \tag{2.8}$$

If the mode and mean are very different, the posterior distribution is too asymmetric to be adequately summarized by a single estimate. An allowed range for a parameter with probability content C is provided by a *credible region*, or highest posterior density (HPD) region, R, defined by

$$\int_R d\theta\, p(\theta \mid D, M) = C, \tag{2.9}$$

with the posterior density inside R everywhere greater than that outside it. We sometimes speak picturesquely of the region of parameter space that is assigned a large density as the "posterior bubble." More sophisticated treatment of best-fit parameter values and ranges is possible using decision theory, but these simple, intuitive summaries will be adequate here.

As an example, consider estimating the mean, μ, of a Gaussian distribution with known standard deviation, σ, from N samples with a sample mean of $\bar{x}$. The likelihood function for such data is proportional to a Gaussian distribution for μ with mean $\bar{x}$ and standard deviation $\sigma/\sqrt{N}$. If the available prior information leads to a prior density for μ that is constant, or at least virtually constant near the peak of the likelihood, the posterior distribution will be just this same Gaussian. This distribution is the full Bayesian solution to the problem of estimating μ. But we will often simply summarize it by its posterior mean or mode, $\bar{x}$, and the 68.3% ("one sigma") credible region of $\bar{x} \pm \sigma/\sqrt{N}$. These are the usual estimates from frequentist statistics. But the equality of Bayesian and frequentist estimates will not generally hold for likelihoods that are not Gaussian. Examples of this are provided in Loredo (1992a) and in the remainder of this work.

2.2. *Nuisance Parameters*

Frequently a parametrized model will have more than one parameter, but we will want to focus attention on a subset of the parameters. For example, in this work we will want to focus on the implications of the data for the frequency of a periodic signal, independent of the signal's amplitude, shape, or phase. The uninteresting parameters are known as *nuisance parameters*. As always, the full Bayesian inference is the full joint posterior distribution for all of the parameters; but its implications for the parameters of interest can be simply summarized by integrating out the nuisance parameters. Explicitly, if model M has two parameters, θ and ϕ, and we are interested only in θ, then it is a simple consequence of the sum and product rules that,

$$p(\theta \mid D, M) = \int d\phi\, p(\theta, \phi \mid D, M). \tag{2.10}$$

For historical reasons, the procedure of integrating out nuisance parameters is called *marginalization*, and $p(\theta \mid D, M)$ is called the marginal posterior distribution for θ. Equation (2.7) for the global likelihood is a special case of marginalization in which *all* of the model parameters are marginalized out of the joint distribution, $p(D, \theta \mid M) = p(\theta \mid M)p(D \mid \theta, M)$.

The use of marginalization to eliminate nuisance parameters is one of the most important technical advantages of Bayesian inference over standard frequentist statistics. Indeed, the name "nuisance parameters" originated in frequentist statistics because there is no general frequentist method for dealing with such parameters; they are indeed a "nuisance" in frequentist statistics. Marginalization plays a very important role in this work. Its broader implications for analyzing astrophysical data will be discussed elsewhere.

2.3. *Model Comparison*

Often more than one parametrized model will be available to explain a phenomenon, and we will wish to compare them. The models may differ in form or in number of parameters. Use of Bayes' theorem to compare competing models by calculating the probability of each model as a whole is called *model comparison*. Bayesian model comparison has a built-in "Ockham's Razor": Bayes' theorem automatically penalizes complicated models, assigning them large probabilities only if the complexity of the data justifies the additional complication of the model.

Model comparison calculations require the explicit specification of two or more specific alternative models, M_i. We take as our background information the proposition that one of the models under consideration is true. Symbolically, we might write this as $I = M_1 + M_2 + \ldots$, where the "+" symbol here stands for disjunction ("or"). Given this information, we can calculate the probability for each model with Bayes' theorem:

$$p(M_i \mid D, I) = p(M_i \mid I)\, \frac{p(D \mid M_i, I)}{p(D \mid I)}. \tag{2.11}$$

The proposition (M_i, I) is true if and only if model M_i is true, that is, it is equivalent to the proposition M_i itself. Thus $p(D \mid M_i, I) = p(D \mid M_i)$, which we recognize as the global likelihood for model M_i, which we can calculate according to equation (2.7). The term in the denominator is again a normalization constant, obtained by summing the products of the priors and the global likelihoods of all models being considered. Model comparison is thus completely analogous to parameter estimation: just as the posterior distribution for a parameter is proportional to its prior times its likelihood, so the posterior probability for a model as a whole is proportional to its prior probability times its global likelihood.

It is often useful to consider the ratios of the probabilities of two models, rather than the probabilities directly. The ratio, $O_{ij} = p(M_i \mid D, I)/p(M_j \mid D, I)$, is called the *odds ratio* in favor of model M_i over model M_j. From equation (2.11),

$$\begin{aligned} O_{ij} &= \frac{p(M_i \mid I)}{p(M_j \mid I)}\, \frac{p(D \mid M_i)}{p(D \mid M_j)} \\ &\equiv \frac{p(M_i \mid I)}{p(M_j \mid I)}\, B_{ij}, \end{aligned} \tag{2.12}$$

where the first factor is the prior odds ratio, and the second factor is called the *Bayes factor*. The Bayes factor is the ratio of the global likelihoods of the models: Bayesian model comparison relies on the ratio of global likelihoods, not maximum likelihoods. Note that the normalization constant in equation (2.11) drops out of the odds ratio; this can make the odds ratio somewhat easier to work with. The odds ratio is also conceptually useful when one particular model is of special interest. For example, in this work we will compare a constant rate model with a class of periodic alternatives, and will thus calculate the odds in favor of each alternative over the constant model.

If we have calculated the odds ratios, O_{i1}, in favor of each model over model M_1, we can find the probabilities for each model by inverting equation (2.12), giving

$$p(M_i \mid D, I) = \frac{O_{i1}}{\sum_{j=1}^{N_{\rm mod}} O_{j1}}, \tag{2.13}$$

where $N_{\rm mod}$ is the total number of models considered, and of course $O_{11} = 1$.

A crucial consequence of the marginalization procedure used to calculate global likelihoods is that the Bayes factor automatically favors simpler models unless the data justify the complexity of more complicated alternatives. Thus even if we assign equal prior probabilities to competing models, simpler models can have larger posterior probabilities than their competitors. This can be understood as follows.

Imagine comparing two models: M_1 with a single parameter, θ, and M_0 with θ fixed at some default value θ_0 (so M_0 has no free parameters). To calculate the Bayes factor B_{10} in favor of model M_1, we will need to perform the integral in equation (2.7) for model M_1. We can find a simple and illuminating approximation to this integral as follows. First, we note that in most cases the likelihood function, $\mathcal{L}(\theta) = p(D \mid \theta, M_1)$, is much more peaked than the prior, so that the prior is essentially constant over the range of significant likelihood. If $\hat{\theta}$ is the maximum likelihood value of of θ, we can thus approximate equation (2.7) as,

$$p(D \mid M_1) \approx p(\hat{\theta} \mid M_1) \int d\theta\, \mathcal{L}(\theta). \tag{2.14}$$

Next, we write the integral of the likelihood function as the product of its maximum amplitude and a characteristic width, $\delta\theta$. An approximate characteristic width might be the full width at half maximum of the peak, or twice the standard deviation of the peak. Alternatively, if the integral can be calculated exactly, the width can be defined as the integral divided by the peak value. Writing the integral in this manner gives,

$$p(D \mid M_1) \approx p(\hat{\theta} \mid M_1)\; \mathcal{L}(\hat{\theta})\; \delta\theta. \tag{2.15}$$

Finally, we write $p(\hat{\theta} \mid M_1)$ as the reciprocal of a characteristic width of the prior, $\Delta\theta$. Thus the global likelihood for M_1 is approximately,

$$p(D \mid M_1) \approx \mathcal{L}(\hat{\theta}) \frac{\delta\theta}{\Delta\theta}. \tag{2.16}$$

Since model M_0 has no free parameters, no integral need be calculated to find its global likelihood, which is simply the likelihood for model M_1 evaluated at the default value,

$$p(D \mid M_0) = \mathcal{L}(\theta_0). \tag{2.17}$$

Thus the Bayes factor in favor of the more complicated model is,

$$B_{10} \approx \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\theta_0)}\, \frac{\delta\theta}{\Delta\theta}. \tag{2.18}$$

The likelihood ratio in the first factor can never be less than unity; that is, it can never favor the simpler model because M_1 contains it as a special case. But since the posterior width, $\delta\theta$, is narrower than the prior width, $\Delta\theta$, the second factor penalizes the complicated model for any "wasted" parameter space that gets ruled out by the data. The Bayes factor will thus favor the more complicated model only if the likelihood ratio is large enough to overcome this penalty.

Equation (2.15) for the global likelihood has the form of the best-fit likelihood times the factor that penalizes M_1. In general, one can always write the global likelihood of a model with one or more parameters (denoted collectively by θ) as the maximum value of its likelihood times some factor, Ω_θ:

$$p(D \mid M) \equiv \mathcal{L}_{\max}\, \Omega_\theta. \tag{2.19}$$

This implicitly defines Ω_θ as equation (2.7) divided by the maximum likelihood. The quantity Ω_θ is called the *Ockham factor* associated with the parameters, θ. It is so named because it corrects the maximum likelihood usually considered in statistical tests in a manner that quantifies the qualitative notion behind "Ockham's Razor": simpler explanations are to be preferred unless there is sufficient evidence in favor of more complicated explanations. Bayes' theorem both quantifies such evidence and determines how much additional evidence is "sufficient" through the calculation of global likelihoods. We will see examples of this in § 5 below.

For model M_1 in our example above, $\Omega_\theta \approx \delta\theta\; p(\hat{\theta} \mid M_1) \approx \delta\theta/\Delta\theta$. If the likelihood function is Gaussian, it is simple to show that the characteristic width of the posterior that leads to the correct value of the integral in equation (2.14) is $\delta\theta = \sigma_\theta\sqrt{2\pi}$, where σ_θ is the standard deviation of the posterior distribution for θ. Thus for likelihood functions that are nearly Gaussian near their peak,

$$\Omega_\theta \approx \sqrt{2\pi}\;\sigma_\theta\; p(\hat{\theta} \mid M). \tag{2.20}$$

We will see Ockham factors of this form arising at several places in this work.

As we noted above, Bayesian model comparison requires the specification of at least two competing models. There is no Bayesian counterpart to frequentist "goodness-of-fit" tests that seek to reject a single "null" hypothesis without specifying an alternative. However, it is often the case that Bayesian model comparison calculations with specific alternatives require evaluation of the same statistic as frequentist tests that are allegedly alternative-free. The present work will provide an example of this. Indeed, rigorous frequentist studies characterize statistics, not only by their ability to reject a default model (often characterized by the "false alarm" probability), but also by their ability to correctly identify one of two alternative models (characterized by the "power" of the test). From the Bayesian point of view, there is no such thing as an alternative-free test, for any choice of test statistic implies a direction of departure from the null hypothesis that implicitly refers to a specific class of alternative models.

3. LIKELIHOOD FUNCTION FOR ARRIVAL TIME DATA

As shown above, the likelihood function plays a key role in Bayesian inference, encoding all the information the data provide about the hypotheses of interest. In this section we derive a general likelihood function for arrival time data with an arbitrary time-dependent rate, $r(t)$. We then specialize to the models of interest in this work in the following section.

The data are the arrival times for each of N events, $D = \{t_i\}$, $i = 1$ to N, over some observing interval of duration T. The probability for D given some specified rate function—the likelihood function—can be calculated as follows.

We divide the observation period into small time intervals, Δt, each containing either no event or one event. From the Poisson distribution, the probability of seeing n events in an interval Δt about time t is,

$$p_n(t) = \frac{[r(t)\Delta t]^n e^{-r(t)\Delta t}}{n!}. \tag{3.1}$$

We have assumed that the rate does not vary substantially within Δt, so the average rate in the interval is approximately equal to the rate at any time within the interval. If N and Q are the number of time intervals in which one event and no event are detected, respectively, then the likelihood function is given by

$$p(D \mid r, I) = \prod_{i=1}^{N} p_1(t_i) \prod_{k=1}^{Q} p_0(t_k). \tag{3.2}$$

From equation (3.1), with $n = 0$ and 1,

$$p_0(t) = e^{-r(t)\Delta t}, \tag{3.3}$$

and

$$p_1(t) = r(t)\Delta t\, e^{-r(t)\Delta t}, \tag{3.4}$$

so the likelihood function takes the form,

$$\begin{aligned} p(D \mid r, I) &= \Delta t^N \left[\prod_{i=1}^{N} r(t_i)\right] \exp\left[-\sum_{k=1}^{N+Q} r(t_k)\Delta t\right] \\ &= \Delta t^N \left[\prod_{i=1}^{N} r(t_i)\right] \exp\left[-\int_T dt\, r(t)\right]. \end{aligned} \tag{3.5}$$

In the last equation we have identified the sum of $r(t)\Delta t$ over all the observed intervals with the integral of the rate over the intervals, with the range of integration $T = (N + Q)\Delta t$. This identification holds even if Δt is long enough that $r(t)$ varies substantially within the intervals, in which case $r(t)$ must be replaced with its average value over the interval in the above equations.

Equation (3.5) is the general likelihood function we use throughout this work. The intervals, Δt, could correspond to the precision of the clock recording the arrival times. But when used in Bayes' theorem, the Δt^N factor in the likelihood will cancel with the same factor in the global likelihood, so inferences will not depend on the size of Δt, and are well-behaved even in the limit in which the time intervals become infinitesimal.

Note that in equation (3.5) the time T is the total duration of the intervals in which it is known that either no event or one event was detected. These intervals need not be contiguous; there can be gaps during which there are no data. These "off" intervals are excluded from the integration range, T. This is the case for intervals that are off due to a known "dead time" associated with the detected events, or intervals that are "off" for any other reason (telemetry delays, intermittent observing, *etc.*). Formally, such intervals should be included in equation (3.2) by adding a factor of the form,

$$\prod_{k=1}^{N_{\rm off}} p_{\rm any}(t_k), \tag{3.6}$$

where k runs over the $N_{\rm off}$ intervals during which there is no information, and $p_{\rm any}(t_k)$ is the probability of seeing any number of events (0, 1, 2, ...) in an interval of size Δt at time t_k. This probability is,

$$p_{\rm any}(t_k) = \sum_{n=0}^{\infty} p_n(t_k), \tag{3.7}$$

where $p_n(t_k)$ is the Poisson probability given in equation (3.1). Since the Poisson distribution is normalized, we find immediately that $p_{\rm any}(t_k) = 1$. Thus the likelihood function given by equation (3.5) is unchanged, so long as we exclude all "off" intervals from our analysis, and particularly from the integration over T. In this way our analysis accounts for known gaps in the data, regardless of the cause of the gaps. (If the gaps are of unknown duration, as is sometimes the case for detectors with complicated dead time behavior, a more complicated analysis is necessary.)

We close this section by pointing out some distinctions between analyzing Poisson distributed event arrival times and analyzing sampled time series with Gaussian noise distributions. First, as is apparent from equation (3.5), the likelihood function for arrival times requires calculation of the *product* of the rates at the times of each event. In contrast, the likelihood function for sampled time series is the product of Gaussians for each sample, which requires calculation of the *sum* of the rates at the times of each sample, because the rates appear in exponents. If, for example, one uses sinusoidal signal models, sums of sinusoids resembling a Fourier series appear in the analysis of sampled time series, but products of sinusoids appear in the analysis of arrival time series. A second distinction is that sampled time series are often uniformly sampled, allowing efficient calculation of the necessary sums with fast Fourier transforms if sinusoidal models are used. In contrast, arrival times are never on a uniform grid. Because of these distinctions, the calculations required for analyzing the two types of time series are very different, and models appropriate for sampled time series may not be appropriate for arrival time series.

4. CONSTANT AND PERIODIC MODELS

In this section we specialize equation (3.5) to the models considered in this paper. In addition, we assign prior probabilities for the parameters of the models, and for the models themselves, to prepare us to use Bayes' theorem to perform model comparison and parameter estimation in later sections.

4.1. *Constant Model*

The simplest model for the data has a constant event rate, A. We denote the information specifying this one-parameter model by the symbol M_1. Setting $r(t_i) = A$ in equation (3.5), the likelihood function for this model is,

$$p(D \mid A, M_1) = \Delta t^N A^N e^{-AT}. \tag{4.1}$$

We will assume that the information M_1 leads to a prior density for A that is constant from $A = 0$ to some upper limit, $A_{\rm max}$. Thus the normalized prior for A is,

$$p(A \mid M_1) = \frac{1}{A_{\rm max}}. \tag{4.2}$$

We will find that our results depend only very weakly (some results not at all) on the form of this prior or on the value of $A_{\rm max}$, and will be well-behaved even in the limit where $A_{\rm max} \to \infty$.

4.2. *Periodic Stepwise Likelihood*

As noted in the introduction, our model for a periodic signal is a stepwise function with a constant rate in each of m bins per period ($m \geq 2$). 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,

$$j(t) = \mathrm{int}[1 + m\{(\omega t + \phi) \bmod 2\pi\}/2\pi]. \tag{4.3}$$

We sometimes denote the full set of m values of r_j by the symbol $\vec{r}$.

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,

$$A = \frac{1}{m}\sum_{j=1}^{m} r_j. \tag{4.4}$$

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

$$\begin{aligned} f_j &= \frac{r_j}{\sum_{k=1}^{m} r_k} \\ &= \frac{r_j}{mA}. \end{aligned} \tag{4.5}$$

Equations (4.4) and (4.5) let us write r_j in terms of A and f_j:

$$r_j = mAf_j. \tag{4.6}$$

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,

$$\sum_{j=1}^{m} f_j = 1, \tag{4.7}$$

so only $(m-1)$ of them are free. We sometimes 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 (3.5), giving

$$\begin{aligned} P(D \mid \omega, \phi, A, \vec{f}, M_m) &= \Delta t^N \left[\prod_{i=1}^{N} mAf_{j(t_i)} \right] \exp\left[-\sum_{k=1}^{N+Q} mAf_{j(t_k)}\Delta t \right] \\ &= \Delta t^N (mA)^N \left[\prod_{j=1}^{m} f_j^{n_j} \right] \exp\left[-mA\sum_{j=1}^{m} f_j \tau_j(\omega,\phi) \right], \end{aligned} \tag{4.8}$$

where $j(t_k)$ is given by equation (4.3), $n_j = n_j(\omega,\phi)$ is the number of events that fall into phase bin j, and $\tau_j(\omega,\phi)$ is the total integration time for bin j. We have implicitly assumed that the intervals Δt are small compared to the width, P/m, of the r_j bins, so that each event has an unambiguous bin assignment for given ω and ϕ.

The integration times $\tau_j(\omega,\phi)$ 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;

$$\begin{aligned} mA\sum_{j=1}^{m} f_j\tau_j &= AT\sum_{j=1}^{m} f_j \frac{\tau_j}{T/m} \\ &= AT\sum_{j=1}^{m} f_j s_j, \end{aligned} \tag{4.9}$$

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

$$s_j(\omega,\phi) = \frac{\tau_j(\omega,\phi)}{T/m}. \tag{4.10}$$

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 (4.9) will be equal to $AT\sum_j f_j = AT$. Then the likelihood function is well approximated by,

$$P(D \mid \omega, \phi, A, \vec{f}, M_m) = \Delta t^N (mA)^N e^{-AT} \left[\prod_{j=1}^{m} f_j^{n_j} \right]. \tag{4.11}$$

We use this likelihood function in the remainder of this work. 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, complicating the analysis. In Appendix B we discuss a simple modification to the method to deal with this problem in an approximate manner.

Note that when $m = 1$, so that the single value of the shape parameter $f_1 = 1$, equation (4.11) takes the same form as equation (4.1). Thus model M_1 is the $m = 1$ case of model M_m, as suggested by our notation.

4.3. *Priors for Periodic Model Parameters*

We will assume that we do not have any prior information linking the frequency, phase, and lightcurve shape, so that the priors for ω, ϕ, and $\vec{r}$ are all independent of one another. We will further assume that there is no prior information linking the shape of the lightcurve to its average value. When reparametrized in terms of A and $\vec{f}$, the joint prior will thus be of the form,

$$p(\omega, \phi, A, \vec{f} \mid M_m) = p(\omega \mid M_m)\, p(\phi \mid M_m)\, p(A \mid M_m)\, p(\vec{f} \mid M_m). \tag{4.12}$$

We may know the frequency or both the frequency and the phase *a priori* (as when searching for X-ray pulsations from a radio pulsar at the radio pulsation frequency), in which case the priors for the frequency and phase are δ-functions (more simply, we could just always keep them to the right of the bar in probability symbols). We discuss this case in the next section. Here we will assign priors assuming that little is known about the model parameters *a priori*, aside from their physical significance.

The prior density for ϕ we take to be uniform over the interval $[0, 2\pi]$,

$$p(\phi \mid M_m) = \frac{1}{2\pi}. \tag{4.13}$$

Formally, this prior can be derived from an invariance argument requiring investigators with different origins of time to reach the same conclusions. A similar invariance argument, this time requiring observers with different units of time to reach the same conclusions (Bretthorst 1988), leads to a prior density for ω of the form,

$$p(\omega \mid M_m) = \frac{1}{\omega \ln \frac{\omega_{\rm hi}}{\omega_{\rm lo}}}, \tag{4.14}$$

where $[\omega_{\rm lo}, \omega_{\rm hi}]$ is a prior range for ω and the $\ln \frac{\omega_{\rm hi}}{\omega_{\rm lo}}$ factor is a normalization constant ensuring that the integral of $p(\omega \mid M_m)$ over the prior range is equal to 1. This density is uniform in the logarithm of ω. Since our likelihood function requires there to be at least several periods in the data set, we usually set $\omega_{\rm lo} = 20\pi/T$, so that there are at least 10 periods in the data set for every frequency considered. We typically set $\omega_{\rm hi} = 2\pi N/T$, corresponding to twice a "Nyquist limit" (since the events do not come at uniform intervals, there is often useful information at considerably higher frequencies than the "Nyquist frequency" of $\pi N/T$). Changing the value of $\omega_{\rm hi}$ by a factor of a few does not greatly affect our results, because $\omega_{\rm hi}$ enters calculations only through the logarithmic factor in equation (4.14).

We noted that this nonuniform prior arises from an invariance property: invariance of conclusions with respect to scale changes in time. In addition, this prior is form-invariant with respect to reparameterization in terms of the period, P. That is, an investigator working in terms of ω, and using a $1/\omega$ prior, will reach the same conclusions as another investigator working in terms of P, and using a $1/P$ prior. This would not be true for a prior of another form. To the extent that parameterizations in terms of ω and P are both "natural," this form invariance is desirable.

The average rate, A, is not the same quantity as the constant rate that appears in model M_1, so technically we should use a different symbol for it. If we knew that the rate was constant, so that $f_j = 1/m$ for all j, we expect our inferences about A under model M_m to be identical to those under model M_1. Combined with our assumption of independent priors for the shape and average rate, this requires that we use the same prior for the average rate under M_m and the constant rate under M_1. Accordingly, our prior is,

$$p(A \mid M_m) = \frac{1}{A_{\max}}. \tag{4.15}$$

Since the A parameters for all of the models enter their respective likelihood functions in the same way, and have the same priors, no confusion is caused by using the same symbol for them.

Finally, we must assign a prior joint density for $\vec{f}$. We simply assume that all values between 0 and 1 are equally probable subject to the constraint that $\sum f_j = 1$. Thus we write,

$$p(\vec{f} \mid I) = K_m \delta \left(1 - \sum_{j=1}^{m} f_j \right), \tag{4.16}$$

where $\delta(x)$ denotes the Dirac Delta-function, and K_m is a normalization constant that depends on the value of m.

We can evaluate K_m from the requirement that $\int d\vec{f}\, p(\vec{f} \mid I) = 1$. The required integral is a special case of the generalized Beta integral that we will require later:

$$\int_0^\infty dx_1 \ldots \int_0^\infty dx_m x_1^{k_1 - 1} \ldots x_m^{k_m - 1}\, \delta \left(a - \sum_{j=1}^{m} x_j \right) = \frac{\Gamma(k_1) \ldots \Gamma(k_m)}{\Gamma(k)}\, a^k, \tag{4.17}$$

where $k = \sum_j k_j$, and $\Gamma(x)$ is the Gamma function, with $\Gamma(n) = (n-1)!$ when n is a positive integer. To evaluate K_m, we consider the case with $a = 1$ and all $k_j = 1$, so $k = m$. Thus $\int d\vec{f}\, p(\vec{f} \mid I) = K_m/(m-1)!$, so

$$K_m = (m-1)!. \tag{4.18}$$

Our specification of model M_m is now complete.

4.4. *Priors for Model Comparison*

Finally, we need to assign a prior probability to each model as a whole in order to perform model comparison calculations. We will consider the hypotheses of the presence and absence of a periodic signal to be equally probable *a priori*. Thus we assign equal prior probabilities of 1/2 to the class of nonperiodic models and to the class of periodic models. Since the nonperiodic class contains only the constant ($m = 1$) model, we have

$$p(M_1 \mid I) = \frac{1}{2}, \tag{4.19}$$

where we use the symbol I to denote the information specifying the classes of models we are comparing, as in § 2. The periodic class will consist of some finite number of stepwise models with m varying from $m = 2$ to $m = m_{\max}$. We consider each member of this class to be equally probable *a priori*, so that the probability of 1/2 assigned to the periodic class is spread equally among the $\nu = m_{\rm max} - 1$ members of this class. Thus,

$$p(M_m \mid I) = \frac{1}{2\nu}. \tag{4.20}$$

Alternatively, we could view our stepwise models as a single model, and m as a discrete parameter in this model; the $1/\nu$ factor in equation (4.20) then plays the role of a flat prior distribution for the parameter m.

We now have all the probabilities we need. We note that either model class could be expanded to more comprehensively cover the possible forms of periodic or nonperiodic signals. For example, we could enlarge the periodic class to contain sinusoidal or other simple functional models (Loredo 1992b). Or we could enlarge the nonperiodic class to include varying but nonperiodic models, such as the stepwise model discussed in Appendix C, or a simple polynomial variation with time. In either case, we would further spread the prior probability of 1/2 over the additional members of the model class. Alternatively, if we knew the precise shape of a possible signal *a priori*, we could shrink the periodic model class, assigning its full prior probability of 1/2 to a single model.

5. ODDS RATIOS FOR SIGNAL DETECTION

In this section we use the likelihoods and priors from the previous section to test the hypothesis that the signal is periodic. We do this by comparing members of the class of stepwise periodic models to the constant model that comprises the nonperiodic class using the model comparison methods of § 2. We calculate the global likelihoods for all models, and use them to find the odds ratios, O_{m1}, in favor of each periodic model over the constant model. The probability for a model can be calculated from the odds ratios using equation (2.13). In particular, the probability for the nonperiodic (constant) model is,

$$p(M_1 \mid D, I) = \frac{1}{1 + \sum_{m=2}^{m_{\rm max}} O_{m1}}, \tag{5.1}$$

and the probability that the signal is periodic is just the sum of the probabilities of the ν periodic models,

$$p(m > 1 \mid D, I) = \frac{\sum_{m=2}^{m_{\rm max}} O_{m1}}{1 + \sum_{m=2}^{m_{\rm max}} O_{m1}}, \tag{5.2}$$

This is simply $[1 - p(M_1 \mid D, I)]$. The ratio of equation (5.2) to equation (5.1) is the odds ratio, $O_{\rm per}$, in favor of the hypothesis that the signal is periodic,

$$O_{\rm per} = \sum_{m=2}^{m_{\rm max}} O_{m1}. \tag{5.3}$$

When $O_{\rm per}$ is greater than unity ($p(m > 1 \mid D, I) > 1/2$), there is evidence for a periodic signal, the magnitude of $O_{\rm per}$ indicating the strength of this evidence.

We discuss three cases: the case when the period and phase of the signal is known, that when only the period is known, and that when neither the period nor the phase is known. In all cases the shape of the lightcurve will be considered unknown.

5.1. *Period and Phase Known*

Our first case is of little practical interest: if we do not know the shape of the lightcurve, it is not meaningful to say we know the phase. However, if the phase is fixed, the needed calculations can be done analytically, and the result can be readily interpreted in terms of the Ockham factors described in § 2, facilitating our understanding of the results for more complicated cases.

As explained in § 2, we need the global likelihoods of the models to perform model comparison. These are calculated by integrating the product of the prior and the likelihood for each model as illustrated in equation (2.7). From equations (4.1) and (4.2), the global likelihood for the constant model is,

$$\begin{aligned} p(D \mid M_1) &= \int_0^{A_{\max}} dA\; p(A \mid M_1)\; p(D \mid M_1, A, I) \\ &= \frac{\Delta t^N}{A_{\max}} \int_0^{A_{\max}} dA\; A^N e^{-AT} \\ &= \frac{\Delta t^N \gamma(N+1, A_{\max}T)}{A_{\max} T^{N+1}}. \end{aligned} \tag{5.4}$$

Here $\gamma(n,x)$ denotes the incomplete Gamma function defined by,

$$\gamma(n,x) = \int_0^x y^{n-1} e^{-y} dy, \tag{5.5}$$

where the usual Gamma function $\Gamma(n) = \gamma(n, \infty)$.

With ω and ϕ known, the global likelihood for a periodic model is similarly calculated by integrating the product of equations (4.11), (4.15), and (4.16) over A and $\vec{f}$. The integral over A is the same as that just performed for the constant model, and the constrained integral over $\vec{f}$ can be performed using the generalized Beta integral of equation (4.17). The result is,

$$\begin{aligned} p(D \mid \omega, \phi, M_m) &= \int d\vec{f} \int_0^{A_{\max}} dA\; p(A \mid M_m)\; p(\vec{f} \mid M_m)\; p(D \mid \omega, \phi, A, \vec{f}, M_m) \\ &= \frac{\Delta t^N m^N (m-1)!}{A_{\max}} \int_0^{A_{\max}} dA\; A^N e^{-AT} \int d\vec{f} \prod_{j=1}^m f_j^{n_j} \delta\left(1 - \sum_{j=1}^m f_j\right) \\ &= \frac{\Delta t^N (m-1)!\; N!\; \gamma(N+1, A_{\max}T)}{A_{\max} T^{N+1} (N+m-1)!} \; \frac{m^N}{W_m(\omega,\phi)}, \end{aligned} \tag{5.6}$$

where $W_m(\omega, \phi)$ is the multiplicity of the binned distribution of events, the number of ways the particular set of n_j values can be made by distributing N events in m bins,

$$W_m(\omega,\phi) = \frac{N!}{n_1!\; n_2!\; \ldots\; n_m!}. \tag{5.7}$$

The multiplicity is a function of ω and ϕ because the n_j are. Note that m^N is the total number of possible arrangements of N events in m bins, so equation (5.6) is inversely proportional to the ratio of the number of ways the observed n_j can be made to the total number of ways N events can be placed in m bins.

The probability of a model is proportional to its global likelihood. Thus from equation (5.6) we find the intuitively appealing result that the probability of a periodic model is inversely proportional to the multiplicity of its resulting binned distribution. Crudely, if the number of ways the binned distribution could have arisen "by chance" is large, the probability that the distribution is due to a genuinely periodic signal is small.

The many factors common to the global likelihoods in equations (5.4) and (5.6) cancel when we calculate the odds ratio comparing one of the periodic models to the constant model. Using the prior probabilities for the models given by equations (4.19) and (4.20), the odds ratio in favor of an m-bin periodic model over the constant model, conditional on ω and ϕ, is,

$$O_{m1}(\omega,\phi) = \frac{1}{\nu}\binom{N+m-1}{N}^{-1}\frac{m^N}{W_m(\omega,\phi)}. \tag{5.8}$$

Note that $A_{\rm max}$ has canceled out of the odds ratio, so that the result of comparing the models is independent of the prior range for A. This is generally the case in model comparison calculations when a parameter is common to all models being considered, and its value is independent of the values of the other parameters: only the ranges associated with the additional parameters affect the outcome.

Values of $O_{m1}(\omega,\phi) > 1$ indicate that model M_m is more probable than the constant rate model for the frequency and phase considered. From the conditional odds ratios we can easily calculate the odds ratio, $O_{\rm per}(\omega,\phi)$, in favor of the periodic class of models over the constant model,

$$O_{\rm per}(\omega,\phi) = \sum_{m=2}^{m_{\rm max}} O_{m1}(\omega,\phi). \tag{5.9}$$

This is the conditional counterpart to equation (5.3). Note that $O_{\rm per}(\omega,\phi)$ could exceed unity even if no single periodic model is more probable than the constant model. This can arise in practice when there is a weak periodic signal present whose shape is not well-modeled by a single stepwise curve, so that probability is spread over several stepwise models.

5.2. *Ockham Factors*

The odds ratios $O_{m1}(\omega,\phi)$ contain Ockham factors that penalize models with larger numbers of bins. As noted in § 2, we can see this by writing the global likelihood for each model as the product of its maximum likelihood and a remaining factor, the Ockham factor.

The constant model has an Ockham factor associated with its single parameter, the rate, A. To find it, we differentiate equation (4.1) with respect to A, and choose A so the derivative vanishes, leading to a maximum likelihood value for A of N/T, as we might have guessed. Thus the maximum value of the likelihood itself is,

$$\mathcal{L}_{\rm max,1} = \Delta t^N N^N T^{-N} e^{-N}. \tag{5.10}$$

From this result and equation (5.4), the global likelihood for M_1 can be written,

$$\begin{aligned} p(D \mid M_1) &= \mathcal{L}_{\rm max,1}\,\frac{e^N N^{-N}\gamma(N+1, A_{\rm max}T)}{A_{\rm max}T} \\ &= \mathcal{L}_{\rm max,1}\,\Omega_A, \end{aligned} \tag{5.11}$$

where we have identified the Ockham factor associated with the parameter A,

$$\Omega_A = \frac{e^N N^{-N}\gamma(N+1, A_{\rm max}T)}{A_{\rm max}T}. \tag{5.12}$$

When $A_{\rm max}T \gg N$ (*i.e.*, the prior upper limit for A is larger than the observed rate), as will usually be the case, the incomplete Gamma function is equal to $N!$ to a very good approximation. Using Stirling's approximation, $N! \approx \sqrt{2\pi N}\, N^N e^{-N}$, we find that,

$$\Omega_A \approx \sqrt{2\pi}\, \frac{N^{1/2}/T}{A_{\rm max}}. \tag{5.13}$$

It is easy to show that the standard deviation of the posterior distribution for A is $N^{1/2}/T$ (the familiar "root-N" result; see Loredo 1992a); thus the Ockham factor associated with A is approximately $\sqrt{2\pi}$ times the posterior standard deviation divided by the prior range, the Gaussian limit described in § 2, equation (2.20).

The Ockham factors associated with the parameters of a periodic model can be found in a similar manner. As with the constant model, the maximum likelihood value of A is N/T. We show below that maximization of equation (4.11) with respect to the f_j leads to maximum likelihood values of $f_j = n_j/N$, as one might expect. Thus the maximum value of the likelihood for a known ω and ϕ is,

$$\mathcal{L}_{{\rm max},m}(\omega,\phi) = \left[\Delta t^N N^N T^{-N} e^{-N}\right] \left[m^N N^{-N} \prod_{j=1}^{m} n_j^{n_j}\right]. \tag{5.14}$$

We have grouped terms to facilitate comparison with the calculations for the constant model. We can now write the global likelihood in equation (5.6) as,

$$\begin{aligned} p(D \mid \omega,\phi,M_m) = \mathcal{L}_{{\rm max},m}(\omega,\phi) &\left[\frac{e^N N^{-N}\gamma(N+1, A_{\rm max}T)}{A_{\rm max}T}\right] \times \\ &\binom{N+m-1}{N}^{-1} \frac{N^N \prod_j n_j!}{N! \prod_j n_j^{n_j}}, \end{aligned} \tag{5.15}$$

where we have identified a combination of factorials that are equal to a binomial coefficient,

$$\binom{N+m-1}{N} = \frac{(N+m-1)!}{N!\,(m-1)!}. \tag{5.16}$$

Comparing equation (5.15) with equation (5.12), we see that we can write the global likelihood for M_m as,

$$p(D \mid \omega,\phi,M_m) = \mathcal{L}_{{\rm max},m}(\omega,\phi)\, \Omega_A\, \Omega_m, \tag{5.17}$$

where Ω_A is given by equation (5.12), and the Ockham factor associated with the m values of f_j is,

$$\Omega_m = \binom{N+m-1}{N}^{-1} \frac{N^N \prod_j n_j!}{N! \prod_j n_j^{n_j}}. \tag{5.18}$$

Using Stirling's approximation for the various factorials, one can show that

$$\Omega_m \approx \sqrt{\frac{N}{2\pi}} \left(1 + \frac{m-1}{N}\right)^{-(m-\frac{1}{2})} \left[(m-1)! \prod_{j=1}^{m} \frac{\sqrt{2\pi n_j}}{N}\right]. \tag{5.19}$$

We show below that the posterior standard deviation for f_j is approximately $\sqrt{n_j}/N$, thus the term in brackets is the product of the prior and $\sqrt{2\pi}$ times the posterior standard deviations, as described in § 2, equation (2.20). The remaining factors arise because the Gaussian approximation is not adequate for the bounded, constrained f_j parameters.

Using equations (5.11) and (5.17), we can write the odds ratio in equation (5.8) as the product of a prior odds ratio $(1/\nu)$, a maximum likelihood ratio, and and Ockham factor:

$$O_{m1}(\omega, \phi) = \frac{1}{\nu} R_{m1}\, \Omega_m, \tag{5.20}$$

where the ratio of the maximum values of the likelihoods of the models is,

$$\begin{aligned} R_{m1} &\equiv \frac{\mathcal{L}_{\max,m}(\omega, \phi)}{\mathcal{L}_{\max,1}} \\ &= \left(\frac{m}{N}\right)^N \prod_{j=1}^{m} n_j^{n_j}. \end{aligned} \tag{5.21}$$

Note that Ω_A has cancelled out of the odds ratio. Our remark above regarding the cancellation of $A_{\max}$ in the odds ratio applies to Ockham factors as well: when a parameter is common to all models being compared, and its inferred value is not correlated with the values of the other parameters, only the Ockham factors associated with the additional parameters affect the outcome of model comparison calculations.

To illustrate the effect of the Ockham factor, Ω_m, assume that we only had one model, with fixed m, in the periodic model class (so $\nu = 1$). Then the prior odds for this model over the constant model would be unity. However, the Ockham factor in equation (5.21) implements an "Ockham's razor," strongly favoring the simpler constant rate model, M_1, unless the data justify the model M_m with its larger number of free parameters. The effect of the Ockham factor can be appreciated most readily for the special case where the total number of events, N, is an integer multiple of the number of bins, m, and the events fall uniformly in the bins, so that $n_j = N/m$, an integer. In this case the likelihood ratio $R_{m1} = 1$, as we expect, so the odds ratio is equal to the Ockham factor, which takes the value,

$$\Omega_m = \binom{N+m-1}{N}^{-1} \frac{m^N}{N!} \left(\frac{N}{m}!\right)^m. \tag{5.22}$$

Figure 2*a* shows a plot of $\log_{10} \Omega_m$ versus the total number of events, N, while the number of bins is held constant at $m = 6$. For $N \approx 500$ events, $\Omega_m \approx 10^{-5}$. Figure 2*b* shows $\log_{10} \Omega_m$ versus m for N fixed at 420 events. In this case for 5 bins ($m = 5$), $\Omega_5 = 9.3 \times 10^{-5}$, and for 10 bins, $\Omega_{10} = 2.0 \times 10^{-8}$. In all cases, the Ockham factor strongly penalizes the complicated models, and the odds ratio favors the constant model, even though the likelihood ratio itself does not favor one model over another for this hypothetical flat data set, and even though the models are assigned equal prior probabilities.

These results may be surprising to those familiar with more traditional frequentist approaches to model comparison based on best-fit likelihood ratios or their logarithms, (e.g., differences in χ^2, such as are used in the F-test). Likelihood ratios can never favor the simpler of two nested models; at best, the ratio can be unity. In such tests "Ockham's Razor" must be invoked to justify the selection of the simpler model when the likelihood ratio does not favor the more complicated one too strongly, the simplicity of the simpler model supposedly making it more plausible *a priori*. In the Bayesian analysis a quantitative *a posteriori* Ockham factor arises as a derivable consequence of the basic sum and product rules of probability theory. Thus model probabilities and odds ratios can favor simpler models even when likelihood ratios do not.

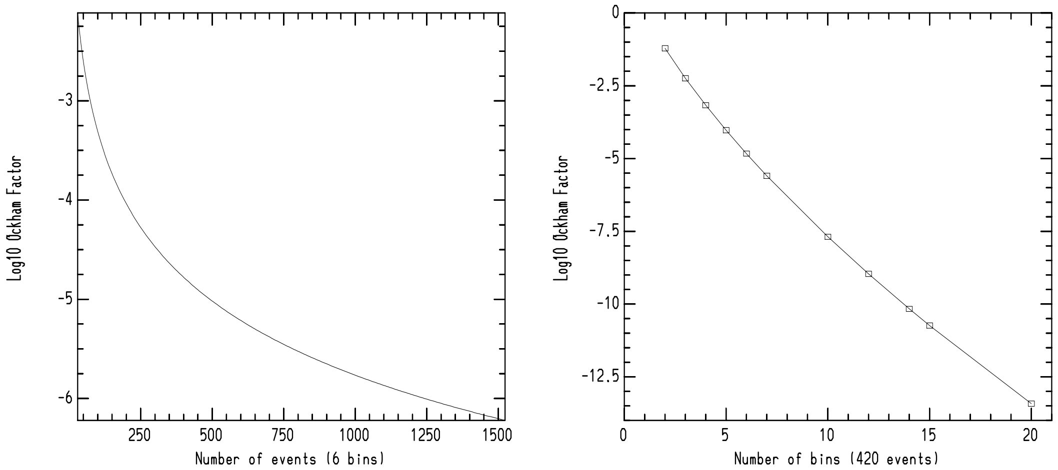

Fig.2 Logarithm of the Ockham factor, Ω_m, for N uniformly distributed events. (a) Ω_m versus N for $m = 6$ bins. (b) Ω_m versus m, for $N = 420$ events. Only points with N/m an integer are plotted. Models with smaller Ockham factors are penalized over models with larger Ockham factors. Thus, the penalty increases with increasing number of bins.

5.3. *Period Known, Phase Unknown*

We now treat the more practically useful case of known period or frequency, but unknown phase and shape. This is the state of knowledge one might be in when, say, searching for X-ray pulsations from a radio pulsar at the known frequency of the radio pulsations. The global likelihood for a model with m bins and known frequency is given by,

$$p(D \mid \omega, M_m) = \int_0^{2\pi} d\phi \int d\vec{f} \int_0^{A_{\max}} dA \; p(\phi \mid M_m)\, p(A \mid M_m)\, p(\vec{f} \mid M_m) \times p(D \mid \omega, \phi, A, \vec{f}, M_m). \tag{5.23}$$

This is simply equation (5.6), multiplied by the prior for ϕ, and integrated over ϕ. Thus,

$$p(D \mid \omega, M_m) = \frac{\Delta t^N (m-1)!\, N!\, \gamma(N+1, A_{\max}T)}{2\pi A_{\max} T^{N+1} (N+m-1)!} \int_0^{2\pi} d\phi \, \frac{m^N}{W_m(\omega, \phi)}. \tag{5.24}$$

The integral must be evaluated numerically. Numerical integrals over ϕ are most efficiently done by integrating over one bin width ($\phi = 0$ to $2\pi/m$), and multiplying the result by m, since phase shifts larger than one bin width simply correspond to cyclicly permuting the f_j, which does not change the value of the multiplicity.

From equations (5.4) and (5.24), the odds ratio in favor of an m-bin model with known frequency is,

$$O_{m1}(\omega) = \frac{1}{2\pi\nu} \binom{N+m-1}{N}^{-1} \int_0^{2\pi} d\phi \, \frac{m^N}{W_m(\omega, \phi)}, \tag{5.25}$$

that is, it is the odds ratio for the case with known period and phase, equation (5.8), averaged over phase. This is smaller than $O_{m1}(\omega, \phi)$, maximized with respect to ϕ, by an additional Ockham factor that penalizes the model for its unknown phase (the $1/2\pi$ factor is just the part of Ω_ϕ arising from the prior for ϕ).

As in equation (5.9), the odds ratios given by equation (5.25) can be used to calculate the odds ratio in favor of the periodic class, $O_{\rm per}(\omega)$, according to,

$$O_{\rm per}(\omega) = \sum_{m=2}^{m_{\rm max}} O_{m1}(\omega). \tag{5.26}$$

This odds ratio, or equivalently, the probability of the periodic class given by equation (5.2), is the quantity one would use to detect a signal at a frequency which is known *a priori* (say, from measurements at another wavelength), when the shape of the lightcurve is unknown.

5.4. *Period and Phase Unknown*

Finally, if the frequency or period is also unknown, we can find the global likelihood for an m-bin model by multiplying equation (5.24) by the prior for ω and integrating over ω;

$$p(D \mid M_m) = \frac{\Delta t^N (m-1)!\; N!\; \gamma(N+1, A_{\rm max}T)}{2\pi A_{\rm max}(N+m-1)!\; T^{N+1} \ln\frac{\omega_{\rm hi}}{\omega_{\rm lo}}} \int_{\omega_{\rm lo}}^{\omega_{\rm hi}} \frac{d\omega}{\omega} \int_0^{2\pi} d\phi\, \frac{m^N}{W_m(\omega,\phi)}. \tag{5.27}$$

Using this global likelihood, the odds ratio in favor of M_m with unknown period and phase is,

$$O_{m1} = \frac{1}{2\pi\nu \ln\frac{\omega_{\rm hi}}{\omega_{\rm lo}}} \binom{N+m-1}{N}^{-1} \int_{\omega_{\rm lo}}^{\omega_{\rm hi}} \frac{d\omega}{\omega} \int_0^{2\pi} d\phi\, \frac{m^N}{W_m(\omega,\phi)}. \tag{5.28}$$

As before, this expression must be evaluated numerically. If there is significant evidence for a periodic signal, the integral is usually completely dominated by a single peak in the inverse multiplicity, even for moderately small numbers of events. The integration introduces an additional Ockham factor penalizing the model for its unknown frequency (the $1/\ln\frac{\omega_{\rm hi}}{\omega_{\rm lo}}$ factor is the part of Ω_ω arising from the normalization constant for the prior for ω). This Ockham factor, which arises from marginalizing ω, is the Bayesian counterpart to the adjustment of the significance of a frequentist period search for the number of independent periods searched. We elaborate on the distinction between these methods below.

The odds ratios given by equation (5.28) can be used to calculate the odds ratio in favor of the hypothesis that a periodic signal of unknown frequency and unknown shape is present by calculating $O_{\rm per}$ using equation (5.3).

If the results of the signal detection calculations just described lead us to conclude that a periodic signal is present, then the problem becomes one of estimating the frequency and shape of the lightcurve. These problems are considered in the next two sections.

6. ESTIMATION OF THE FREQUENCY

Our signal detection calculations have focused on the global likelihoods for various models. For parameter estimation, we will focus instead on the posterior distribution for model parameters, in which the global likelihood merely plays the role of a normalization constant.

Assuming the truth of a particular model, M_m, Bayes' theorem for the posterior distribution for the frequency reads,

$$p(\omega \mid D, M_m) = p(\omega \mid M_m)\, \frac{p(D \mid \omega, M_m)}{p(D \mid M_m)}. \tag{6.1}$$

The probabilities on the right hand side are given by equations (4.14), (5.24), and (5.27). The result is,

$$p(\omega \mid D, M_m) = \frac{C}{\omega} \int d\phi \, \frac{1}{W_m(\omega, \phi)}. \tag{6.2}$$

where C is a normalization constant,

$$C = \left[\int_{\omega_{\rm lo}}^{\omega_{\rm hi}} \frac{d\omega}{\omega} \int_0^{2\pi} d\phi \, \frac{1}{W_m(\omega, \phi)} \right]^{-1}. \tag{6.3}$$

If we wish to estimate the phase and frequency jointly, a similar calculation, using equations (4.13), (4.14), (5.6), and (5.27), gives,

$$p(\omega, \phi \mid D, M_m) = \frac{C}{\omega} \frac{1}{W_m(\omega, \phi)}. \tag{6.4}$$

These marginal distributions show that the multiplicity, $W_m(\omega, \phi)$, of the binned arrival times contains all the information provided by the data about the frequency and phase (it is a "sufficient statistic" for the frequency and phase) when we are not interested in the exact shape of the lightcurve. These distributions straightforwardly summarize the information provided by the data about the frequency and phase. In particular, a "best" frequency can be found by locating the maximum of equation (6.2) (giving the mode) or by calculating the posterior mean for ω (as illustrated in equation (2.8)). The probability that the true frequency is in any specified range can be found simply by integrating equation (6.2) over that range.

Note that the normalization constants for the priors, $A_{\rm max}$, 2π, and $\ln(\omega_{\rm hi}/\omega_{\rm lo})$, do not appear, making the posterior very insensitive to the prior ranges for the parameters. This is a general characteristic of Bayesian calculations: the results of parameter estimation tend to be very insensitive to the prior range of parameters, being essentially equal to limiting results obtained with infinite range. This is in contrast to model comparison calculations, which can depend more sensitively on the range through the Ockham factor (*e.g.*, the odds ratio O_{m1} is inversely proportional to $\ln(\omega_{\rm hi}/\omega_{\rm lo})$).

These calculations are all conditional on the choice of a particular model, M_m. But the Bayesian model comparison calculations of the previous section do not isolate a single model; rather, they assign a probability to each possible model (just as Bayesian parameter estimation does not produce a single point in parameter space, but rather a posterior "bubble"). Formally, our estimate of the frequency should include the information provided by all of the models, essentially marginalizing over m, which we can consider a discrete nuisance parameter. We can perform this calculation as follows.

As we did in equation (5.2), let $(m > 1)$ stand for the proposition, "the signal is periodic, not constant." Then a complete description of our knowledge of the frequency of the periodic signal is given by the marginal distribution,

$$\begin{aligned} p(\omega \mid m > 1, D, I) &= \sum_{m=2}^{m_{\rm max}} p(M_m, \omega \mid m > 1, D, I) \\ &= \sum_{m=2}^{m_{\rm max}} p(M_m \mid D, I) \, p(\omega \mid D, M_m), \end{aligned} \tag{6.5}$$

where $p(M_m \mid D, I)$ can be calculated from the odds ratios in equation (5.28) using equation (2.13), and $p(\omega \mid D, M_m)$ is given by equation (6.2). Equation (6.5) is a weighted sum of the marginal posteriors for all the periodic models being considered, from those with $m = 2$ to $m = m_{\max}$.

Though equation (6.5) is the full Bayesian estimate for the frequency, we find in practice that it is often adequate simply to compute equation (6.2), conditional on the most probable choice of m (that with the largest O_{m1}).

Though in principle the parameter estimation problem of determining the frequency should be treated after the model comparison problem of determining whether a periodic signal is present, in practice one will begin by calculating the inverse multiplicity averaged over phase in equation (6.2) throughout a range of frequencies, and for each of several values of m. The locations of important peaks will be noted, and additional values should be calculated in their vicinity to enable accurate calculation of the integrals required for calculating the normalization constant, C, and for calculation of the odds ratios needed for signal detection. We discuss this calculation procedure further in Appendix A.

7. ESTIMATION OF THE LIGHTCURVE SHAPE

It may be of interest to estimate the shape of the lightcurve. The lightcurve is entirely specified by the stepwise function $r(t) = r_{j(t)}$. Since $r_{j(t)} = mAf_{j(t)}$, the uncertainty in the rate at some time will depend on the uncertainties of both A and the f_j value for the specified time. However, the uncertainty in the *shape* of the lightcurve does not depend on how well its overall amplitude has been determined, so we will seek to estimate $r(t)/A = mf(t)$ (with $f(t) \equiv f_{j(t)}$), instead of $r(t)$.

The estimate of $f(t)$ at one particular time is not simply the estimate of a particular f_j parameter, because the value of j corresponding to t depends on the frequency and phase, which are usually not known. Additionally, the estimate will depend on which model (number of bins) is chosen. Formally, since the shape of the signal is determined by the choice of the model and by the values of *all* of the model parameters except A, the full Bayesian shape estimate is the joint posterior distribution, $p(M_m, \omega, \phi, \vec{f} \mid D, I)$, with A eliminated by marginalizing. We will call this distribution the *shape distribution*. The shape distribution is so complicated as to be almost useless by itself. In this section we seek to summarize its implications in an intuitively accessible manner.

The first summary we will seek is the most probable shape. In principle this shape should be found by jointly maximizing the shape distribution with respect to all of its parameters. In practice, this is a formidable task, so instead we factor the shape distribution according to,

$$p(M_m, \omega, \phi, \vec{f} \mid D, I) = p(M_m \mid D, I)\, p(\omega \mid D, M_m)\, p(\phi \mid \omega, D, M_m)\, p(\vec{f} \mid \omega, \phi, D, M_m). \tag{7.1}$$

Then we maximize each factor in sequence. First we identify the most probable number of bins from the odds ratio calculations of § 5. Then we find the most probable frequency for this model from the marginal distribution for ω given by equation (6.2). Then we find the most probable phase given these choices for the number of bins and the frequency. Finally, we find the f_j values that are most probable, given these choices for the other parameters. The resulting shape should be nearly the most probable shape. Not surprisingly, we show below that it is the shape that assigns $f_j = n_j(\omega, \phi)/N$, with m, ω, and ϕ chosen as just specified.

Unfortunately, though the most probable shape is easily found from the results of previous detection and estimation calculations, it tells us little about the *extent* of the shape distribution, the variety of shapes that are consistent with the data. So we seek an alternative summary that accessibly portrays the "width" of the shape distribution. The particular summary we will offer can be motivated as follows. Imagine creating a movie by performing a Monte Carlo calculation that samples the shape distribution. For each frame of the movie, we sample a choice of m from $p(M_m \mid D, I)$, then sample a frequency and phase from $p(\omega \mid D, M_m)$ and $p(\phi \mid \omega, D, M_m)$, then sample f_j values from $p(\vec{f} \mid \omega, \phi, D, M_m)$, and finally draw the curve specified by $(m, \omega, \phi, \vec{f})$ in black on the frame. Viewing this movie, we will not see any particular stepwise curve, but will instead see a blurred greyscale "image" that is darkest near points that many shape samples went through, and lightest near points that lie only on shapes with small probabilities, and which are thus seldom sampled. This blurred image will more completely portray the extent of the shape distribution than a single highly probable curve.

It is not feasible to make such a movie, but we show how straightforward calculations can be used to effectively draw "contours" of the image one might see in such a movie. We do this by calculating, for each time, t, the mean and standard deviation of $r(t)/A$, averaging over the entire shape distribution. Nearly all of the analytical and numerical integrals required are already available from detection and estimation calculations, so this shape estimate can be readily calculated. It is a weighted superposition of stepwise curves with various numbers of bins and various frequencies, phases, and f_j values, and can be considerably smoother than a stepwise curve. It is somewhat similar in spirit to the recently proposed kernel density estimates for lightcurves (de Jager, Swanepoel and Raubenheimer 1986).

In the remainder of this section, we first show how to find the most probable values of the f_j parameters, assuming that optimum values of m, ω, and ϕ have been chosen as outlined above. Then we show how to find the mean and standard deviation estimate of $r(t)/A$ at a specified time in three steps. First, we find the estimate for specified m, ω, and ϕ. Then we average over frequency and phase. Finally, we average over m, superposing models with various numbers of bins.

7.1. *Most Probable Stepwise Model*

Before actually finding the most probable stepwise model, we elaborate on when it may provide an adequate shape estimate. We saw earlier that the Ockham factor strongly penalizes more complicated models with larger numbers of bins (Fig. 2(b)). Therefore we expect that if the odds ratio, O_{m1}, is computed for two different choices of m, its value for the smaller m will be greater unless the data provide sufficient justification for a larger number of bins. Thus peaks in a plot of O_{m1} versus m indicate evidence for structure on timescales of order P/m.

Since the evidence provided by the data depends on the number of events counted, it is to be expected that as N increases, peaks can occur at larger and larger values of m if the data arise from a sufficiently structured signal. Thus peaks in O_{m1} versus m must be interpreted as indicating the degree of complexity of the lightcurve supported by the data.

If O_{m1} exhibits a single strong peak, this is evidence that the lightcurve is well represented by a single stepwise model. In such cases, the most probable stepwise model may adequately represent the information in the data. In other cases, O_{m1} may be large for several values of m, indicating that no particular stepwise curve exhausts the structure in the data. It is only in the former case that the most probable shape we are about to derive should be considered to be a useful shape estimate.

We assume most probable values of m, ω, and ϕ have been selected from the marginal and conditional distributions in equation (7.1). We thus seek the f_j values which maximize $p(\vec{f} \mid \omega, \phi, D, M_m)$. This distribution is the joint conditional distribution for $\vec{f}$ and A, marginalized with respect to A. The joint distribution is the product of the likelihood given by equation (4.11) and the priors for A and $\vec{f}$ given by equations (4.15) and (4.16), divided by the global likelihood (conditional on ω and ϕ) given by equation (5.6). Integrating out A gives,

$$p(\vec{f} \mid \omega, \phi, D, M_m) = \frac{(N+m-1)!}{N!} \, W_m(\omega, \phi) \, \prod_{j=1}^{m} f_j^{n_j}. \tag{7.2}$$

To find the mode, we can concentrate on the $\vec{f}$-dependent part,

$$F = \prod_{j=1}^{m} f_j^{n_j}. \tag{7.3}$$

When we maximize this with respect to the f_j, we must enforce the constraint that $\sum_j f_j = 1$. We can perform the constrained maximization by using a Lagrange multiplier, maximizing the expression,

$$F' = F + \lambda \left(1 - \sum_{j=1}^{m} f_j \right), \tag{7.4}$$

with respect to the f_j and the Lagrange multiplier, λ. This leads to the not too surprising result,

$$\hat{f}_j = \frac{n_j}{N}, \tag{7.5}$$

where the circumflex denotes the mode.

7.2. *Conditional Mean of f_j*

As a first step toward a more informative summary of the shape distribution, we now seek the posterior mean of f_j when m, ω, and ϕ are specified. We can also consider this as an alternative summary to the mode if we are seeking a single curve to summarize the shape distribution. The difference between the mode and the mean will give us a sense of the asymmetry in the posterior (since each f_j is bounded below by 0 and above by 1, if the mode is not at $f_j = 1/2$ for all j, which is only possible for $m = 2$, the posterior must be asymmetric).

The mean (expected value) of f_j is given by,

$$\langle f_j \mid \omega, \phi, m \rangle \equiv \int d\vec{f} \, f_j \, p(\vec{f} \mid \omega, \phi, D, M_m). \tag{7.6}$$

We can perform the integral easily using the generalized Beta integral, equation (4.17). The result is,

$$\langle f_j \mid \omega, \phi, m \rangle = \frac{n_j + 1}{N + m}. \tag{7.7}$$

The mean is different from the mode, as expected. However, when $n_j \gg 1$ and $N \gg m$ the mean and the mode essentially coincide, so in such cases we need not worry about the asymmetry; the mean and mode give nearly indistinguishable estimates. We prefer the mean because it is easier to perform more complicated calculations (such as those required when ω or ϕ are unknown) with the mean.

The standard deviation of the marginal posterior for f_j, defined by

$$\begin{aligned} \sigma_j &\equiv \langle (f_k - \langle f_k \rangle)^2 \rangle^{1/2} \\ &= \left[\langle f_k^2 \rangle - \langle f_k \rangle^2 \right]^{1/2}, \end{aligned} \tag{7.8}$$

provides a measure of the uncertainty of the value of f_j (we have suppressed the conditioning on ω and ϕ for clarity). To calculate it, we need the second moment of f_j, which again is easily calculated using the generalized Beta integral: $\langle f_j^2\rangle = (n_j+2)(n_j+1)/(N+m+1)(N+m)$. This, combined with equations (7.7) and (7.8), give

$$\sigma_j = \sqrt{\frac{\langle f_j\rangle(1-\langle f_j\rangle)}{N+m+1}}. \tag{7.9}$$

If $\langle f_j\rangle \ll 1$ and $N \gg m$, the uncertainty in f_j is approximately $\sqrt{n_j}/N$, the "root-n" result we might have guessed.

7.3. *Averaging Over Frequency and Phase*

Recall that all of these results are conditional on ω and ϕ (and the choice of m). When ω or ϕ are unknown, we need to estimate $f(t)$ at a particular time, t, which may not correspond to estimating a particular f_j, since the uncertainty in ω and ϕ make $j(t)$ uncertain. We can take into account our uncertainty in ω and ϕ by marginalizing. This can all be done numerically using the results already found. For example, for $\langle f(t)\mid m\rangle$ we have,

$$\begin{aligned}\langle f(t)\mid m\rangle &= \int d\omega \int d\phi \int d\vec{f}\; f_{j(t)}\; p(\omega,\phi,\vec{f}\mid D, M_m)\\ &= \int d\omega \int d\phi \left[\int d\vec{f}\; f_{j(t)}\; p(\vec{f}\mid \omega,\phi, D, M_m)\right] p(\omega,\phi\mid D, M_m)\\ &= \int d\omega \int d\phi\; \langle f_{j(t)}\mid \omega,\phi,m\rangle\; p(\omega,\phi\mid D,M_m),\end{aligned} \tag{7.10}$$

that is, we just average the conditional result we found above over the marginal posterior for ω and ϕ (conditional on m) given by equation (6.4), choosing the appropriate value of j in the integral as ω and ϕ vary. Calculation of the second moment proceeds analogously, allowing us to calculate the standard deviation of the estimate of $f(t)$. Of course, if the frequency is known, the integrals over ω should be eliminated from equation (7.10).

Note that the estimate provided by equation (7.10) is essentially a weighted superposition of stepwise lightcurves, each with different phase and frequency, with weights given by the probability densities for various choices of frequency and phase. It is thus not a stepwise function, but rather a somewhat smoothed version of a stepwise function.

7.4. *Superposing Stepwise Models*

All of our lightcurve estimates so far have been conditional on the choice of a single best (most probable) number of bins. But, as we noted in our derivation of equation (6.5), Bayesian model comparison leads to a probability distribution for the models, $p(M_i\mid D,I)$; it does not isolate a single model. Thus a more complete estimate of the rate should marginalize over m as well as the various model parameters. We can calculate the marginal posterior mean for $r(t)/A = mf(t)$ as follows;

$$\begin{aligned}\langle r(t)/A\mid m>1\rangle &= \sum_{m=2}^{m_{\max}} m \int d\omega \int d\phi \int d\vec{f}\; f_{j(t)}\; p(M_m,\omega,\phi,\vec{f}\mid D,I)\\ &= \sum_{m=2}^{m_{\max}} m\, p(M_m\mid D,I) \int d\omega \int d\phi \left[\int d\vec{f}\; f_{j(t)}\; p(\vec{f}\mid\omega,\phi,D,M_m)\right]\times\\ &\qquad p(\omega,\phi\mid D,M_m)\\ &= \sum_{m=2}^{m_{\max}} m\, p(M_m\mid D,I)\; \langle f(t)\mid m\rangle,\end{aligned} \tag{7.11}$$

where $\langle f(t) \mid m \rangle$ is given by equation (7.10), and $p(M_m \mid D, I)$ can be calculated from the odds ratios in equation (5.28) using equation (2.13). An analogous equation holds for the second moment, allowing calculation of the standard deviation for $f(t)$. The resulting estimate is even more "smoothed" than that produced by equation (7.10) because it contains contributions with different numbers of bins. It may still have a significantly "boxy" shape, however.

8. RELATION TO EXISTING STATISTICS

Our Bayesian calculation bears a close relationship to other statistical methods already in use. In particular, the multiplicity that plays a key role in our analysis is related to entropy and to the χ^2 statististic used in the epoch folding method. In this section we elucidate these relationships, and highlight some of the distinctions between our Bayesian calculation and more familiar frequentist methods. A more extensive comparison will appear elsewhere (Loredo 1992b).

8.1. *Multiplicity and Entropy*

Using Stirling's approximation for the factorials appearing in equation (5.7), we can approximate the logarithm of the multiplicity as follows:

$$-\ln W(\omega, \phi) \approx \ln \frac{C}{\sqrt{N}} + \frac{1}{2}\sum_{j=1}^{m} \ln n_j + N \sum_{j=1}^{m} \frac{n_j}{N} \ln \frac{n_j}{N}, \tag{8.1}$$

where $C = (2\pi)^{m-1}$. The last, leading order term is proportional to the configurational entropy, $H(\omega, \phi)$, of the folded lightcurve (binned data) which is defined as,

$$H(\omega, \phi) = -\sum_{j=1}^{m} \frac{n_j}{N} \ln \frac{n_j}{N}. \tag{8.2}$$

Thus the joint posterior for the frequency and phase is approximately inversely proportional to the exponential of N times the entropy,

$$p(\omega, \phi \mid D, M_m) \propto e^{-NH(\omega,\phi)}. \tag{8.3}$$

Thus the most probable frequency and phase, for a given choice of the number of bins, are those leading to the lightcurve with minimum entropy.

8.2. *Multiplicity and "Chi-Squared"*

We can connect our results with the χ^2 statistic by making a further approximation. Consider evaluating the multiplicity when the binned data are nearly uniform, with $n_j \approx N/m$. This would be the case if there was no periodic signal, or if there was a signal but ω was not close to a harmonic of the true frequency. Writing,

$$\ln \frac{n_j}{N} = \ln \frac{n_j}{N/m} - \ln m, \tag{8.4),}$$

we expect the argument of the first logarithm to be near unity. Using the expansion $\ln x \approx (x-1) - \frac{1}{2}(x-1)^2$, the entropy can be approximated by,

$$\begin{aligned} H(\omega, \phi) &= +\frac{1}{N}\sum_{j=1}^{m}\left[n_j \log m + n_j \log \frac{N/m}{n_j}\right] \\ &\approx \log m + \frac{1}{N}\sum_{j=1}^{m} n_j \left(\frac{n_j - N/m}{n_j}\right) - \frac{1}{2N}\sum_{j=1}^{m} n_j \left(\frac{n_j - N/m}{n_j}\right)^2. \end{aligned} \tag{8.5}$$

The first sum is just $\sum_j n_j - \sum_j (N/m) = N - N = 0$, so we find,

$$H(\omega, \phi) \approx \log m - \frac{1}{2N} \sum_{j=1}^{m} \frac{(n_j - N/m)^2}{n_j}$$
$$= \log m - \frac{1}{2N} \chi^2(\omega, \phi), \tag{8.6}$$

where $\chi^2(\omega, \phi)$ is the same χ^2 statistic used in the epoch folding method. Using this approximation in equation (8.3), we have shown that, roughly,

$$p(\omega, \phi \mid D, M_m) \propto e^{\chi^2/2}. \tag{8.7}$$

Note, however, that the approximation breaks down just in the interesting case where the binned data are nonuniform so there is evidence for a period. Note also that, were we interested in the frequency alone, marginalizing the phase corresponds to averaging the exponential of $\chi^2/2$ over phase, not averaging χ^2 itself, as has been advocated in the literature (Collura, *et al.* 1987).

It is interesting to compare how the multiplicity is used in Bayesian calculations with how the popular epoch folding (EF) method uses the χ^2 statistic, an approximation of the logarithm of the multiplicity. In EF (see, *e.g.*, Leahy *et al.* 1983) one assumes that there is indirect evidence for a periodic signal if we can reject the null hypothesis that the data is consistent with a constant rate model. The null hypothesis is tested by examining values of the χ^2 misfit between a flat distribution and the histogram resulting from binning the events for various periods. The percent confidence level, C, associated with the test is found by identifying the maximum observed χ^2 value, $\chi^2_{\rm obs}$, and calculating

$$C = 100[1 - Q_{m-1}(\chi^2_{\rm obs})]^{N_{\rm per}}$$
$$\approx 100[1 - N_{\rm per} Q_{m-1}(\chi^2_{\rm obs})], \tag{8.8}$$

where $Q_{m-1}(\chi^2_{\rm obs})$ is the area above $\chi^2_{\rm obs}$ in the χ^2 distribution for $m-1$ degrees of freedom, and $N_{\rm per}$ is the number of independent χ^2 values found. If C is near 100%, there is evidence against the null hypothesis, which is interpreted as evidence for a periodic signal. Typically, $C \gtrsim 95\%$ is considered to be the threshold of believability; signals with a smaller confidence level are discounted as fluctuations.

Several problems complicate such analyses (Leahy *et al.* 1983, Collura *et al.* 1987). The most widely recognized problem is the choice of $N_{\rm per}$. One would like to examine a large number of frequencies to ensure finding any peaks, yet it is obvious that values of χ^2 found for two neighboring frequencies cannot really be independent, since the χ^2 values must eventually coincide as the frequencies get closer and closer together. Thus some criterion must be proposed to quantify the notion of independence. Once $N_{\rm per}$ is determined, a related problem arises: which $N_{\rm per}$ frequencies should be examined? Again, one wants to make sure that a significant peak is found, but it is not fair to choose the frequencies *a posteriori* to maximize χ^2. This would lead one to overestimate the significance of a possible pulsation.

Rigorously, both $N_{\rm per}$ and the location of the frequencies must be determined *a priori*. In the most careful studies, this is done by examining many hypothetical data sets in a Monte Carlo study. Usually, the frequencies are chosen on a regular grid, and $N_{\rm per}$ is chosen to be the largest number of examined periods for which the confidence level is accurately given by equation (8.8). Unfortunately, this number can vary by a factor of 2 to 3 depending on what underlying "true" lightcurve is used in the simulations, so serious subjectivity remains in the choice of $N_{\rm per}$.

An additional problem is the choice of number of bins. Only one choice can be fairly used, for if one were to examine χ^2 values found with different numbers of bins, difficult problems regarding the independence of the resulting values would arise. Again, the only existing guidelines for the choice of bin number are results from Monte Carlo simulations, where the bin number is chosen to optimize the ability to detect the underlying signal. Again, the result depends on the assumed underlying shape, and varies by a factor of 2 to 3 depending on what shapes are assumed.

Finally, only recently have Collura *et al.* (1987) pointed out, in another context, that the choice of phase can greatly affect EF results. Typically, analysts have ignored the signal phase, implicitly setting $\phi = 0$. In our own simulations, we have found that χ^2 can easily vary by as much as a factor of 2 as ϕ varies over $[0, 2\pi/m]$. Thus accounting for possible values of ϕ is crucial. Rather than examining a grid of ϕ values and facing the resulting problems of independence, Collura *et al.* average χ^2 over phase, and use simulations to show that, at least asymptotically, the phase-averaged χ^2 still follows the χ^2 distribution. Unfortunately, we are unaware of any published period searches that have recognized the problems associated with choosing ϕ, and the need to use a phase-averaged χ^2 or some other method to account for uncertainty in ϕ.

From the Bayesian point of view, all of these problems arise because of the inability of frequentist methods to handle nuisance parameters. It is true that not knowing the frequency (or phase or number of bins) for the signal should decrease the significance of the result. Bayesian methods accomplish this in a unique, objective manner through marginalizing. The unknown frequency is accounted for by averaging (integrating) over frequency rather than by maximizing. Only the range of frequencies searched is relevant; the actual number of frequencies used in the integration is irrelevant. The unknown phase is also accounted for by integrating, but the Bayesian calculation tells us exactly what to integrate: a probability distribution for the phase. This distribution is approximately the *exponential* of the χ^2 quantity averaged in the frequentist treatment. Finally, the unknown number of bins can also be accounted for by summing; alternatively, an optimum number of bins can be found. But this optimum number is determined *a posteriori* by the *observed* data, not by *a priori* Monte Carlo simulations of *hypothetical* data. It is a measure of the complexity of the observed data, and is not determined to optimize detection in a class of unobserved, hypothetical data sets.

The Bayesian calculation also clearly distinguishes the tasks of signal detection and parameter estimation. We are not aware of any astronomical work on using the EF method to find rigorous estimates and confidence regions for the period of a detected signal. Most studies seem to consider the width of the χ^2 peak to give a rough estimate of the frequency resolution. Since χ^2 is roughly the logarithm of the probability for ω, this greatly overestimates the uncertainty of the frequency. The Bayesian marginal distribution for ω determines ω with an accuracy that can be orders of magnitude greater than that found from the width of the χ^2 peak.

An apparent advantage of the EF method, and other frequentist methods based on rejecting a constant signal model with a "goodness-of-fit" test, is that they do not assume a particular periodic model for the signal (though they require such a model to determine $N_{\rm per}$ and the number of bins). In contrast, any Bayesian signal detection calculation must make explicit assumptions about the nature of the periodic or nonperiodic signal that might have produced the data. However, we have demonstrated that a Bayesian calculation *with an explicit choice of signal models* leads one to consider essentially the same statistic examined in an apparently alternative-free frequentist test: our piecewise constant model leads one to bin the data and calculate a quantity related to χ^2. Similar results are known for many other popular goodness-of-fit statistics. For example, Loredo (1992b) demonstrates that the frequentist Rayleigh and Z_n^2 tests for periodicity use exactly the same statistics required in Bayesian model comparisons with specific smooth models for the signal. In this way, Bayesian thinking illuminates the assumptions implicit in the choice of a goodness-of-fit statistic.

We believe that the ability of Bayesian methods to treat nuisance parameters, and the need to explicit identify alternatives in a Bayesian analysis, make Bayesian results both more precise and more intuitive than their frequentist counterparts. Because the signal model must be specified, Bayesian methods force one to identify all nuisance parameters relevant to the problem being considered. And because unknown parameters can be correctly accounted for by marginalizing, Bayesian probabilities can be taken at face value: they indicate directly the intuitive degree of certainty one should have in the stated conclusions, given the assumptions of the analysis. In contrast, it is not unusual to consider a "2σ" (or even "5σ") frequentist result to be only just believable, despite the fact that the probability of such results being false alarms is formally only 0.05 (or 6×10^{-7} for 5σ). This nonintuitive behavior arises, we believe, because nuisance parameters are often unidentified and always incorrectly eliminated in frequentist analyses.

Bayesian calculations can correctly handle nuisance parameters because they consider probabilities of *hypotheses*, conditional on the observed data. When the hypothesis of interest is the disjunction of several simpler hypotheses, the probability of the interesting hypthesis is simply the sum of the probabilities of its constituents (here the periodic hypothesis is the disjunction of many simpler periodic models). For historical and philosophical reasons, frequentist theory deems probabilities of hypotheses to be meaningless, and so must rely on probabilities of *hypothetical data*, conditional on a single hypothesis. This causes problems when the hypothesis of interest is the disjunction of many simpler hypotheses, because one particular hypothesis must be assumed true in the analysis. Thus the superiority of Bayesian methods for treating nuisance parameters arises from very basic and fundamental distinctions between the Bayesian and frequentist viewpoints. But even apart from such formal, somewhat philosophical considerations, the Bayesian notion of the probability of an hypothesis seems to us to be a much more intuitive measure of the plausibility of an hypothesis than frequentist probabilities of more extreme data.

9. APPLICATION TO SIMULATED DATA

In this section we illustrate the use of our Bayesian method by applying it to simulated data. We perform detection and estimation calculations both for data from a stepwise signal and data from a smooth, sinusoidal signal.

For our first example, we simulated data from a stepwise lightcurve with 7 bins. The period of the signal was $P = 2.05633$ s ($\omega = 3.05553$ s^{-1}), and we simulated data over an observing interval of $T = 60$ s (about 29 periods). The average signal rate was 4.6 s^{-1} (≈ 9.5 events per period). Approximately 76% of this rate was in a constant background; 16% (≈ 1.5 events per period) was in a 1-bin pulse in bin 2, and 8% ($\approx .75$ events per period) was in a 1-bin pulse in bin 4. A total of 276 events were simulated. The simulated data were analyzed "blind," though by construction the signal period was chosen to be in the "twice Nyquist" range we standardly search, and the number of bins was chosen to be in the range 2 to 12 (we typically use $m_{\max} = 12$ to 15, a computationally convenient range that is capable of representing the simulated and real lightcurves we have analyzed).

In Figure 3 we have plotted $O_{m1}(\omega)$ (eqn. (5.25)) in the range $\omega_{\rm lo} = 20\pi/T$ to $\omega_{\rm hi} = 2\pi N/T$ (twice effective Nyquist frequency), for $m = 7$ and $\nu = 1$. This is the quantity that

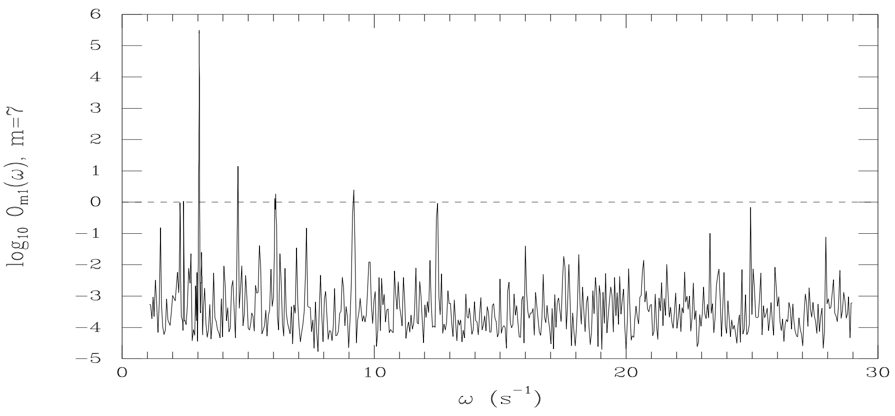

Fig.3 $O_{m1}(\omega)$ versus ω for $m = 7$ bins, for data simulated from a stepwise rate function with 7 bins and $\omega = 3.05553$ described in the text.

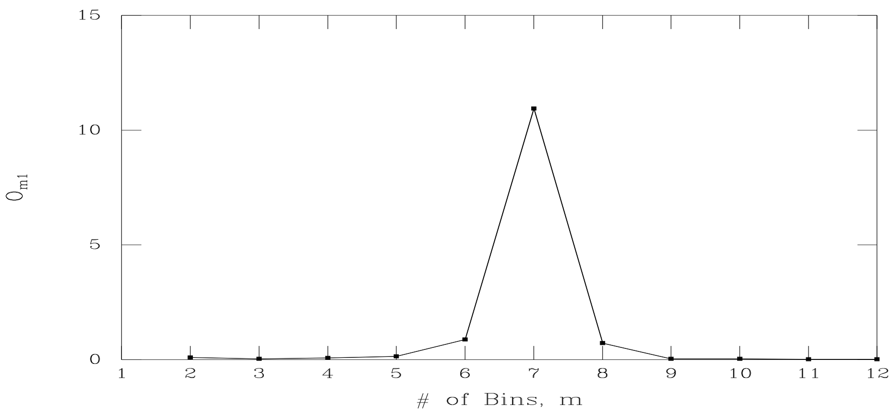

Fig.4 O_{m1} versus m for data simulated from a stepwise rate function with 7 bins.

one would calculate were the frequency and number of bins known, and one would then evaluate it only at the known frequency.

We plot it for illustrative purposes. Were the frequency and number of bins known, Figure 3 shows that the odds ratio in favor of a periodic signal would be greater than 2.5×10^5 to 1. But if the frequency and number of bins are not known, the evidence for a periodic signal decreases due to the Ockham factors associated with m and ω. This is illustrated by Figure 4, which shows O_{m1} versus m, as given by equation (5.28). Here the odds ratio in favor of the $m = 7$ model has dropped from 2.5×10^5 to 10.9. The odds ratio in favor of a periodic signal is the sum of the odds ratios for the models considered; here we find $O_{\rm per} = 13.03$, corresponding to a 93% probability that the signal is periodic, given the class of models assumed.

Since this is strong evidence for a signal, we considered a signal to be present, and estimated its frequency and shape. Figure 5 shows the marginal posterior density for ω, for $m = 7$. The large-scale plot shows a strong and extremely narrow peak near the true frequency.

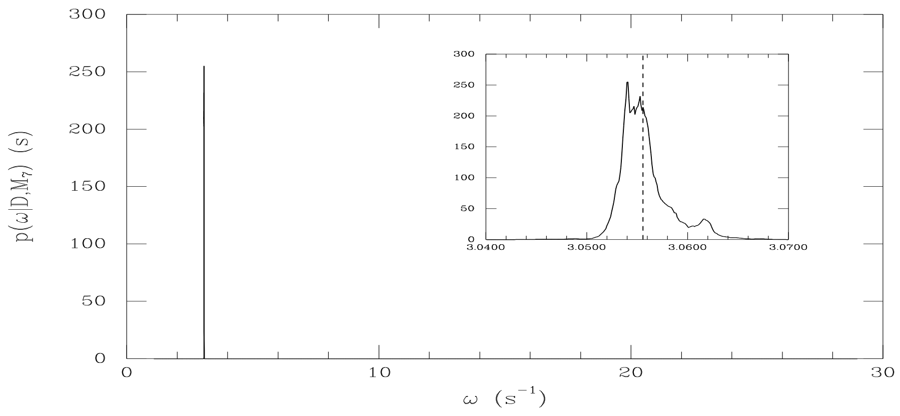

Fig.5 Marginal posterior density for ω, for $m = 7$ bins, for data simulated from a stepwise rate function with 7 bins and $\omega = 3.05553$. Inset shows a detail of the narrow peak; the dashed vertical line indicates the true frequency.

The inset details this peak, which is indeed at the true frequency; the frequency is measured with an accuracy of approximately 0.6 mHz (for comparison, the "Nyquist frequency" $1/T = 17$ mHz). Figure 6 shows the estimated lightcurve shape, calculated according to equation (7.11). We calculated the mean and standard deviation for $r(t)/A$ at 49 times, and plotted the two resulting (mean±standard deviation) curves as solid lines. The true underlying rate is shown with a dotted line. The shape estimate satisfactorily reproduces the true shape to within the uncertainties.

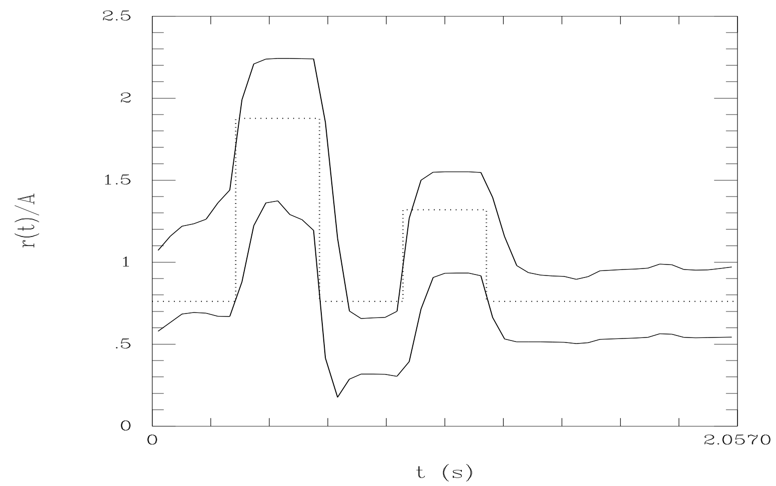

Fig.6 Shape estimate for data simulated from a stepwise rate function with 7 bins and 2.05633 s period. Solid curves show ±1 standard deviation estimates; dotted curve shows true shape.

The same data were then analyzed using the epoch folding (EF) method. In Figure 7 the χ^2 statistic is plotted versus ω for $m = 7$ bins. The values of χ^2 were calculated on a grid spacing of $\Delta\omega = \pi/4T$, for a total of 1065 "independent" frequencies in the same frequency range as used for the $O_{m1}(\omega)$ calculation. To account for the unknown phase, at

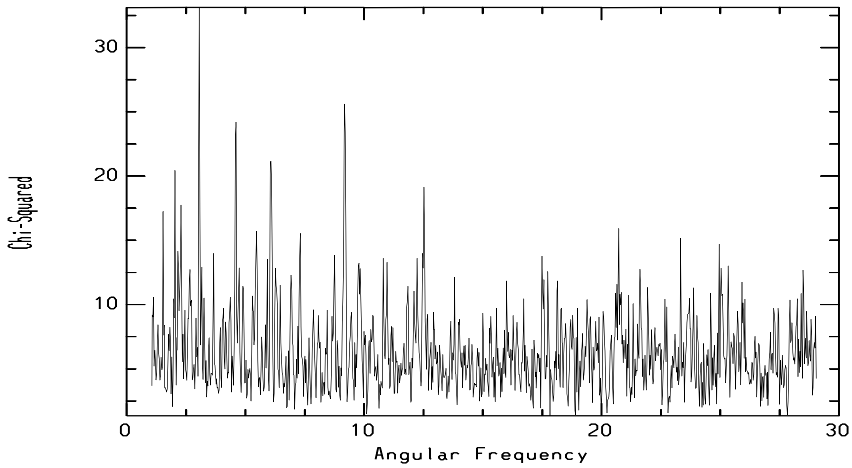

Fig.7 Phase-averaged χ^2 versus ω for $m = 7$ bins, from an epoch folding analysis of the same data used in Figure 3.

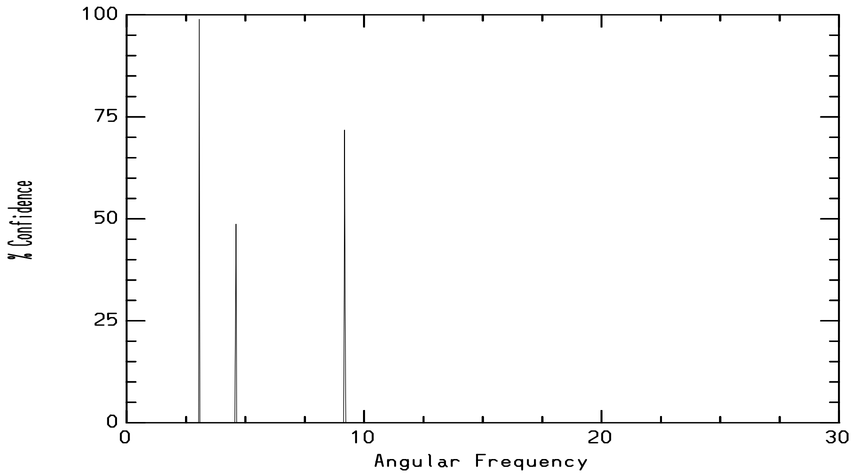

Fig.8 Percent confidence associated with phase-averaged χ^2, versus ω, for $m = 7$ bins, from an epoch folding analysis of the same data used in Figure 3.

each trial frequency we averaged χ^2 itself over phase, as advocated by Collura *et al.* (1987).

The χ^2 values were converted to percent confidence levels, C, using (8.8) , and are shown plotted in Figure 8 versus ω (rigorously, a confidence level should be associated only with the single highest peak).

The highest peak at $\omega = 3.0507$ s^{-1} ($P = 2.0596$ s), has a $C = 98.7\%$, corresponding to 2.5 σ, which is not usually considered a significant detection. In addition the confidence associated with the largest peak is only 1.4 times that of the next highest peak. In contrast the second highest peak in $O_{m1}(\omega)$ is down by four orders of magnitude.

Next we considered data simulated using a smooth, sinusoidal lightcurve, to see how the method behaves when the lightcurve is not in the stepwise model class. The rate function we used was of the form,

$$r(t) = A\left[1 + f\sin(\omega t + \phi)\right], \tag{9.1}$$

where A is the time-averaged rate, f is the pulsed fraction, which must lie in the interval

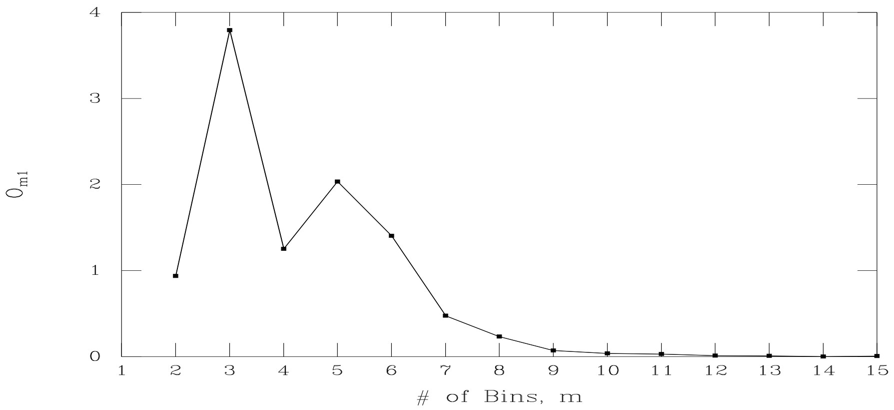

Fig.9 O_{m1} versus m for data simulated from the sinusoidal rate function described in the text.

$[0, 1]$, and ω and ϕ are the usual frequency and phase. For the simulations, we set $A = 1$ s^{-1}, $f = 0.5$, $\omega = 2\pi$ s^{-1} ($P = 1$ s), and $\phi = \pi/4$. Thus on the average there is one pulsed event for every two periods. We simulated several 200 s observations. In approximately half of the observations, no signal was detected (using either the Bayesian calculation or the EF method), and so no parameter estimates were made. Figures 9–11 illustrate the results from one of the simulations in which the signal was detected. Figure 9 shows the odds ratio O_{m1} versus m; the prior range for the frequency was the usual Nyquist range. Note that no particular model is significantly more probable than its competitors; the smooth underlying shape causes the probability to be distributed over several models. Figure 10 shows the marginal distribution for the frequency; this distribution peaks within 0.5 mHz of the true frequency, and is only ≈ 1.7 mHz wide. Finally, Figure 11 shows the "1σ" region for the estimated signal shape, with the true sinusoidal shape shown as a dotted curve. The superposition of stepwise curves with various numbers of bins and various phases and frequencies has nicely estimated the sinusoidal shape to within the uncertainty provided by the sparse data.

We wish to emphasize that we do not consider our model to be optimal for detecting pulsations from smooth lightcurves; other smooth parametric models (such as the sinusoidal and exponentiated sinusoidal models considered in Loredo 1992b) can more sensitively detect such signals because they can model the signal with fewer parameters, and so pay a smaller Ockham penalty. The best way to search for a signal whose shape is truly unknown is to include smooth models in the periodic class in addition to the stepwise model, so that both "spiky" and smooth signals can be optimally detected. However, our simulations demonstrate that our general model can usefully detect smooth signals, and accurately estimate their frequency and shape.

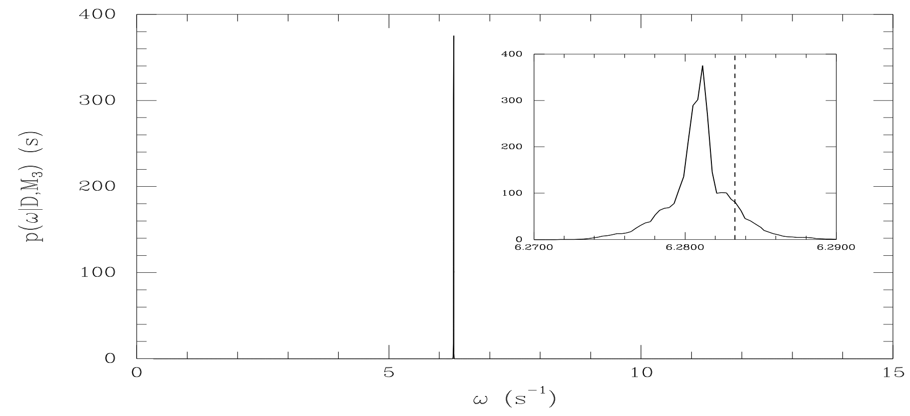

Fig.10 Marginal posterior density for ω, for $m = 3$, for data simulated from a sinusoidal rate function with $\omega = 2\pi$. Inset shows a detail of the narrow peak; the dashed vertical line indicates the true frequency.

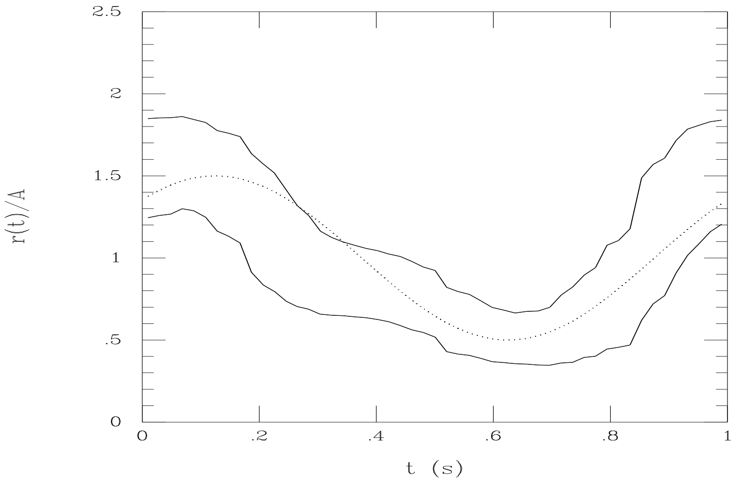

Fig.11 Shape estimate for data simulated from a sinusoidal rate function with 1 s period. Solid curves show ± 1 standard deviation estimates; dotted curve shows true shape.

10. CONCLUSIONS

Bayesian probability theory has yielded a solution to the important problem of the detection of a periodic signal in a data set consisting of the locations or times of individual events, when we have no *a priori* knowledge about the nature of the periodic signal. This has been accomplished by using Bayes' theorem to calculate the probabilities of nonperiodic and periodic models. The class of periodic models we adopt spans a wide variety of shapes, and the probability of the class as a whole is found by marginalizing (integrating) over all the shapes it can describe. At first sight this might appear to be an overwhelming computational problem. But our choice of models allows us to perform the needed marginalizations analytically, leading to an algorithm with computational speed comparable to that of the popular epoch folding method based on χ^2.

The probability of a model contains a factor—the Ockham factor—that is a quantitative expression of Ockham's razor. It penalizes the periodic models for their greater complexity in a manner determined by the data and the structure of the models, and not by subjective criteria. In traditional statistical tests, Ockham's razor is invoked to justify considering simpler models to be more plausible *a priori*. In the Bayesian analysis the Ockham factor is a derivable consequence of the basic sum and product rules of probability theory and arises *a posteriori*; our calculations assume the nonperiodic and periodic model classes to have *equal* prior probabilities, and the members of the periodic class to also have 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.

When there is appreciable probability for the periodic models, the problem becomes one of estimating the period and shape of the lightcurve. We have derived Bayesian solutions to both of these problems. With our models, the posterior probability density for the signal frequency is shown to be inversely proportional to the multiplicity or configurational entropy of the binned events. The probability is a maximum for the period corresponding to minimum entropy.

When there are gaps in the observing period (*e.g.*, from dead time, or intermittent observing) our method can be applied without modification. However, to simplify the calculations, we have assumed that the gaps affect the phase bins nearly uniformly. This may not be the case if a strong periodic signal is observed with significant dead time, since the dead time will accumulate preferentially in highly populated bins. Appendix B outlines a simple modification to deal with significant data gaps which is computationally efficient.

The ability of Bayesian methods to straightforwardly handle gaps, determine the optimum number of bins, and quantify Ockham's razor, are features which are unique to the Bayesian approach. We believe these features commend the use of our method, and in Appendix A we provide a "recipe" for those interested in applying the method. In Appendix C we show how the same approach can also be used to test for nonperiodic source variability. Further, since the features mentioned above are generic to Bayesian inference, we believe our work motivates the application of Bayesian methods to a much wider variety of statistical problems arising in astrophysical data analysis.

ACKNOWLEDGMENTS

We thank David MacKay for helping us better understand Bayesian model comparison and Ockham factors. 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 (NAG 5-1758), NASA grant NAGW–666, and NSF grants AST–87–14475 and AST–89–13112 at Cornell University.

APPENDIX A.

RECIPE FOR USE OF THE METHOD

We present here a brief outline of how one might perform the numerical calculations required to implement our method. We outline calculations for searching for pulsations at an *unknown* frequency. When the frequency is known *a priori*, several steps can be eliminated.

Logically, one should address the model comparison problem before actually estimating the parameters of any particular model, and our paper followed this logical path. But practically, solution of the model comparison problem requires one to calculate the quantities needed to address the estimation problem for each model, so numerical calculations must proceed in a somewhat reversed order, as follows.

1. Decide on the frequency range $(\omega_{\rm hi}, \omega_{\rm lo})$ to search. We typically use $\omega_{\rm lo} = 20\pi/T$ and $\omega_{\rm hi} = \pi N/T$ to $2\pi N/T$. Also, determine the maximum number of bins that will be considered, $m_{\rm max}$. We routinely use $m_{\rm max} = 12$ to 15 for the sparse data sets we have analyzed; long data sets with a lot of structure may require a larger value.

2. For each value of m being considered, calculate $O_{m1}(\omega)$ given by equation (5.25) on a grid of ω values. These values should be stored, as they will be required for estimating ω in step 4. A grid spacing $\Delta\omega \sim \pi/10T$ is usually adequate to locate any significant peaks; we then fill in more points to resolve strong peaks. The numerical integral in equation (5.25) can be adequately calculated with the trapezoidal rule, but for sparse data sets ($N \lesssim 10^3$) the integral can be calculated exactly, as follows. The multiplicity $W_m(\omega, \phi)$ is a stepwise function of ϕ; its value changes only when ϕ changes enough to move an event across a bin boundary. Thus, starting at $\phi = 0$, as one bins events to calculate $W_m(\omega, \phi)$, note the smallest distance in phase between an event and the following bin boundary. Multiply the resulting $W_m(\omega, \phi)$ value by this distance, increment ϕ by the distance, and iterate until ϕ reaches its upper limit, accumulating the desired integral. Whether finding the integral exactly or with a quadrature rule, the integration is most efficiently done by integrating over the interval $[0, 2\pi/m]$ and multiplying the result by m, since other phases correspond to cyclicly permuting the bins, which does not change the value of $W_m(\omega, \phi)$.

3. For each m, O_{m1} is found by multiplying the $O_{m1}(\omega)$ values found in step 2 by the prior for ω, equation (4.14), and integrating over ω. Equation (5.2) or (5.3) can then be used to calculate the probability or odds for the periodic class to determine if there is evidence for a periodic signal, in which case the following steps are performed to get frequency and shape estimates.

4. For each m, the marginal distribution for ω is the product of $O_{m1}(\omega)$ and the prior for ω, normalized with respect to ω (compare equations (5.25) and (6.2)). The information needed to calculate this distribution is already available from steps 2 and 3. The marginal distribution conditional on the most probable value of m is usually an adequate summary of the information about ω. Alternatively, the results for various m can be averaged together using equation (6.5) and the values of O_{m1} found in step 3. This is the most complete summary of the information provided by the data about the frequency.

5. Finally, shape estimates can be found as follows. For each m, the phase and frequency averaged estimates are calculated according to equation (7.10). The innermost phase integral is performed as outlined in step 2, and the outermost frequency integral is performed by summing over frequency. To speed the calculation, we only consider frequencies that are, say, at least 1% as probable as the most probable frequency. Also, the calculation is most efficiently done if the mean and second moment of $f(t)$ are calculated simultaneously, for all desired values of t, in the innermost integration over ϕ. The resulting moments for each m should then be combined using equation (7.11), with the model probabilities already available from step 3. The standard deviation is then found from the mean and second

moment using equation (7.8), and curves are plotted as a function of t indicating the mean $\pm$ the standard deviation.

On the question of computational speed, we have found that the Bayesian detection calculations can be performed at least as quickly as the calculations required by the EF method (with phase-averaged χ^2). Most of the computational burden arises from binning the events, which is required for both methods. By pre-computing an array of natural logarithms of factorials that are repeatedly required in the evaluation of the multiplicity, the remaining calculations required for the Bayesian method can be performed relatively quickly.

APPENDIX B.

A MODIFIED SOLUTION FOR SIGNIFICANT DATA GAPS

An important feature of our method is that it is computationally tractable, because we can analytically marginalize over the f_j parameters. However, this relies on the assumption that any data gaps are not concentrated in certain bins, so that the argument of the exponential in equation (4.8) becomes independent of the f_j. This corresponds to assuming that the bin time factors, s_j, are all nearly equal to one. If the number of periods in the observing time is small, or if a strong periodic signal is present, the data gaps may not affect the phase bins uniformly, leading to s_j values differing significantly from unity. We now consider a useful solution for this case, which again permits an analytic marginalization of the f_j parameters. It is not a rigorous Bayesian solution, in a sense elaborated upon below; but it should prove useful in situations where our earlier approximation fails.

We will redefine the f_j parameters to guarantee that the argument of the exponential in equation (4.8) is independent of the f_j, regardless of the values of the s_j. We do this by writing r_j as,

$$r_j = mAf_j/s_j, \tag{B.1}$$

where r_j, m, A, and s_j have the same meaning as before. The shape of the lightcurve is now described by $h_j = f_j/s_j$, and the f_j are related to the r_j according to,

$$f_j = \frac{r_j s_j}{mA} = \frac{r_j \tau_j}{AT}. \tag{B.2}$$

Thus f_j is now the fraction of events expected to fall into bin j, and again $\sum_j f_j = 1$. This new definition of f_j insures that $\sum f_j \tau_j = AT$, so that equation (4.8) becomes

$$P(D \mid \omega, \phi, A, \vec{f}, M_m) = \Delta t^N (mA)^N e^{-AT} S(\omega,\phi) \left[\prod_{j=1}^{m} f_j^{n_j}\right], \tag{B.3}$$

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

$$S(\omega,\phi) = \prod_{j=1}^{m} s_j^{-n_j}. \tag{B.4}$$

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

So far we have simply reparametrized our earlier model. In fact, when $s_j = 1$ for all j, this is an equivalent parametrization to the one used in the body of our paper. But our treatment of this model will differ from that of the original models with respect to the prior probabilities we assign to the parameters. We take all the priors to be of the same form as those presented in § 4. But since we have redefined the f_j parameters, this corresponds to a different state of prior information than that assumed above. We again assume the same broad prior joint density for $\vec{f}$ (equation (4.16)), which enables the marginalization of $\vec{f}$ to be carried out using a generalized Beta integral as before. The result for $p(D \mid \omega, \phi, M_m)$ is the same as equation (5.6) with the addition of the binning factor in the numerator.

The results of § 5 for the odds ratio, and § 6 for the posterior probability of the frequency, follow as before with the inclusion of the binning factor. For example, equation (5.25) becomes

$$O_{m1}(\omega) = \frac{1}{2\pi\nu} \binom{N+m-1}{N}^{-1} \int_0^{2\pi} d\phi \, \frac{S(\omega,\phi)m^N}{W_m(\omega,\phi)}. \tag{B.5}$$

It is of interest to consider the effect of the binning factor on $O_{m1}(\omega, \phi)$:

$$O_{m1}(\omega,\phi) = \frac{1}{\nu} \binom{N+m-1}{N}^{-1} \frac{S(\omega,\phi)m^N}{W_m(\omega,\phi)}. \tag{B.6}$$

When the binning factor is unity, $O_{m1}(\omega, \phi)$ is inversely proportional to the probability of the distribution occurring by chance. For example, if $N = 100$ events and $m = 6$ bins, the probability of all events falling into one bin by chance would be $m^{-N} = 1.5 \times 10^{-78}$. If this distribution occurred for some combination of ω and ϕ, it would be strong evidence in favor of a periodic signal. However, if this configuration arose simply because all of the recorded time intervals of data happened to correspond to one bin for a particular choice of ω and ϕ, we would not consider the configuration improbable, it would simply be a consequence of the gaps in the data. For this case $S(\omega, \phi) = m^{-N}$, which when multiplied by the inverse of the probability gives 1 for the last factor in equation (B.6). Thus the binning factor ensures that the odds ratio behaves as expected, not favoring the periodic model when apparent periodicity is due to gaps in the data.

Finally we consider the effect of the binning factor on the estimation of the lightcurve shape. In § 7 this was discussed in terms of the mode, mean and variance of the joint probability density distribution of the f_j. Since the shape is now described by $h_j = f_j/s_j$, the replacements for equations (7.5), (7.7) and (7.9) are as follows.

$$\hat{h}_j = \frac{n_j}{s_j N}, \tag{B.7}$$

$$\langle h_j \mid \omega, \phi, m \rangle = \frac{n_j + 1}{s_j(N+m)}. \tag{B.8}$$

$$\sigma_j = \frac{1}{s_j} \sqrt{\frac{s_j \langle h_j \rangle (1 - s_j \langle h_j \rangle)}{N+m+1}}. \tag{B.9}$$

The full estimate of the shape obtained by marginalizing over ω, ϕ and m, can be computed from equations (7.10) and (7.11) by replacing $f(t)$ by $h(t)$.

This treatment of significant gaps is equivalent to the analysis presented in the body of this work when there are no gaps, and behaves reasonably when large gaps are present, as noted in the discussion following equation (B.6). But it is not a rigorous Bayesian solution for the following reason. This analysis has assumed a constant prior for the reparameterized f_j values. Using equation (B.1), we can map this prior onto the rate parameters, r_j. It is clear from this equation that the resulting prior will depend on the s_j values, which are determined by the data. However, a prior probability cannot depend on the data. It is for this reason that this calculation is not rigorously Bayesian.

APPENDIX C.

STEPWISE MODEL FOR SOURCE VARIABILITY

Many astrophysical sources are observed to exhibit nonperiodic variability on a variety of temporal scales. The stepwise model, used in this paper, can just as easily represent a nonperiodic variable lightcurve. 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. This provides a Bayesian counterpart to the method proposed by Collura *et al.* (1987). All that is necessary is to set the period , $P = 2\pi/\omega$, equal to the total duration of the observations (including any gaps). With ω fixed in this way the desired odds ratio is given by equation (5.25), or if the data contain gaps, by equation (B.5). In this problem data gaps from intermittent observing will likely result in bin integration factors, s_j, which differ significantly from unity.

When the odds ratio is greater than unity, indicating that the source is variable, the problem becomes one of estimating the characteristic time scale(s) of the variability and the shape of the lightcurve. The former can be determined from the dependence of the odds ratio on the number of bins, m, which will exhibit one or more maxima as a result of the Ockham factor. The lightcurve shape can be obtained from the equations of § 7, modified in the case of data gaps according to Appendix B.

In this variability model, the phase, ϕ, merely determines where the bin boundaries are. As ϕ is increased from zero, the bin at the end of the observed interval "wraps around" and picks up events at the beginning of the interval. Thus this model has a periodic boundary condition. If this is not desirable, the period can be lengthened by the width of one bin to prevent "wrap around." If this is done, however, the modification of Appendix B must be implemented, because the bins on the boundaries will have significant gaps.

REFERENCES

Bretthorst, G. L. 1988, Bayesian Spectrum Analysis and Parameter Estimation (New York: Springer-Verlag)

Broadhurst, T. J., Ellis, R. S., Koo, D. C., & Szalay, A. S. 1990, Nature, 343, 726

Cleveland , T. 1983, Nuc. Instr. and Meth., 214, 451

Collura, A., Maggio, A., Sciortino, S., Serio, S., Vaiana, G. S., & Rosner, R. 1987, ApJ, 315, 340

de Jager, O. C., Swanepoel, J. W. H., & Raubenheimer, B. C. 1986, A&A, 170, 187

Jeffreys, H. 1961, Theory of Probability (Oxford: Oxford University Press)

Leahy, D. A., Darbro, W., Elsner, R. F., Weisskopf, M. C., Sutherland, P. G., Kahn, S., & Grindley, J. E. 1983, Ap. J. 266, 160

Loredo, T. J. 1990, in Maximum Entropy and Bayesian Methods, ed. P. F. Fougere (Dordrecht: Kluwer), 81

Loredo, T. J. 1992a, in Statistical Challenges in Modern Astronomy, ed. E. Feigelson & G. Babu (New York: Springer-Verlag), in press

Loredo, T. J. 1992b, in preparation.

11. FIGURE CAPTIONS

Figure 1: Periodic model, M_m, assumes the periodic signal plus background can be modeled by a stepwise distribution in m bins as illustrated here.

Figure 2: Logarithm of the Ockham factor, Ω_m, for N uniformly distributed events. (a) Ω_m versus N for $m = 6$ bins. (b) Ω_m versus m, for $N = 420$ events. Only points with N/m an integer are plotted. Models with smaller Ockham factors are penalized over models with larger Ockham factors. Thus, the penalty increases with increasing number of bins.

Figure 3: $O_{m1}(\omega)$ versus ω for $m = 7$ bins, for data simulated from a stepwise rate function with 7 bins and $\omega = 3.05553$ described in the text.

Figure 4: O_{m1} versus m for data simulated from a stepwise rate function with 7 bins.

Figure 5: Marginal posterior density for ω, for $m = 7$ bins, for data simulated from a stepwise rate function with 7 bins and $\omega = 3.05553$. Inset shows a detail of the narrow peak; the dashed vertical line indicates the true frequency.

Figure 6: Shape estimate for data simulated from a stepwise rate function with 7 bins and 2.05633 s period. Solid curves show ± 1 standard deviation estimates; dotted curve shows true shape.

Figure 7: Phase-averaged χ^2 versus ω for $m = 7$ bins, from an epoch folding analysis of the same data used in Figure 3.

Figure 8: Percent confidence associated with phase-averaged χ^2, versus ω, for $m = 7$ bins, from an epoch folding analysis of the same data used in Figure 3.

Figure 9: O_{m1} versus m for data simulated from the sinusoidal rate function described in the text.

Figure 10: Marginal posterior density for ω, for $m = 3$, for data simulated from a sinusoidal rate function with $\omega = 2\pi$. Inset shows a detail of the narrow peak; the dashed vertical line indicates the true frequency.

Figure 11: Shape estimate for data simulated from a sinusoidal rate function with 1 s period. Solid curves show ± 1 standard deviation estimates; dotted curve shows true shape.