A Bayesian method for the analysis of deterministic and stochastic time series

C. A. L. Bailer-Jones

Max Planck Institute for Astronomy, Heidelberg, Germany
e-mail: calj@mpia.de

ABSTRACT

I introduce a general, Bayesian method for modelling univariate time series data assumed to be drawn from a continuous, stochastic process. The method accommodates arbitrary temporal sampling, and takes into account measurement uncertainties for arbitrary error models (not just Gaussian) on both the time and signal variables. Any model for the deterministic component of the variation of the signal with time is supported, as is any model of the stochastic component on the signal and time variables. Models illustrated here are constant and sinusoidal models for the signal mean combined with a Gaussian stochastic component, as well as a purely stochastic model, the Ornstein-Uhlenbeck process. The posterior probability distribution over model parameters is determined via Monte Carlo sampling. Models are compared using the "cross-validation likelihood", in which the posterior-averaged likelihood for different partitions of the data are combined. In principle this is more robust to changes in the prior than is the evidence (the prior-averaged likelihood). The method is demonstrated by applying it to the light curves of 11 ultra cool dwarf stars, claimed by a previous study to show statistically significant variability. This is reassessed here by calculating the cross-validation likelihood for various time series models, including a null hypothesis of no variability beyond the error bars. 10 of 11 light curves are confirmed as being significantly variable, and one of these seems to be periodic, with two plausible periods identified. Another object is best described by the Ornstein-Uhlenbeck process, a conclusion which is obviously limited to the set of models actually tested.

Key words. methods: statistical – brown dwarfs

1. Introduction

When confronted with a univariate time series, we are often interested in answering one or more of three questions. Which model best describes the data? What values of the parameters of this model best explain the data? What range of values does the model predict for the signal at some arbitrary time? These are questions of inference from data, and can be summarized as model comparison, parameter estimation and prediction, respectively.

Probabilistic modelling provides a self-consistent and logical framework for answering these questions. In this article I introduce a general method for time series model comparison and parameter estimation. The principle is straight forward. The time series data comprise a set of measurements of the signal at various times, with measurement uncertainties generally in both signal and time. We write down a parametrized model for the variation of the signal as a function of time. This could be a deterministic function or a stochastic model or, more generally, a combination of the two. An example of a combined model is a sinusoidal variation of the mean of the signal on top of which is a Gaussian stochastic variation in the signal itself, which is *not* measurement noise. A purely stochastic model is one in which the expected signal evolves according to a random distribution, e.g. a random walk. Given this generative model and a noise model for the measurements, we then calculate the likelihood distribution of the data for different values of the model parameters. Rather than identifying just the single best fitting parameters, I use a Monte Carlo method to sample the posterior probability density function (PDF) over the model parameters. In addition to providing uncertainties on the inferred parameters, this also provides a measure of the goodness-of-fit of the overall model, in the form of the marginal likelihood (evidence), or the cross-validation likelihood (defined here). In this way we can identify the best overall model from a set, something which frequentist hypothesis testing can be notoriously bad at (e.g. Berger & Sellke 1987; Kass & Raftery 1996; Jaynes 2003; Christensen 2005; Bailer-Jones 2009).

There of course exist numerous time series analysis methods which attempt to answer one or more of the questions posed, so the reader may wonder why we need another one. For example, if we focus on periodic (Fourier) models, then we can calculate the power spectrum or periodogram in order to identify the most significant periods and to estimate the amplitudes of the components. If we work in the time domain, we could do least squares fitting of a parametrized model (e.g. Chatfield 1996; Brockwell & Davis 2002). However, many of these methods can only answer one of the posed questions, are limited to a restricted set of models or specific types of problems, do not take into account uncertainties in the signal and/or time, are limited to equally-spaced data, do not provide uncertainty estimates on the parameters, or make other restrictive assumptions. The method introduced in the present work is quite general, and firmly embedded in a probabilistic approach to data modelling (see, e.g., von Toussaint 2011, for an introduction). This makes it powerful, but at the price of considerably higher computational cost. Yet in many applications this is a price we should be willing to pay for hard-won data, and should often be preferred to ad hoc, suboptimal recipes.

The first two sections of this article – occupying about a quarter of its length – are dedicated to a complete description

Table 1. Primary notation.

Symbol	Definition
s_j	measured time of jth event
σ_{s_j}	standard deviation in s_j
t_j	(unknown) true time of the jth event
y_j	measured signal of jth event
σ_{y_j}	standard deviation in y_j
z_j	(unknown) true signal of the jth event
D_j	$= (s_j, y_j)$ measurements for the jth event
σ_j	$= (\sigma_{s_j}, \sigma_{y_j})$ estimated uncertainties in D_j
D	$= \{D_j\}$ set of measurements for J events
σ	$= \{\sigma_j\}$ estimated uncertainties in D
D_k	set of measurements for events in partition k
D_{-k}	set of measurements for all events not in partition k
M	time series model
θ	$= (\theta_1, \theta_2, \theta_3)$, parameters of the time series model
$\eta(t; \theta_1)$	deterministic model of the expected true signal (TSMod1)
log	logarithm base 10
$\mathcal{N}(x; \mu, V)$	Gaussian in x with mean μ and variance V

of the method: Sect. 2 covers the method itself and the likelihood calculation, whereas the model comparison method described in Sect. 3 is general. The mathematics here is relatively simple and intuitive. More involved is the introduction of the Ornstein-Uhlenbeck process into the method. This, as well as the numerical approximations which allow the likelihood integrals to be evaluated more rapidly, are presented in the appendices. Section 4 summarizes how to use the method. Most of the rest of the paper (about a third in total) is concerned with the application of the method, first to a simulated time series (Sect. 5), and then to real astronomical data (Sect. 6). These are the light curves of 11 ultra cool dwarfs (low mass stars or brown dwarfs), which an earlier study has claimed show statistically significant variability. Although these data are used here primarily for demonstration purposes, this reanalysis of these data is astrophysically interesting, identifying a possible model and possible periods in two cases. I summarize and conclude in Sect. 7.

The method developed here is related to the `artmod` method introduced in Bailer-Jones (2011; hereafter CBJ11), which is a model for time-of-arrival time series. The present method extends this to model time series with noisy signal values at each measured time.

The notation used is summarized in Table 1.

2. The time series method

2.1. Data and model definition

We have a set of J events, each defined by its time, t, and signal, z. For each event j, our measurement of the time of the event, t_j, is s_j with a standard deviation (estimated measurement uncertainty) σ_{s_j}, and our measurement of the signal of the event, z_j, is y_j with a standard deviation (estimated measurement uncertainty) σ_{y_j} (see Fig. 1). That is, t_j and z_j are the true, unknown values, not the measurements. Define $D_j = (s_j, y_j)$ and $\sigma_j = (\sigma_{s_j}, \sigma_{y_j})$. The *measurement model* (or noise model) describes the probability of observing the measured values for a single event given the true values and the estimated uncertainties: it gives $P(D_j|t_j, z_j, \sigma_j)$. The σ_j are considered fixed parameters of the measurement model, and the conditioning on the measurement model is implicit (because I do not want to compare measurement models in this work).
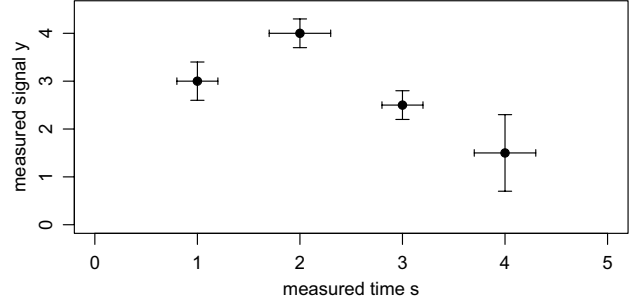
Fig. 1. Example of a measured data set with four events.

M is a stochastic *time series model* with parameters θ. It specifies $P(t_j, z_j|\theta, M)$, the probability of observing an event at time t_j with signal z_j.

The goal is (1) to compare the posterior probability of different models M, and (2) to determine the posterior PDF over the model parameters for a given M. After describing the measurement and time series models in the next two subsections, I will then show how to combine them in order to calculate the likelihood.

2.2. Measurement model

If t and z have no bounds, or can be approximated as such, and the known measurement uncertainties are standard deviations, then an appropriate choice for the measurement model is a two-dimensional (2D) Gaussian in the variables (s_j, y_j) for event j. If we assume no covariance between the variables then this reduces to the product of two 1D Gaussians

$$P(D_j|t_j, z_j, \sigma_j) = \frac{1}{\sqrt{2\pi}\sigma_{s_j}} e^{-(s_j-t_j)^2/2\sigma_{s_j}^2} \frac{1}{\sqrt{2\pi}\sigma_{y_j}} e^{-(y_j-z_j)^2/2\sigma_{y_j}^2}.$$

(1)

(The two terms are normalized with respect to s_j and y_j respectively.) If we had other information about the measurement, e.g. asymmetric error bars, strictly positive signals, or uncertainties which are not standard deviations, then we should adopt a more appropriate distribution.

2.3. Time series model

Without loss of generality, the time series model can be written as the product of two stochastic components

$$P(t_j, z_j|\theta, M) = P(z_j|t_j, \theta, M)P(t_j|\theta, M)$$

(2)

which I will refer to as the signal and time components respectively. For many processes it is appropriate to express the signal component using two independent subcomponents: the stochastic model itself and a deterministic function which defines the time-dependence of its mean. This stochastic subcomponent describes the intrinsic variability of the true signal of the physical process at a given time, with the PDF $P(z_j|t_j, \theta', M)$. I refer to this as TSMod2. An example is a Gaussian

$$P(z_j|t_j, \theta', M) = \frac{1}{\sqrt{2\pi}\omega} e^{-(z_j-\eta[t_j])^2/2\omega^2}$$

(3)

where $\theta' = (\eta, \omega)$ are the parameters of the distribution: $\eta[t_j]$ is the expected true signal at true time t_j; ω is a parameter which
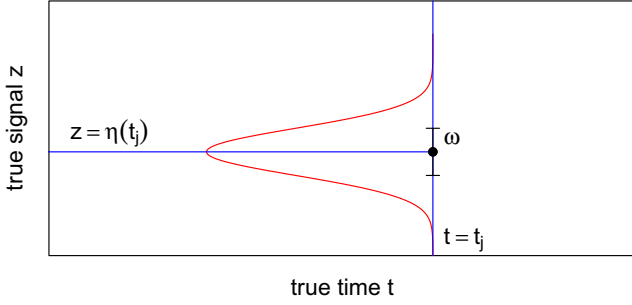
Fig. 2. Conceptual representation of the stochastic nature of the signal component of the time series model, $P(z_j|t_j, \theta', M)$ (in red) of the true signal, z, at a true time t_j (here shown as a Gaussian).

reflects the degree of stochasticity in the process. This is illustrated schematically for a single point in Fig. 2. The Gaussian is just an example, and would be inappropriate if z were a strictly non-negative quantity.

The relationship between the expected true signal and the true time is given by a deterministic function, $\eta(t; \theta_1)$, where θ_1 denotes another set of parameters. I refer to this deterministic subcomponent as TSMod1. A simple example is a single frequency sinusoid

$$\eta = \frac{a}{2}\cos[2\pi(\nu t + \phi)] + b \qquad (4)$$

which has parameters $\theta_1 = (a, \nu, \phi, b)$, the amplitude, frequency, phase, and offset. Having parametrized the signal component of the time series model in this way, it is convenient to write $\theta' = (\theta_1, \theta_2)$ in general, where $\theta_2 = \omega$ in the example of Eq. (3).

The second component of the time series model in Eq. (2) describes any intrinsic randomness in the time of the events which make up the physical process. This is represented by $P(t_j|\theta_3, M)$, which I refer to as TSMod3. If there is no variation in the probability when an event could occur, we should make this constant by using a flat distribution in t between the earliest possible and latest possible times, T_1 and T_2,

$$P(t_j|\theta_3, M) = \begin{cases} \frac{1}{T_2 - T_1} & \text{if } T_1 < t_j < T_2 \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

where $\theta_3 = (T_1, T_2)$ are its parameters. This is generally appropriate to modelling light curves (for example), where there is no concept of intrinsically discrete events: the discreteness arises only because we take measurements at certain times. There is then no sense in which the probability of an "event occuring" could vary. In contrast, when modelling a discrete process, such as the times and energies of large asteroid impacts (see CBJ11), we generally will have a time varying probability of an event occuring.

This three-subcomponent approach (TSMod1,2,3) to the time series model is conceptually a little complex, so let us consider what it means. We have a physical process in which the expected value of the signal varies with time in a deterministic manner. This is given by $\eta(t; \theta_1)$, e.g. Eq. (4). At any given true time, the true signal of the process can vary due to intrinsic randomness in the process. This is described by $P(z_j|t_j, \theta_1, \theta_2, M)$, an example of which is Eq. (3). Finally, while the mean of the process signal is considered to vary continuously in time, there may be a time varying probability that an event could occur at all (e.g. an asteroid impact). This is described by $P(t_j|\theta_3, M)$. The stochasticity in the time series model has nothing to do with measurement noise. It is intrinsic to the process.

This description of the signal component as a stochastic model with a time-independent variance and a (deterministic) time-dependence for the mean we might refer to as a *partially* stochastic process. A *fully* stochastic process, in contrast, is one in which *all* the parameters of the PDF $P(z_j|t_j, \theta, M)$ have a time-dependence, in which case this decomposition of the signal component into TSMod1 and TSMod2 is not possible. An example is the Ornstein-Uhlenbeck process, which will be used in this work. It is described in Appendix A.

The overall time series model is the combination of these three subcomponents

$$P(t_j, z_j|\theta, M) = P(z_j|t_j, \theta_1, \theta_2, M)P(t_j|\theta_3, M) \qquad (6)$$

where $\theta = (\theta_1, \theta_2, \theta_3)$. For the cases shown above, this model has seven parameters, $\theta = (a, \nu, \phi, b, \omega, T_1, T_2)$, although probably we would fix (T_1, T_2) based on inspection of the time range of the data.

Later in Sect. 3.3 we will look at the specific models and their parametrizations as used in this paper.

2.4. Likelihood

The probability of observing data D_j from time series model M with parameters θ when the uncertainties are σ_j, is $P(D_j|\sigma_j, \theta, M)$, the *event likelihood*. This is obtained by marginalizing over the true, unknown event time and signal

$$\begin{aligned} P(D_j|\sigma_j, \theta, M) &= \int_{t_j, z_j} P(D_j, t_j, z_j|\sigma_j, \theta, M) \; \mathrm{d}t_j \mathrm{d}z_j \\ &= \int_{t_j, z_j} P(D_j|t_j, z_j, \sigma_j, \theta, M)P(t_j, z_j|\sigma_j, \theta, M) \mathrm{d}t_j \mathrm{d}z_j \\ &= \int_{t_j, z_j} \underbrace{P(D_j|t_j, z_j, \sigma_j)}_{\text{Measurement model}} \underbrace{P(t_j, z_j|\theta, M)}_{\text{Time series model}} \; \mathrm{d}t_j \mathrm{d}z_j \quad (7) \end{aligned}$$

where the time series model and its parameters drop out of the first term because D_j is independent of this once conditioned on the true variables, and the measurement model (via σ_j) drops out of the the second term because it has nothing to do with the predictions of the time series model. For specific, but common, situations, this 2D integral can be approximated by a 1D integral or even a function evaluation (see Appendix B).

If we have a set of J events for which the ages and signals have been estimated independently of one another, then the probability of observing these data $D = \{D_j\}$, the *likelihood*, is

$$P(D|\sigma, \theta, M) = \prod_j P(D_j|\sigma_j, \theta, M) \qquad (8)$$

where $\sigma = \{\sigma_j\}$.

3. Model comparison

3.1. Evidence

In order to compare different models, M, we would like to know $P(M|D, \sigma)$, the model posterior probability. We can use Bayes' theorem to write this down in terms of the evidence, $P(D|\sigma, M)$. This is the probability of getting the observed data from model M, regardless of the specific values of the model parameters. Adopting equal prior probabilities, $P(M)$, for different models, the evidence is the appropriate quantity to use to compare models. One may be tempted to use instead the maximum value of the likelihood for model comparison, but this is

wrong, because it will just favour increasingly complex models, as these can increasingly overfit the data. For more discussion of this point see, for example, MacKay (2003) or CBJ11.

The *evidence*, E, is obtained by marginalizing the likelihood over the parameter prior probability distribution, $P(\theta|M)$.

$$
\begin{aligned}
E \equiv P(D|\sigma, M) &= \int_{\theta} P(D, \theta|\sigma, M)\, d\theta \\
&= \int_{\theta} \underbrace{P(D|\sigma, \theta, M)}_{\text{likelihood}}\ \underbrace{P(\theta|M)}_{\text{prior}}\, d\theta.
\end{aligned} \tag{9}
$$

Note that the evidence is conditioned on both the measurement model (via σ) and the time series model, M. For a given set of data, we calculate this evidence for the different models we wish to compare, each parametrized by θ. The parameter prior, $P(\theta|M)$, encapsulates our knowledge of the prior plausibility of different parameters. (This is independent of σ, which is why it was removed in the above equation.) As the evidence has an uninterpretable scale, we usually examine the ratio of the evidences of two models, the Bayes factor.

We evaluate the above integral using a Monte Carlo approximation

$$
E \approx \frac{1}{N} \sum_{n=1}^{n=N} P(D|\sigma, \theta_n, M) \tag{10}
$$

where the parameter samples, $\{\theta_n\}$, have been drawn from the prior, $P(\theta|M)$. Often this prior is a product of simple, 1D functions (e.g. Gaussian or gamma PDFs), so it is easy to sample from without having to employ more sophisticated methods.

3.2. Cross validation likelihood

The evidence is often sensitive to the parameter prior PDF. For example, in a single-parameter model, if the likelihood were constant over the range $0 < \theta < 1$ but zero outside this, then the evidence calculated using a prior uniform over $0 < \theta < 2$ would be half that calculated using a prior uniform over $0 < \theta < 1$. In a model with p such parameters the factor would be 2^{-p}. If we had no reason to limit the prior range, then the evidence would be of limited use in this example. Conversely, in cases where the parameters have a physical interpretation and/or where we have reasonable prior information, then we may be able to justify a reasonable choice for the prior. But in any case we should always explore the sensitivity of the evidence to "fair" changes in the prior. A fair change is one which we have no reason not to make. For example, if there were no reason to prefer a prior which is uniform over frequency rather than period, then this would be a fair change. (See CBJ11 for an illustration of this on real data.) If the evidence changes enough to alter significantly the Bayes factors when making fair changes, then the evidence is over-sensitive to the choice of prior, making it impossible to draw robust conclusions without additional information.

In such situations we might resort to one of the many "information criteria" which have been defined, such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC; e.g. Kadane & Lazar 2004) or the deviance information criterion (DIC; Spiegelhalter et al. 2002). The advantage of these is that they are simpler and quicker to calculate. But they all make (possibly unreasonable) assumptions regarding how to represent the complexity of the model, and all have been criticized in the literature.

An alternative approach is a form of K-fold cross validation (CV). We split the data set (J events) into K disjoint partitions, where $K \leq J$. Denote the data in the kth partition as D_k and its complement as D_{-k}. The idea is to calculate the likelihood of D_k using D_{-k}, without having an additional dependence on a specific choice of model parameters. That is, we want $P(D_k|D_{-k}, \sigma, M)$, which tells us how well, in model M, some of the data are predicted using the other data. Combining these likelihoods for all K partitions gives an overall measure of the fit of the model. By marginalization

$$
\begin{aligned}
P(D_k|D_{-k}, \sigma, M) &= \int_{\theta} P(D_k|D_{-k}, \sigma, \theta, M) P(\theta|D_{-k}, \sigma, M)\, d\theta \\
&= \int_{\theta} \underbrace{P(D_k|\sigma_k, \theta, M)}_{\text{likelihood}}\ \underbrace{P(\theta|D_{-k}, \sigma_{-k}, M)}_{\text{posterior}}\, d\theta
\end{aligned} \tag{11}
$$

where D_{-k} drops out of the first term because the model predictions are independent of these data when θ is specified. (σ_{-k} and σ_k drop out of the first and second terms, respectively, also for reasons of independence.) (cf. Eq. (10) of Vehtari & Lampinen 2002.) If we draw a sample $\{\theta_n\}$ of size N from the posterior $P(\theta|D_{-k}, \sigma_{-k}, M)$, then the Monte Carlo approximation of this integral is

$$
L_k \equiv P(D_k|D_{-k}, \sigma, M) \approx \frac{1}{N} \sum_{n=1}^{n=N} P(D_k|\sigma_k, \theta_n, M) \tag{12}
$$

the mean of the likelihood of the data in partition k. I will call L_k the *partition likelihood*. Note that here the posterior is sampled using the data D_{-k} only.

Because L_k is the product of event likelihoods, it scales multiplicatively with the number of events in partition k. An appropriate combination of the partition likelihoods over all partitions is therefore their product

$$
L_{\text{CV}} = \prod_{k=1}^{k=K} L_k \quad \text{or} \quad \log L_{\text{CV}} = \sum_k \log L_k \tag{13}
$$

which I call the *K-fold cross validation likelihood*, for $1 \leq K \leq J$. If $K > 1$ and $K < J$ then its value will depend on the choice of partitions. If $K = J$ there is one event per partition (a unique choice). This is *leave-one-out CV (LOO–CV)*, the likelihood for which I will denote with $L_{\text{LOO–CV}}$. If $K = 1$, we just use all of the data to calculate both the likelihood and the posterior. This is not a very correct measure of goodness-of-fit, however, because it uses all of the data both to draw the posterior samples and to calculate the likelihood.

The posterior PDF required in Eq. (11) is given by Bayes' theorem. It is sufficient to use the unnormalized posterior (as indeed we must, because the normalization term is the evidence), which is

$$
P(\theta|D_{-k}, \sigma_{-k}, M) \propto P(D_{-k}|\sigma_{-k}, \theta, M) P(\theta|M) \tag{14}
$$

i.e. the product of the likelihood and the prior. L_{CV} therefore still depends on the choice of prior (discussed in Sect. 3.3). However, the likelihood will often dominate the prior (unless the data are very indeterminate), in which case L_{CV} will be less sensitive to the prior than is the evidence.

There is a close relationship between the partition likelihood and the evidence. Whereas the evidence involves integrating the likelihood (for D) over the *prior* (Eq. (9)), the partition likelihood involves integrating the likelihood (for D_k) over the *posterior* (for D_{-k}) (Eq. (11)). This is like using D_{-k} to build a new

Table 2. Time series model components with parameters θ used in this work.

Name	Function	θ	Prior PDF, $P(\theta; \alpha)$	$\theta_{\min}, \theta_{\max}$
TSMod1				
Off	b	b	$\mathcal{N}(b; \text{mean, sd})$	$-\infty, \infty$
Sin	$\dfrac{a}{2} \cos[2\pi(\nu t + \phi)]$	a	Gamma(a; shape, scale)	$0, \infty$
		ν	Gamma(ν; shape, scale)	$0, \infty$
		ϕ	$U(\phi)$	$0, 1$
TSMod2				
Stoch	$\dfrac{1}{\sqrt{2\pi}\omega} e^{-(z-\eta[t;\theta_1])^2/2\omega^2}$	ω	Gamma(ω; shape, scale)	$0, \infty$
OUprocess	see appendix A	τ	Gamma(τ; shape, scale)	$0, \infty$
		c	Gamma(c; shape, scale)	$0, \infty$
		b	$\mathcal{N}(b; \text{mean, sd})$	$-\infty, \infty$
		$\mu[z_1]$	$\mathcal{N}(\mu[z_1]; \text{mean, sd})$	$-\infty, \infty$
		$\sqrt{V[z_1]}$	Gamma($\sqrt{V[z_1]}$; shape, scale)	$0, \infty$

Notes. The penultimate column shows the prior PDFs over these parameters, which themselves have (hyper)parameters α. These prior PDFs respect the limits on θ listed in the final column. $U()$ is the uniform distribution between 0 and 1, so has no free parameters. Note that a in the Sin model is the peak-to-peak amplitude of the sinusoid. The uniform model for TSMod3 is used throughout (Eq. (5)).

prior from "previous" data. We can use the product rule to write the partition likelihood as

$$L_k \equiv P(D_k|D_{-k}, \sigma_k, M) = \frac{P(D|\sigma, M)}{P(D_{-k}|\sigma_{-k}, M)} \tag{15}$$

which shows that it is equal to the ratio of the evidence calculated over all the data to the evidence calculated on the subset of the data used in the posterior sampling. As the same prior PDF enters into both terms, it will, in some vague sense, "cancel" out, although I stress that there is still a prior dependence.

It is important to realize that the model complexity is taken into account by the model comparison with the K-fold CV likelihood (and therefore the LOO-CV likelihood), just as it is with the Bayesian evidence. That is, more complex models are not penalized simply on account of having more parameters. It is, as usual, the prior plausibility of the model which counts.

3.3. Parameter priors

The model measures mentioned – the evidence, the K-fold CV likelihood, also the DIC – are calculated by averaging the likelihood over the model parameter space. This parameter space must therefore be sampled, and this requires that we specify a prior PDF, $P(\theta|M)$, over these.

We invariably have some information about values of the parameters, such as bounds or plausible values. For example, standard deviations, frequencies and amplitudes cannot be negative, and a phase (as defined here) must lie be between 0 and 1. Nonnegative quantities are common, and for these I adopt the gamma distribution in the applications which follow. This is characterized by two parameters, shape and scale (both positive). The mean of the gamma distribution is shape times scale, and its variance is the mean times scale.

The different components of the time series models used in the applications, along with the prior distributions over their parameters, are shown in Table 2.

We have to assign values for the (hyper)parameters, α, of the prior PDFs. Although we rarely have sufficient knowledge to specify these precisely, we can use our knowledge of the problem and the general scale of the data to assign them. I adopt the
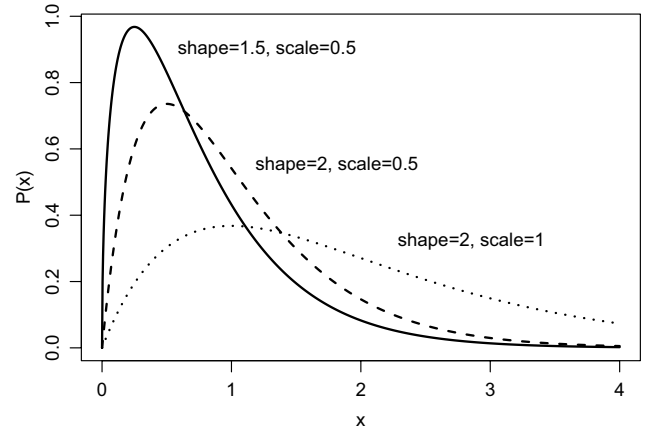

Fig. 3. Three examples of the gamma PDF, used as the prior for non-negative model parameters. The solid line, with shape = 1.5, scale = 0.5, is used as the prior PDF over frequency in units of inverse hours.

following procedure for assigning what I call the *canonical priors*, appropriate for the data which will be analysed in Sect. 6. Some parameters are set according to the standard deviation of the signal values, $\varsigma_y = \sqrt{\frac{1}{J-1} \sum_j (y_j - \overline{y_j})^2}$, where $\overline{y_j}$ is the mean signal.

– For the Off model (parameter b), I use a Gaussian with zero mean and standard deviation 1–2 times ς_y. The exact value is determined by visual inspection of the light curve.
– For the Sin model, I use a gamma prior on the frequency, ν, with shape = 1.5 and scale = 0.5 (Fig. 3). This assigns significant prior probability to a broad range of frequencies believed to be plausible based on knowledge of the problem, the temporal sampling, and the total span of the light curves. For the amplitude, a, I use a gamma prior with shape = 2 and scale 1–3 times ς_y. The prior over the phase is uniform.
– For the Stoch model (parameter ω), I use a gamma prior with shape = 2 and scale 1–2 times ς_y.
– For the OU process, I use a gamma prior on both τ and c with shape = 1.5. τ is a decay time scale, so I set its scale parameter to one quarter of the duration of the time series.

The long-term variance of the OU process is $c\tau/2$. Equating this to ς_y^2, I therefore set the scale of the diffusion coefficient, c, to be $2\varsigma_y^2/\tau$. The parameters b and $\mu[z_1]$ are both assigned Gaussian priors with a standard deviation equal to ς_y. The mean of the former is set to zero, and the mean of the latter to y_1, the signal value of the first data point. The final parameter, $\sqrt{V[z_1]}$, a standard deviation, is assigned a gamma distribution with shape = 1.5 and scale = ς_y.

This scheme of "data-based" priors was arrived at after some experimentation, and generally the calculating LOO-CV likelihoods are robust to small changes in the priors (as demonstrated later).

3.4. Markov chain Monte Carlo (MCMC)

For sampling the posterior I use the standard Metropolis algorithm with a Gaussian sampling distribution with diagonal covariance matrix. Those model parameters which do not naturally have an infinite range are transformed in order to be commensurate with Gaussian sampling: parameters with a range zero to infinity (such as frequency) are logarithmically transformed; phase (which has a range 0–1) is transformed using the logit (inverse sigmoid) function. The standard deviations of the sampling covariance matrix are set to fixed, relatively small values, typically 0.05–0.1 for the logarithmically transformed parameters (these are then scale factors). A consequence of this scaling is that the parameter can never exactly reach the extreme values (zero for the log transformation), but this is not necessarily a disadvantage. I experimented with instead using a circular transformation rather than logit for the phase parameter (by taking ϕ mod 1). While this has an affect on the posterior phase distribution, it barely changed the resulting model average likelihoods.

4. Practical application

Given a data set and a time series model we wish to evaluate, the procedure for applying the model is as follows: (1) define the (hyper)parameters of the prior parameter PDFs, as well as the standard deviations of the MCMC sampling PDF and its initial values; (2) select a partitioning of the data (normally we will use LOO-CV, so the choice is unique); (3) for each partition of the data, use MCMC to sample the posterior PDF, retaining the value of the likelihood at each parameter sample. Average these likelihoods to get the partition likelihood (Eq. (12)); (4) sum the logarithms of the partition likelihoods to get the K-fold CV log likelihood (Eq. (13)). Note that each partition provides a posterior PDF, which we could plot and summarize. In order to calculate the evidence for a model (Eq. (9)), then after step (1) we sample the prior PDF and use Eq. (10).

In the applications in this article I adopt the uniform model for the temporal component of the time series model, TSMod3 (Eq. (5)). The two parameters of this are fixed to the start and end points of the measured light curve to include all of the data. (The exact values are otherwise irrelevant, as they are the same for all models for a given light curve. This effectively removes TSMod3 from the model.) I also assume that the signal component of the measurement model is a Gaussian. I further assume that the uncertainties on the times are small compared to the time scale on which the time series model varies. This allows us to replace the 2D integration in the expression for the event likelihood (Eq. (7)) with an analytic expression, as shown in Sect. B.2. This results in significantly reduced computational times.
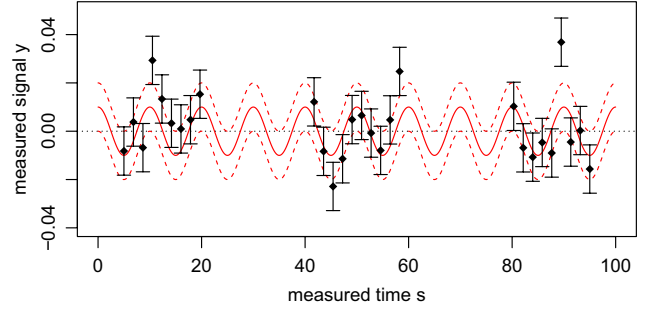
Fig. 4. Simulated sinusoidal data set. The solid red line is the true model; the dashed lines show the $\pm\sigma_{\text{true}}$ variation about this. The black points show the data set drawn from this model along with their $\pm\sigma_{\text{true}}$ error bars.

In Sect. B.2 I define the *no-model*, the model which assumes that the data are just Gaussian variations – with standard deviation given by the error bars – about the mean of the data. As this model has no parameters, its likelihood, L^{NM}, is equal to its LOO-CV likelihood and its evidence. This is therefore a convenient baseline against which to compare all other models, so in the text and tables I report the LOO-CV likelihood/evidence for all models relative to this, i.e. $\log L_{\text{LOO–CV}} - \log L^{\text{NM}}$ and $E - \log L^{\text{NM}}$ (the latter is the logarithm of the Bayes factor).

Once we have calculated these quantities for a number of models, we need to compare them. It is somewhat arbitrary how large the difference in the log likelihoods must be before we bother discussing them. Clearly very small differences are not "significant", as small changes in the priors would produce "acceptable" changes in the likelihoods. Here I identify two models as being "significantly different" if their log (base 10) likelihoods differ by more than 1. I use this term merely for the sake of identifying which differences are worth discussing.

5. Application to simulated time series

In this section I demonstrate the method by applying it to data drawn from a known model. The true model, $z(t)$, is a sinusoid (Eq. (4)) with amplitude $a = 0.02$, frequency $\nu = 0.1$, phase $\phi = 0$, and background $b = 0$. Fifty event times are drawn from a uniform random distribution between $t = 5$ and $t = 95$. There is no stochastic component. The measured signal, y, at each time is simulated by adding to z Gaussian random noise with zero mean and standard deviation $\sigma_{\text{true}} = 0.01$. In order to simulate a typical astronomical light curve – one with long gaps corresponding to day time – events lying in the range $t = 20$–40 and $t = 60$–80 were removed, leaving 28 events. (We can consider t as being in units of hours and the signal in units of magnitudes.) The true model and measured data are shown in Fig. 4. (Numerous other experiments on other light curves have also been performed.)

I apply five models to the data, constructed by combining components in Table 2.

– Off+Sin+Stoch: the single frequency sinusoidal model with offset for TSMod1 (Eq. (4)), and the Gaussian stochastic subcomponent for TSMod2 (Eq. (3)). The five adjustable parameters are $\theta = (a, \nu, \phi, b, \omega)$.
– Sin+Stoch: as Off+Sin+Stoch, but with the offset b fixed to zero (four adjustable parameters).
– Sin: as Sin+Stoch, but with the stochastic component in the signal removed ($\omega \rightarrow 0$; see Sect. B.1 for how this is achieved) (three adjustable parameters).

Table 3. Log (base 10) LOO-CV likelihood of each model relative to that for the no-model ($\log L_{\text{LOO-CV}} - \log L^{\text{NM}}$), simulated sinusoidal light curve.

$\{\sigma_{y_j}\}$	OUprocess	Off+Stoch	Sin	Sin+Stoch	Off+Sin+Stoch	No-model
0.01	–0.74	0.19	2.63	2.04	2.40	–21.95
0.005	22.01	24.16	12.67	25.93	25.04	–45.85

Notes. The last column gives the log likelihood for the no-model, $\log L^{\text{NM}}$. The first row is for models using the true value of the signal standard deviation, $\{\sigma_{y_j}\} = \sigma_{\text{true}} = 0.01$. The second row is for models using half the true value.

- Off+Stoch: a simple offset (TSMod1) with a Gaussian stochastic subcomponent for TSMod2. This model has two adjustable parameters $\theta = (b, \omega)$.
- OUprocess: the time series is modelled using the Ornstein-Uhlenbeck stochastic process, as described in Appendix A. This has five adjustable parameters $\theta = (\tau, c, \mu[z_1], V[z_1], b)$.

Model parameter priors and MCMC parameters are set following the criteria described in Sect. 3. The method is then applied – as described in Sect. 4 – to each of the five models, for two cases: (1) the standard deviations in the measured signal, $\{\sigma_{y_j}\}$, which are our estimated uncertainties, are all set equal to the true value, $\sigma_{\text{true}} = 0.01$; (2) $\{\sigma_{y_j}\} = \sigma_{\text{true}}/2 = 0.005$, i.e. the uncertainties are underestimated by a factor of two (the light curve itself is not changed). The resulting values of the LOO-CV log likelihoods are shown in Table 3.

Looking at the first row in this table, we see that all three sinusoidal models are significantly favoured (difference greater than 1) over the no-model (which has a relative LOO-CV log likelihood of zero by construction), the OU process and the Off+Stoch model. The true model, Sin, achieves the highest likelihood, although the likelihood is not significantly lower in the other two sinusoidal models. Inspection of the 1D posterior PDFs of the Sin model for the 28 partitions shows that the posterior peaks around the true frequency and amplitude in most of the partitions, although it is relatively broad over amplitude. The phase posterior PDF peaks sharply near to 0 or 1 in about half the partitions, but in the rest is broader or at intermediate values. Inspection of the posterior PDFs over the two extra parameters in the Off+Sin+Stoch model – the offset, b, and the stochastic component standard deviation, ω – shows that the mean for the offset has an average across the partitions of about 0.0025. This just reflects the fact that this particular data set has a small positive mean signal (of 0.0019). $b = 0$ generally lies within 1 standard deviation of the mean of the posterior PDF. The mean value of ω in this five parameter model is also not zero, but ranges between 0.004 and 0.007 across the 28 partitions.

Overall, we see a reasonably confident detection of the true model, by a factor of 100 in likelihood relative to the non-sinusoidal models, and this with a relatively sparse, non-uniform data set and low signal-to-noise ratio ($a/\sigma_{y_j} = 2$ for all j). Unsurprisingly it is not possible to distinguish between the different sinusoidal models. Although these provide some evidence for a non zero b and ω, it is not enough to formally favour Sin+Stoch or Off+Sin+Stoch over Sin.

In the second case (second row of Table 3) the true model is now Sin+Stoch, because the supplied signal standard deviation is now half the true value: an extra stochastic term is needed to explain the missing variance. This reduces the likelihood of the no-model. The LOO-CV likelihoods of all models are now much larger relative to this. The Sin model is poorer than the other models, because it too lacks the stochastic term needed to explain the missing variance. In contrast, Off+Stoch has a far higher likelihood: it is more important to have the stochastic component than the sinusoidal one in order to explain these data. The most favoured model is the true one, Sin+Stoch. Its posterior PDF over ω is bell-shaped with a mean of 0.01 and a standard deviation of 0.002 in essentially all partitions. Given that the error bars supplied with the data were 0.05 and not zero, we might expect that a value of ω less than 0.01 would be sufficient to explain the variance on top of the sinusoid. A larger value is needed, however, because this data set just happens to have a larger actual variance than explained by σ_{true}, as we also saw in the first case.

In this example we have seen that the models lacking the stochastic component have likelihoods which are very sensitive to the estimated signal standard deviations. As these are usually hard to estimate accurately, we should generally use a model with a stochastic component (TSMod2) with a free parameter in order to accommodate such missing variance. Comparing results from this with those from a model without such a component will help us establish whether the additional variance is required.

6. Application to astronomical light curves

6.1. Background and data

I now apply the method to a set of (sub)stellar light curves. Each light curve shows the variation over time of the total light received (in the I band) from a very low mass star or brown dwarf, objects collectively referred to as ultra cool dwarfs (UCDs). The variability of these sources has been the subject of numerous studies, because the light curves may reveal something about the processes operating in these objects' atmospheres (e.g. Morales-Calderón et al. 2006; Bailer-Jones 2008; Radigan et al. 2012). Variability could plausibly occur on time scales of hours to tens of hours due to the evolution of surface features, which might be star spots or dust clouds. If the opacity or brightness of these surface features differs from the rest of the (sub)stellar photosphere, a change in the coverage of the features with time would modulate the integrated light received by the observer (the stars are not spatially resolved). Another plausible cause of variability is the star's rotation when it has inhomogeneous (but otherwise stable) surface features. (The rotation periods of these objects have been measured to be between a few hours and a few days, e.g. Bailer-Jones 2004; Reiners & Basri 2008.) In general, both mechanisms could generate observable variability.

Here I use a set of 11 light curves previously obtained and analysed by Bailer-Jones & Mundt (2001; hereafter BJM)[1]. The data are shown in Fig. 5. In BJM, the light curves were subject to a simple orthodox hypothesis test using the χ^2 statistic. The null

---

[1] Tables 1 and 2 of BJM lists the properties of the objects. There are two light curves obtained at different times for the object 2M1145. The label 2m1145a is used in the present work to indicate the shorter duration light curve in Table 2 of BJM, i.e. the one with $t_{\text{max}} = 76$ h. 2m1145b labels the longer duration one.
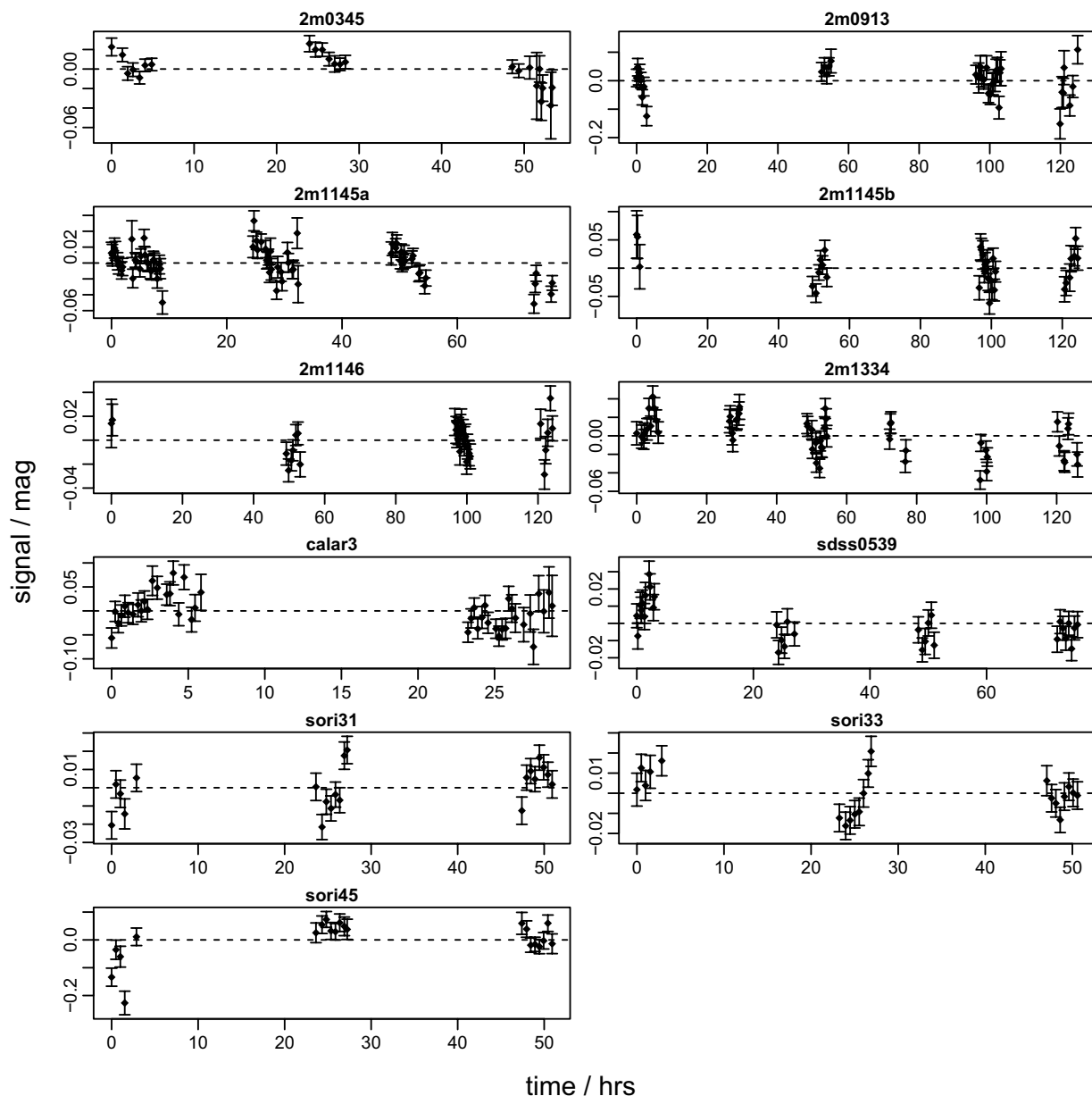
Fig. 5. Ultra cool dwarf light curves, showing the variation in brightness over time. Note that increases in brightness are *downwards* in the y axis (more negative magnitudes). 2m1145a and 2m1145b are light curves of the same object, but observed more than a year apart.

hypothesis was that the light curve was constant, with fluctuations due only to heteroscedastic Gaussian noise, with zero mean and standard deviations estimated in the data reduction process. (This is equivalent to the no-model in the present paper.) The p-value of the χ^2 statistic was calculated, and if less than 0.01 the object was declared as being "variable"[2]. Of the 22 light curves analysed in BJM, 12 were declared as variable in this way, of which 11 are analysed here. (The other 11 light curves are no longer available unfortunately.)

Although this statistical test is widely used in this and other contexts, it is vulnerable to some of the standard – and valid – criticisms of orthodox hypothesis testing (see CBJ11 and

references therein for further discussion). These are: the p-value is defined in terms of the probability of the χ^2 being as large as or larger than the value observed, i.e. it is defined in terms of unobserved and therefore irrelevant data; the p-value does not measure the probability of the hypothesis given the data, so does not answer the right question; a low p-value is used to reject the null hypothesis and therefore accept the (implicit) alternative, but without ever actually testing any alternative, even though the alternative may explain the data even less well. This final point is important, because all but the most trivial models generally give a very low probability for any particular data set, so a low p-value per se tells you little. What is important is the relative likelihoods of different models. At best, a small p-value is just an indication that the null hypothesis may not be adequate to explain the data, but it is not a substitute for proper model assessment, i.e. model comparison. The onus is then on us to define alternative models and compare them in a suitable way, which is what I do here. I use the same five models as were used in Sect. 5

---

[2] BJM then go on to look for significant peaks in the CLEAN periodogram, and identify the variability as non-periodic if there is no significant peak in the periodogram. This suggests that the variability is not due to a simple rotational modulation, according to what was later called the "masking hypothesis": the rapid evolution of surface features "erases" the rotational modulation signature from the periodogram.

Table 4. Log (base 10) LOO-CV likelihood of each model relative to that for the no-model for each light curve ($\log L_{\text{LOO-CV}} - \log L^{\text{NM}}$).

Light curve	OUprocess	Off+Stoch	Sin	Sin+Stoch	Off+Sin+Stoch	No-model	p-value
2m0345	3.26	2.07	0.15	2.06	2.66	–13.60	4e-4
2m0913	0.44	0.72	0.23	0.97	0.10	–53.39	7e-4
2m1145a	15.23	8.59	3.01	12.26	11.70	–63.83	<1e-9
2m1145b	–0.73	1.96	2.00	2.69	2.95	–39.71	1e-3
2m1146	0.67	0.56	–0.08	0.21	1.17	–26.83	3e-3
2m1334	14.95	12.82	4.06	16.86	16.12	–65.88	1e-9
sdss0539	5.50	1.99	4.93	4.48	4.67	–19.62	3e-5
calar3	3.60	1.43	5.65	5.11	4.28	–28.06	6e-4
sori31	2.04	2.12	1.02	2.59	1.90	–11.16	4e-5
sori33	1.49	0.66	2.14	1.85	2.12	–8.39	2e-3
sori45	6.70	4.32	5.08	6.23	6.32	–29.93	5e-9

Notes. The penultimate column gives the value of the log likelihood for the no-model, $\log L^{\text{NM}}$. The last column is the p-value for the hypothesis test from BJM.

Although the measured data points have negligible timing uncertainties, they do have a finite duration (the integration time of the observations), either 5 or 8 min (fixed for a given light curve). This could be accommodated into the measurement model (Sect. 2.2) by using a top-hat distribution instead of a Gaussian. I nonetheless approximate this as a delta function, for two reasons: (1) it accelerates considerably the likelihood calculations, because it allows us to replace the 2D likelihood integral (Eq. (7)) with an analytic calculation (Sect. B.2); (2) the method of calculating the likelihood of the OU process (derived in Sect. A.2) is only defined for this limit.

6.2. Results: LOO-CV likelihood

I follow the procedure outlined in Sect. 4 to define the priors and to sample the posterior with MCMC. The results are summarized in Table 4. A first glance over the table shows that for ten of the light curves, most of the models are significantly better than the no-model at explaining the data, often by a large amount.

According to the χ^2 test of BJM, all of these objects have a variability which is inconsistent with Gaussian noise on the scale of the error bars, so there should be a better model than the no-model (although it may not be among those I have tested). We see from the Table that the no-model is not favoured for any light curve. However, for 2m0913, none of the models is *significantly* more likely than the no-model, so there is no reason to "reject" it. As the no-model is equivalent to the null hypothesis of BJM's χ^2 test, and this gave a p-value of $7e^{-4}$, this shows that the p-value is not a reliable metric for "rejecting" the null hypothesis.

On the other hand, in the three cases where the p-value is very low – 2m1145a, 2m1334, sori45 – the relative log likelihood for at least one model is high. This suggests that a very low p-value *sometimes* correctly indicates that another model explains the data better, although this is of limited use as we do not know *how* low the p-value has to be. But at least it might motivate us to define and test other models. The converse is not true: a relatively high p-value does not indicate that the null hypothesis is the best fitting model.

We turn now to identifying the best models. For all light curves, there is no significant difference between Off+Sin+Stoch and Sin+Stoch, which just means that the offset is not needed. That is not surprising, because the light curves have zero mean by construction. For eight of the light curves, the LOO-CV likelihood for Sin+Stoch is significantly larger than for Sin, implying there is a source of (Gaussian) stochastic variability which is not accounted for by the error bars in the data, $\{\sigma_{y_j}\}$. This

indicates either an additional source of variance (variability), or that the error bars have been underestimated. (In only two of these cases – 2m1145a and 2m1334 – are the differences between Sin and Sin+Stoch very large.)

Of course, there is no reason a priori to assume that a sinusoidal model is the appropriate one. In 9 of the 11 light curves, the sinusoidal models give a higher likelihood than the Off+Stoch model, and in the other two cases the value is not significantly lower. We can therefore state that for none of the 11 light curves is Off+Stoch significantly better than the sinusoidal models. But only with five or six light curves can we say that a sinusoidal model is *significantly* better than Off+Stoch. For the remaining light curves, the data (and priors!) do not discriminate sufficiently between the models, so neither can be "rejected".

Turning now to the OU process, we see that this is significantly better than all other models only for 2m1145a, but by a confident margin. In seven other cases the OU process is still better than the other models, or at least not significantly worse than the best model, so cannot be discounted as an explanation. In the remaining three cases – 2m1145b, 2m1334, calar3 – at least one other model is significantly better than the OU process.

The results for 2m1145a and 2m1145b are interesting, as these are light curves of the same object observed a year apart. At one time the OU process is the best explanation, at the other either a sinusoidal model or Off+Stoch. Although it is plausible that the object shows different behaviour at different times, e.g. according to the degree of cloud coverage, we should not over-interpret this. We should also not forget that another, untested model could be better than any of these.

To summarize: based just on the LOO-CV likelihood, I conclude that 10 of 11 light curves are explained much better by some model other than the no-model, by a factor of 100 or more in likelihood. The exception is 2m0913, for which all models are equally plausible (likelihoods within a factor of ten). Three light curves can be associated with one particular model: 2m1145a is best described by the OU process; 2m1334 and calar3 are best described by a sinusoidal model, the former requiring an additional stochastic component (Sin+Stoch), the latter could be either with or without it (Sin). This would seem to be consistent with a rotational modulation of the light curve (but see the next section). For the remaining seven light curves, no single model emerges as the clear winner, although some models are significantly disfavoured. In three of these seven cases – 2m0345, sdss0539, sori45 – both the OU process and a sinusoid model explain the data equally well (for 2m0345 and sori45
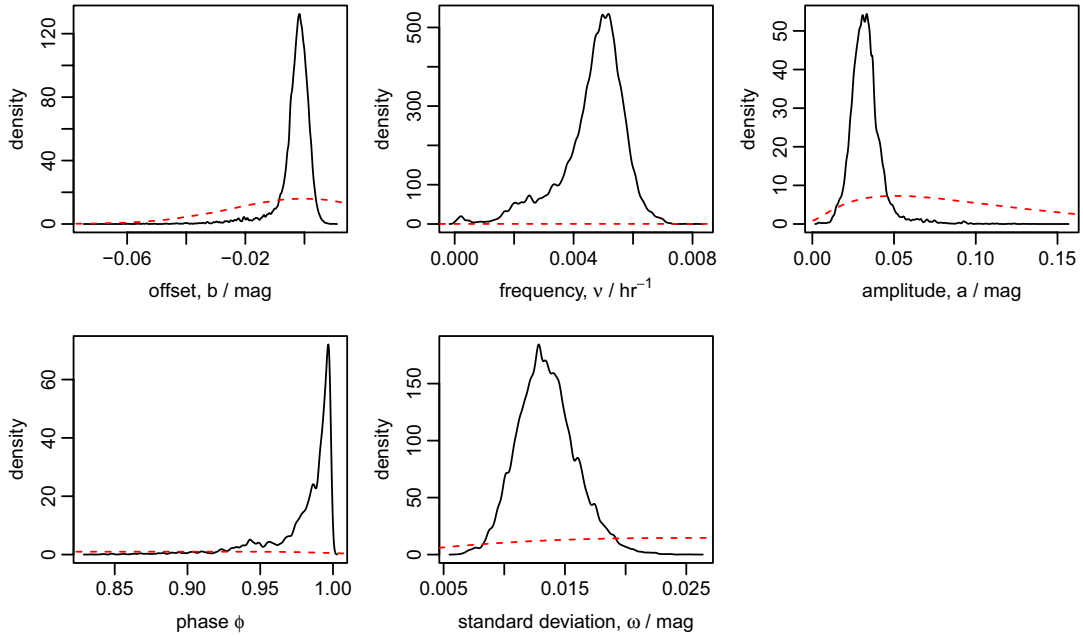
Fig. 6. Posterior PDF (black solid line) and prior PDF (red dashed line) over the five parameters in the Off+Sin+Stoch model of 2m1334, for one of the data partitions. The posterior PDFs for most of the other partitions look similar.

the sinusoidal model needs a stochastic component). For the remaining four light curves – 2m1145b, 2m1146, sori31, sori33 – the Off+Stoch model is at least as plausible as the other models. This model describes the data as having a larger Gaussian variance than is described by the error bars (with a possible constant offset to the light curve in addition). This could betray a variance intrinsic to the UCD, but it could equally well indicate that the error bars have been underestimated, something which is quite plausible given the multiple stages of the data reduction and approximations therein.

6.3. Results: posterior PDFs

To calculate the LOO-CV likelihood for a light curve with J events, we had to sample from J different posterior PDFs – one per partition – each given by Eq. (14). Here I examine the posteriors for the three light curves which could be associated with one particular model: 2m1334, calar3, and 2m1145a.

Figure 6 shows the 1D projections of the 5D posterior PDF for the Off+Sin+Stoch model on the 2m1334 light curve. The most probable model was Sin+Stoch, and the PDFs over the parameters this model has in common with Off+Sin+Stoch are similar. In the first panel we see that the offset is consistent with being zero, as expected. Most of the probability for the frequency (second panel) lies between 0.004 and $0.006\,h^{-1}$, or periods of 170–250 h. This is considerably larger than the periods $6.3 \pm 0.4\,h$ ($\nu = 0.16\,h^{-1}$) and $1.01 \pm 0.08\,h$ identified by BJM (at a signal-to-noise ratio of 6 and 7 respectively). $170-250\,h$ is also relatively long for a rotation period for a UCD and longer than the duration of the light curve. Thus the models to which this frequency range corresponds are in fact not periodic (no complete cycle) but just long term trends. A visual inspection of the light curve supports this. This could be intrinsic to the UCD or could be a slow drift of the zero point of the photometric calibration.

A more detailed study could overcome this by introducing an explicit trend model which is distinct from a periodic model. The prior PDF over frequency of the sinusoidal model would

then be truncated at low values to ensure that such a model is truly periodic. This was done in CBJ11, where the evidence was calculated by averaging the likelihoods over a limited range of the period parameter.

Examining further Fig. 6, we see strong evidence for a nonzero value of ω (the prior permits much smaller values), indicating that we need a stochastic component to explain variance on top of the (low frequency) sinusoidal component. Note finally that the posterior PDFs are far narrower than the corresponding prior PDFs (plotted as red dashed lines), which are essentially flat for three of the parameters (cf. Fig. 3). This suggests that the priors are relatively uninformative, so the results are not very sensitive to their exact choice.

The other light curve for which a sinusoidal model is significantly favoured is calar3. The posterior PDFs of the three parameters of the Sin model are shown in Fig. 7 for two partitions. The PDFs are very similar for these and all other partitions. We see two distinct peaks in frequency, at around $0.075\,h^{-1}$ and $0.12\,h^{-1}$, or periods of 13.3 h and 8.3 h. Both are plausible rotation periods. In comparison, BJM found peaks in the CLEAN periodogram at 14.0 h and 8.5 h, although at less than five times the noise they were described as "barely significant". These nonetheless agree with the periods found in the present analysis to within the uncertainties. As the integrated probability in these peaks in Fig. 7 is very large, then if Sin is the correct model (and not just the most probable of those tested) then these periods are significant. The double-peaked posterior PDF means that there is evidence supporting models at both periods (or one is an alias), but Sin is still a single component model. In particular, the posterior PDF over amplitude in Fig. 7 is the distribution over amplitudes for all models, i.e. for the entire frequency range. In order to determine the best fitting amplitudes for a model with two sinusoidal components we would need to calculate the posterior PDF for a six parameter model.

Of the 11 light curves explored here, BJM identified significant periods for 2m1146, 2m1334, sdss0539, and sori31. The present analysis suggests all of these *could* be explained by a periodic model, but only for 2m1334 is the periodic model
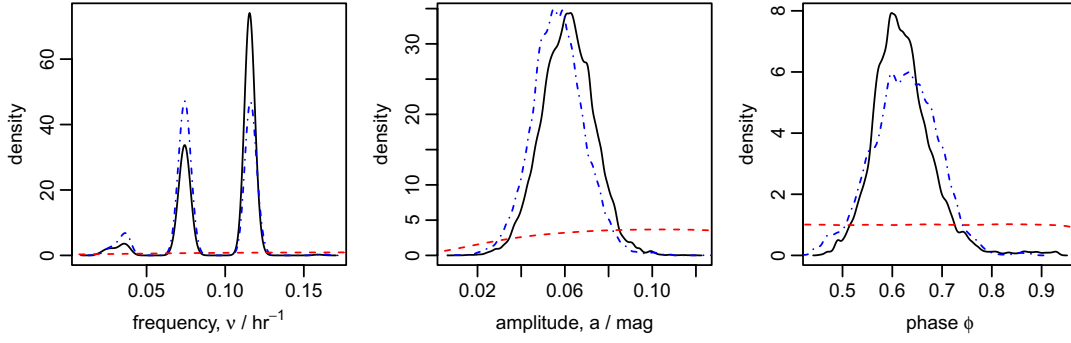
Fig. 7. Posterior PDF (black/solid line for one partition, blue/dot-dashed for another) and prior PDF (red dashed line) over the three parameters of the Sin model of calar3. The posterior PDFs for the other partitions are likewise very similar.
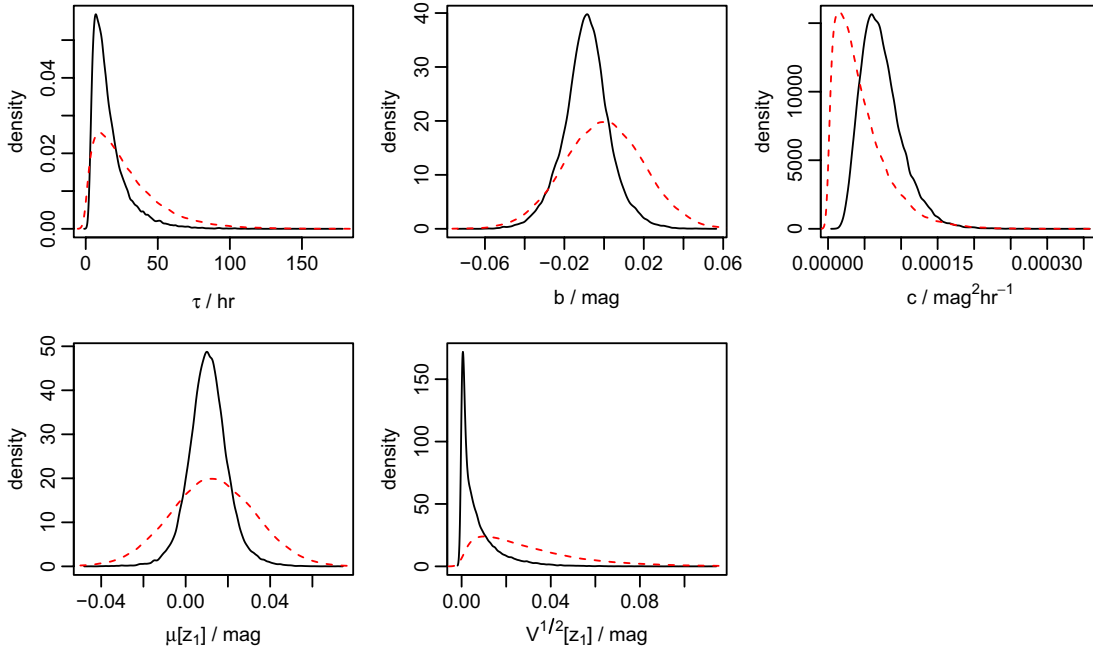

Fig. 8. Posterior PDF (black solid line) and prior PDF (red dashed line) over the five parameters in the Off+Sin+Stoch model of 2m1145a, for one of the partitions. The posterior PDFs for the other partitions look virtually identical.

significantly better than the others (although we just saw that this "period" is actually a trend). Furthermore, for 2m1146 only the Off+Sin+Stoch model is more significant than the no-model, and then only barely (LOO-CV log likelihood difference of just 1.17). So there is no strong evidence supporting periodicity in these four light curves.

The only other light curve with a clear "winner" model in terms of the LOO-CV likelihood is 2m1145a, for which the OU process was identified. The PDFs over one partition are shown in Fig. 8. Noticeable here is that the posterior PDFs are not much narrower than the prior PDFs. In Bayesian terms, the Occam factor (the ratio of volume of the parameter space occupied by the posterior to that occupied by the prior) is not much smaller than one, which means the data are not providing a strong discrimination over parameter solutions. This can be interpreted to mean that the model is quite flexible: a wide range of parameter settings are able to explain the data. This is perhaps not surprising given the nature of the OU process and the low signal-to-noise ratio of the data (the standard deviation of the signal, ς_y, is only 1.7 times the mean error bar, $\overline{\sigma_{y_j}}$). Alternatively, we can interpret this similarity between prior and posterior to mean that we used a comparatively informative (narrow) prior – see next section.

6.4. Results: sensitivity of the LOO-CV likelihood to the prior

An important issue discussed in Sect. 3.2 is the sensitivity of results to the priors adopted. I investigate this here by increasing the scale over which the prior PDFs extend. Specifically, I multiply by 2, then by 4, the standard deviation of Gaussian priors and the scale parameter of beta priors, then re-run the MCMC to recalculate the LOO-CV likelihood. The results of doing this for the sdss0539 light curve are shown in Table 5. Although the likelihoods do change, they do not change by more than the factor of ten adopted here to indicate a significant difference. This pattern is generally seen with the other light curves also. However, there are a few cases in which we can get larger changes in the likelihood for apparently innocuous changes in the priors. This requires further study.

6.5. Results: Bayesian evidence

I have introduced the LOO-CV likelihood as an alternative to the Bayesian evidence, on the basis that it is less sensitive to changes in the prior. I investigate this here.

As the no-model has no adjustable parameters, its evidence is equal to its likelihood and its LOO-CV likelihood. I therefore

Table 5. Log (base 10) LOO-CV likelihood of each model relative to that for the no-model ($\log L_{\text{LOO-CV}} - \log L^{\text{NM}}$), for the light curve sdss0539.

Priors	OUprocess	Off+Stoch	Sin	Sin+Stoch	Off+Sin+Stoch
canonical	5.50	1.99	4.93	4.48	4.67
×2 scale	5.14	2.02	4.51	4.11	4.13
×4 scale	4.60	1.98	5.08	4.51	4.83

Notes. Each line corresponds to different settings of the parameters of the prior PDFs.

Table 6. Log (base 10) evidence of each model relative to that for the no-model, $\log E - \log L^{\text{NM}}$, for the light curve sdss0539.

Priors	OUprocess	Off+Stoch	Sin	Sin+Stoch	Off+Sin+Stoch
canonical	5.38	1.07	3.24	2.31	2.03
×2 scale	4.72	0.30	2.42	0.65	–0.90
×4 scale	3.97	–0.53	1.81	–0.87	–2.63

Notes. Each line corresponds to different settings of the parameters of the prior PDFs (cf. Table 5).

use this again as a baseline against which to report the evidence. This relative evidence is reported in Table 6 for sdss0539, for the same models and priors as shown in Table 5. Comparing the first lines between these tables (canonical priors), we arrive at a similar conclusion based on the evidence as based on the LOO-CV likelihood: the OU process is the single best model; Off+Stoch and the no-model are significantly less likely. However, based on the evidence we would now say that the sinusoidal models are also significantly less likely. Examining how the evidence changes with the priors, we see significant (more than unity) changes for all models, something we do not see for the LOO-CV likelihood in Table 5. The evidence is indeed more sensitive to the priors in this case. We observe a similar behaviour for other light curves (although there are cases where doubling the scale of the priors changes the evidence by less than a factor of ten). Prior sensitivity nonetheless remains an issue for LOO-CV likelihood, and this should always be investigated in any practical application.

7. Summary and conclusions

This article has introduced three ideas

1. a fully probabilistic method for modelling time series with arbitrary temporal spacing. It can accommodate any kind of measurement model (error bars) on both the time and signal variables, as well as any functional model for the signal-time dependence and the stochastic variations in both of these. In contrast to many other time series modelling methods, it can model a stochastic variation in the time axis too. It can in fact be used to model any 2D data set, and not just temporal data: with a linear model it offers an alternative solution to the total least squares solution for data sets with errors in both variables, for example.

2. a cross-validation alternative to the Bayesian evidence, which is based on the posterior-averaged likelihood (combined over partitions of the data) as opposed to the prior-averaged likelihood. In theory this is less sensitive to the prior parameter PDFs than the evidence, something confirmed by the initial experiments reported here. Experiments

on simulated data suggest that this metric is an effective means of model comparison. Its main drawback in comparison to the evidence is that it takes longer to calculate.

3. the use of the Ornstein-Uhlenbeck process in a Bayesian time series model, i.e. one in which we sample rather than maximize the posterior. (Theoretical results similar to the event likelihood and the recurrence relation for the posterior PDF – derived in Appendix A – have been published elsewhere.)

The main purpose of this article was to give a detailed theoretical exposition of the model. A more comprehensive application to time series analysis problems will be published elsewhere. In the present work I have demonstrated the method using simulated data, and through an analysis of 11 brown dwarf light curves. The main conclusions of this study are as follows

- 10 of 11 light curves are explained significantly better by one of the models tested than by the no-model, the "null hypothesis" that the variability is just due to Gaussian fluctuations with standard deviation given by the error bars about the mean of the data. "Significantly better" here means the LOO-CV likelihood is at least 100 times larger, something we might interpret as a 99% confidence level. For comparison, in BJM all 11 light curves were flagged as variable at a (different) 99% confidence level based on an orthodox χ^2 test of the same null hypothesis. The Bayesian model comparison performed here has a sounder theoretical basis, gives us more confidence in the results, and supplies more information. The two methods disagree on 2m0913, which is adequately explained by the no-model here.

- three light curves are described significantly better by one model than any of the others: 2m1145a by the OU process; 2m1334 by a sinusoid with an additional Gaussian stochastic component; calar3 by a sinusoid either with or without an additional stochastic component. However, the probable periods for 2m1334 are longer than the duration of the time series, so this is best interpreted as a long-term trend rather than a periodic variation. For calar3 we see two distinct and significant peaks in the posterior PDF over the frequency, at $0.12\,\text{h}^{-1}$ (period = 8.3 h) and $0.075\,\text{h}^{-1}$ (period = 13.3 h). Both had been identified in earlier work, but as barely statistically significant.

- the other 8 light curves can be described by more than one model, either the OU process, or a constant with stochastic component or a sinusoid with stochastic component.

It must be remembered that we can only comment on models we have explicitly tested: it remains possible that other plausible models exist which could explain the data better. Future work with this method will focus on its practical application to scientific problems, the inclusion of more time series models, as well as further testing of the LOO-CV sensitivity to the prior PDFs.

Appendix A: Fully stochastic time series processes

The signal component of the time series model is the PDF $P(z_j|t_j, \theta, M)$. For a physical process which has a well-defined, time-variable signal on top of which there is some randomness, Sect. 2.3 shows a convenient way of expressing this as

two independent subcomponents: one which describes the time-dependence of the mean of the PDF and the other which describes the PDF itself and its time-independent parameters, e.g. its variance.

A fully stochastic process, in contrast, is one in which all of the parameters of the PDF can have a time dependence. Given a functional form for this time dependence, we can in principle just introduce this into θ and calculate the likelihood as before. A simple fully stochastic process is one with a constant mean and variance, a white noise process. This is achieved by setting TSMod1 to a uniform model ($\eta = b$ in Eq. (4)) with TSMod2 a Gaussian.

However, incorporating a stochastic process which has memory, such as a Markov process, is more complicated. Here I show how to introduce a particular but widely used stochastic process, the Ornstein-Uhlenbeck process.

A.1. The Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck (OU) process (Uhlenbeck & Ornstein 1930) is a stochastic process which describes the evolution of a scalar random variable, z. The equation of motion (Langevin equation) for $t > 0$ can be written

$$dz(t) = -\frac{1}{\tau}z(t)dt + c^{1/2}\mathcal{N}(t; 0, dt) \tag{A.1}$$

where τ and c are positive constants, the *relaxation time* (dimension t) and the *diffusion constant* (dimension $z^2 t^{-1}$) respectively, dt is an infinitesimally short time interval, $\mathcal{N}(t; 0, dt)$ is a Gaussian random variable with zero mean and variance dt, and $dz(t) = z(t + dt) - z(t)$. The OU process is the continuous-time analogue of the discrete-time AR(1) (autoregressive) process, and is sometimes referred to as the CAR(1) process. In the context of Brownian motion, $z(t)$ describes the velocity of the particle. There are alternative, equivalent forms of this equation of motion. For more details see Gillespie (1996a).

The OU process is stationary, Gaussian and Markov[3]. The PDF of $z(t)$ is Gaussian with mean and variance given by

$$\mu_z = z_0 \upsilon \tag{A.2a}$$
$$V_z = \frac{c\tau}{2}(1 - \upsilon^2) \tag{A.2b}$$

respectively, for any $t > t_0$, where $z_0 = z(t = t_0)$ and

$$\upsilon = e^{-(t-t_0)/\tau}. \tag{A.3}$$

Given the initial condition z_0 at t_0, we know the PDF of the process at any subsequent time. The relaxation time, τ, determines the time scale over which the mean and variance change. The diffusion constant determines the amplitude of the variance. The OU process $z(t)$ is a mean-reverting process: for $t - t_0 \gg \tau$ the mean tends towards zero and the variance asymptotes to $c\tau/2$ (for finite τ). From this we can derive an update equation to give the value of the process at time t,

$$z(t) = z_0 \upsilon + n_1 \sqrt{V_z} \tag{A.4}$$

---

[3] Put loosely: Stationary means that the joint PDF of a set of events from the process is invariant under translations in time; Markov means that the present value of the process depends only on the value at one previous time step; Gaussian means that the joint PDF of any set of points is a multivariate Gaussian, in particular the PDF of a single point is Gaussian.

where n_1 is unit random Gaussian variable (Gillespie 1996a). This is just the sum of the mean and a random number drawn from a zero-mean Gaussian with the variance at time t. For a given sequence of time steps, $(t_0, t_1, \ldots)$, we can use this to generate an OU process. Because the time series is stochastic and must be calculated at discrete steps, then even for a fixed random number seed, the generated time series depends on the actual sequence of steps.

The reader may be more familiar with the Wiener process. This can be considered a special case of the OU process in which $\tau \to \infty$ (Gillespie 1996b), in which case $\upsilon \to 1$. The update equation becomes $z(t) = z_0 + n_1 \sqrt{V_z}$, where now $V_z = c(t - t_0)$.

A.2. Likelihood of the Ornstein-Uhlenbeck process

A Markov process is one in which we can specify the PDF of the state variable, z_j at time t_j, using $P(z_j|t_j, z_{j-1}, t_{j-1}, \theta, M)$, i.e. there is a dependence on the previous state variable, z_{j-1}. For the OU process, this PDF is a Gaussian with mean and variance given by Eq. (A.2). Clearly, the nearer t_{j-1} is to t_j the better a measurement of z_{j-1} will constrain z_j.

We could therefore write the signal component of the time series model (see Eq. (2)) as

$$P(z_j|t_j) = \int_{t_{j-1}, z_{j-1}} P(z_j|t_j, z_{j-1}, t_{j-1})P(z_{j-1}, t_{j-1}|t_j)\, dt_{j-1}dz_{j-1}$$
$$= \int_{t_{j-1}, z_{j-1}} P(z_j|t_j, z_{j-1}, t_{j-1})P(z_{j-1}|t_{j-1})P(t_{j-1})\, dt_{j-1}dz_{j-1} \tag{A.5}$$

where conditional independence has been applied in the second line to remove the t_j dependence from the second two terms. Note that everything is implicitly conditioned on M and its parameters θ, but these have been omitted for brevity. The first term under the integral is the PDF for the Markov process we aimed to introduce. The second term is also a PDF for the Markov process but referred to the previous event. We could replace that with another 2D integral over (t_{j-2}, z_{j-2}) of exactly the same form as Eq. (A.5). We could then continue recursively to achieve a chain of nested 2D integrals going back to the beginning of the time series, and use that in our likelihood calculation. Although this is a plausible and general solution for a Markov process, it is not very appealing.

Fortunately a significant simplification is possible. Let us first neglect the time uncertainties. In that case the event likelihood (Eq. (7)) becomes

$$P(D_j|\sigma_j, \theta, M) = \int_{z_j} P(y_j|z_j, \sigma_{y_j})P(z_j|t_j, \theta, M)P(t_j|\theta, M)\, dz_j \tag{A.6}$$

with $t_j = s_j$. $P(y_j|z_j, \sigma_{y_j})$ is the signal part of the measurement model (the time part has dropped out). If this is Gaussian in $y_j - z_j$ (cf. Eq. (1)) and $P(z_j|t_j, \theta, M)$ is Gaussian in z_j, then Eq. (A.6) is a convolution of two Gaussians, which is another Gaussian, multiplied by $P(t_j|\theta, M)$ (which is independent of z_j). A general result is that if f is a Gaussian with mean μ_f and variance V_f, and g is a Gaussian with mean μ_g and variance V_g then

$$\int_{-\infty}^{+\infty} f(y - z)g(z)dz = f \otimes g \tag{A.7}$$

is a Gaussian with mean $\mu_f + \mu_g$ and variance $V_f + V_g$. For the Gaussian measurement model, f is Gaussian in the argument $y_j - z_j$ with $\mu_f = 0$ and $V_f = \sigma_{y_j}^2$. g is then the time series model.

We now turn specifically to the OU process in order to determine its time series model, $P(z_j|t_j, \theta, M)$. We can derive this from Eq. (A.4). With the state variable now written as z_j rather than $z(t)$, this update equation is

$$z_j = z_{j-1}v + n_1\sqrt{V_z} \tag{A.8}$$

with

$$v = e^{-(t_j - t_{j-1})/\tau} \tag{A.9a}$$

$$V_z = \frac{c\tau}{2}\left(1 - v^2\right). \tag{A.9b}$$

z_j has a Gaussian distribution (by definition of the OU process) with mean and variance[4]

$$\mu[z_j] = \mu[z_{j-1}]v \tag{A.10a}$$

$$V[z_j] = V[z_{j-1}]v^2 + V_z \tag{A.10b}$$

respectively (see also Berliner 1996). Specifically, $P(z_j|t_j, \theta, M)$ is a Gaussian with this mean and variance, which are specified by the parameters $\theta = (\mu[z_{j-1}], V[z_{j-1}], v, \tau, c)$[5]. We will look in a moment at how we estimate $\mu[z_{j-1}]$ and $V[z_{j-1}]$.

We can now write the likelihood, the result of the Gaussian convolution, Eqs. (A.6) and (A.7), as

$$P(D_j|\sigma_j, \theta, M) = P(t_j|\theta, M) \int_{z_j} P(y_j|z_j, \sigma_{y_j})P(z_j|t_j, \theta, M)\,dz_j$$

$$= P(t_j|\theta, M)\frac{1}{\sqrt{2\pi V[y_j]}}\exp\left(\frac{-(y_j - \mu[y_j])^2}{2V[y_j]}\right) \tag{A.11}$$

where the mean and variance of this Gaussian are

$$\mu[y_j] = 0 \quad + \mu[z_j] \tag{A.12a}$$

$$V[y_j] = \sigma_{y_j}^2 + V[z_j] \tag{A.12b}$$

respectively. Recall that $P(t_j|\theta, M)$ is just the time component of the time series model with $t_j = s_j$. Normally we will use a uniform model (Eq. (5)), so this is just a constant.

To estimate $\mu[z_{j-1}]$ and $V[z_{j-1}]$ we make use of the data, y_{j-1}. For an event t_j, $P(z_j|t_j, \theta, M)$ – which has mean and variance given by Eq. (A.10) – is our estimate of the PDF of the state variable at t_j prior to taking into account the measurement y_j. It is therefore the appropriate thing to use to calculate the likelihood of y_j, as we have done in Eq. (A.11). But in parallel to this we want to use y_j to improve our estimate of z_j, i.e. we want to calculate the posterior PDF of z_j. This is given by Bayes' theorem

$$P(z_j|y_j, t_j) \propto P(y_j|z_j, t_j)P(z_j|t_j) \tag{A.13}$$

(ignoring the normalization constant $1/P(y_j|t_j)$, and omitting a lot of dependencies). These two terms are again the measurement model (so the dependence on t_j drops out) and the time

---

[4] These we calculate explicitly from Eq. (A.8). The variance is just the sum of the variances of the two terms in that equation. Recall that in general $V(fg) = f^2V(g) + g^2V(f)$, and that $V(v) = 0$.

[5] If we instead had an actual value of z_{j-1}, then $P(z_j|t_j, \theta, M)$ would be Gaussian with mean $z_{j-1}v$, variance V_z, and $\theta = (z_{j-1}, v, \tau, c)$.

series model, both of which are Gaussian in z_j. Thus the posterior PDF over z_j is also a Gaussian with mean and variance[6]

$$\mu'[z_j] = \frac{y_j V[z_j] + \mu[z_j]\sigma_{y_j}^2}{V[z_j] + \sigma_{y_j}^2} \tag{A.15a}$$

$$V'[z_j] = \frac{V[z_j]\sigma_{y_j}^2}{V[z_j] + \sigma_{y_j}^2} \tag{A.15b}$$

respectively, where the prime symbol is used to distinguish these posterior moments from the prior ones in Eq. (A.10). It is these quantities which we then use at the *next* event as the estimates of the mean and variance of the state variable. Thus, at iteration (event) j, when we calculate Eq. (A.12) and hence the likelihood, we use $\mu'[z_{j-1}]$ and $V'[z_{j-1}]$ as our estimates of $\mu[z_{j-1}]$ and $V[z_{j-1}]$. This is how we introduce a dependence on the previous measurement (the Markov property). We then calculate the mean and variance of the posterior for z_j using Eq. (A.15), and will then use these in the next iteration. Thus we have a recurrence relation for the posterior PDF of z_j, at each iteration siphoning off the relevant quantities in order to calculate the event likelihood.

To initialize the process we must specify initial values $\mu[z_1]$ and $V[z_1]$. We use these in Eq. (A.12) to calculate $\mu[y_1]$ and $V[y_1]$ and hence the likelihood for the first event, y_1, from Eq. (A.11). We then calculate the posterior moments using Eq. (A.15). For the next event, $j = 2$, these posterior moments are assigned to $\mu[z_{j-1}]$ and $V[z_{j-1}]$ in Eq. (A.10) and the likelihood calculated. The procedure is iterated through all the events.

The model prediction of the OU process is a Gaussian distribution at each event (at time t_j) with mean and variance given by Eq. (A.10). Unlike the memoryless time series models, the OU process requires the measurement of the one previous event in addition to the model parameters in order to predict the next event (this is the Markov property). The relevant model prediction of event j is therefore given by the prior distribution of Eq. (A.10) – which has not yet looked at y_j – and not by the posterior distribution of Eq. (A.15), which has.

The parameters of the process are $\theta = (\mu[z_1], V[z_1], \tau, c)$ (and implicitly the initial time, t_1). Figure A.1 shows an example of a simulated OU process and the model predictions thereof.

As it stands, the long-term mean of this OU process is zero. We can introduce this long-term mean as an additional parameter, b, of the model. Equation (A.10a) then becomes

$$\mu[z_j] = \mu[z_{j-1}]v + b(1 - v). \tag{A.16}$$

The variance is unchanged. (See also Brockwell & Davis 2002, Sect. 10.4.) Note that this corresponds to solving a different differential equation, namely one in which we have the additional term $(b/\tau)dt$ on the right-hand-side of Eq. (A.1). (Note that b is the *long-term* mean of the process rather than the mean of the data.)

Now that we can calculate the likelihood, we can calculate the evidence or sample the posterior PDF. By partitioning the data set we can also use posterior sampling to evaluate the cross-validation likelihood, as described in Sect. 3.2. Note that whatever partitioning we do, when it comes to calculating the partition likelihood for data D_k we must still use all of the data to

---

[6] The product of two Gaussians f and g with means μ_f and μ_g and variances V_f and V_g is another Gaussian with

mean $\dfrac{\mu_f V_g + \mu_g V_f}{V_f + V_g}$ and variance $\dfrac{V_f V_g}{V_f + V_g}$. (A.14)
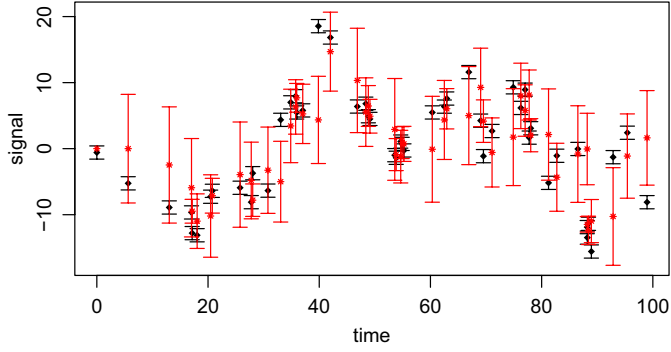
Fig. A.1. Example simulated OU process. The black points (with a uniform random time distribution) have been simulated from an OU process with parameters $\tau = 10$, $c = 20$ and initial conditions $t_1 = 0$, $z_1 = 0$, to which Gaussian measurement noise with mean zero and unit standard deviation has been added (as indicated by the black error bars). This red points show the predictions of this process using the OU process time series model with parameters $\tau = 10$, $c = 20$, $\mu[z_1] = 0$, $V[z_1] = 0$. The prediction for each event is a Gaussian in the signal with mean and variance given Eq. (A.10).

predict the full sequence of events. That is, for a given θ_n, we predict the entire sequence of J events using all the data, but only make use of those event likelihoods which are appropriate. Specifically, to calculate the posterior to use in MCMC sampling (Eq. (14)) we just select the likelihood for the events in D_{-k}, and to calculate the likelihoods in Eq. (11) (or (12)) we just use the events in D_k. The OU process depends not only on the model parameters but also on the state at the previous time step, so we should not be changing these time steps by removing events when predicting the sequence.

A.3. Literature note

I am not aware of an explicit derivation in the literature either of the above posterior recurrence relation (although Berliner 1996 outlines the same thing) or of the event likelihood for the OU process. Kelly et al. (2009) write down similar equations for the latter (their Eqs. (6)–(12)), but in a significantly rearranged form. Kelly et al. also assume a specific initial value (zero) for the initial state variable, z_0 (their Eq. (8)), whereas I give this a distribution. (In my formulation we can achieve a specific initial value by setting $V[z_1] = 0$.) My expression for the variance (Eq. (A.10b)) therefore has an additional term compared to theirs (their Eq. (A.5), which is V_z in my notation), because they are conditioning on an fixed value of the process at the previous step whereas I assume this itself has a variance, $V[z_{j-1}]$. (See also Sect. 10.4 of Brockwell & Davis 2002.) Closely related formulations of this process – but not the likelihood calculation – are given in Jones (1986; Sects. 4 and 5) and Kozlowski et al. (2010; appendix).

Appendix B: Simplifying the event likelihood integration

The calculation of the likelihood for a single event in principle requires a 2D integration (Eq. (7)). This can, however, be reduced to a 1D integration or even just a function evaluation under certain circumstances.

B.1. Dropping the stochastic signal component of the time series model (TSMod2 bypass)

If TSMod2 is a Gaussian (Eq. (3)) in which ω is very small compared to the scale of signal variations, then the only contribution

to the event likelihood is at the prediction of the signal by TSMod1. For given θ_1, the magnitude of the event likelihood is then dictated only by the measurement model, i.e. how close the measured y_j is to the prediction z_j. In this limit $\omega \to 0$, the signal part of the time series model (Eq. (2)) becomes $P(z_j|t_j, \theta_1, \theta_2, M) = \delta(z_j - \eta[t_j; \theta_1])$. This gives us a purely deterministic signal in the time series model; we "bypass" TSMod2. The event likelihood integration (Eq. (7)) then becomes a 1D integration[7]

$$P(D_j|\sigma_j, \theta, M) = \int_{t_j} \underbrace{P(D_j|t_j, z_j = \eta[t_j; \theta_1], \sigma_j)}_{\text{Measurement model}} \, \underbrace{P(t_j|\theta_3, M)}_{\text{Time series model}} \, \mathrm{d}t_j. \tag{B.1}$$

B.2. Small uncertainties on the measured times

If the uncertainty on the measured time, σ_{s_j}, is very small compared to the time scale over which the time series model varies, then the integral over t_j in the event likelihood will have a significant contribution only for times t_j close s_j. This must hold for any sensible measurement model or definition of uncertainties. The time part of the measurement model can then be approximated by the delta function $\delta(t_j - s_j)$, and the integration over t_j is just unity. If the signal part of the measurement model is Gaussian, the event likelihood equation (Eq. (7)) becomes

$$P(D_j|\sigma_{y_j}, \theta, M) = \int_{z_j} \underbrace{\frac{1}{\sqrt{2\pi}\sigma_{y_j}} e^{-(y_j-z_j)^2/2\sigma_{y_j}^2}}_{\text{Measurement model}} \underbrace{P(t_j = s_j, z_j|\theta, M)}_{\text{Time series model}} \mathrm{d}z_j. \tag{B.2}$$

When TSMod2 is the Gaussian model this becomes

$$P(D_j|\sigma_{y_j}, \theta, M) =$$
$$\int_{z_j} \underbrace{\frac{1}{\sqrt{2\pi}\sigma_{y_j}} e^{-(y_j-z_j)^2/2\sigma_{y_j}^2}}_{\text{Measurement model}} \underbrace{\frac{1}{\sqrt{2\pi}\omega} e^{-(z_j-\eta[s_j;\theta_1])^2/2\omega^2} P(s_j|\theta_3, M)}_{\text{Time series model}} \mathrm{d}z_j. \tag{B.3}$$

This can be written as

$$P(D_j|\sigma_{y_j}, \theta, M) = P(s_j|\theta_3, M) \int_{z_j} f(y_j - z_j) g(z_j) \mathrm{d}z_j. \tag{B.4}$$

This is just a convolution of two Gaussian functions, f and g, which is another Gaussian with mean equal to the sum of the means of f and g and variance equal to the sum of the variances of f and g. The event likelihood is therefore

$$P(D_j|\sigma_{y_j}, \theta, M) =$$
$$P(s_j|\theta_3, M) \frac{1}{\sqrt{2\pi(\sigma_{y_j}^2 + \omega^2)}} e^{-(y_j-\eta[s_j;\theta_1])^2/2(\sigma_{y_j}^2+\omega^2)} \tag{B.5}$$

i.e. involves no integration. The time part of the time series model, $P(t_j = s_j|\theta_3, M)$, is simply evaluated at the measured

---

[7] As we now have no stochastic element in either t or z, the reader may wonder why this integral is over t rather than z, i.e. why there is an asymmetry. The point is that we need to integrate along the path of the (deterministic) function $z_j = \eta[t_j; \theta_1]$. As this only requires one parameter, we only have a one-dimensional integral. Whether we parametrize this with t_j or z_j is unimportant, but having written the function as $z_j = \eta[t_j; \theta_1]$ rather than $t_j = \eta'[z_j; \theta_1]$, t_j is the more natural choice.

time, s_j. One particular application of this is to calculate the event likelihood for the *no-model*, the time series model in which there is no stochastic component and the deterministic component is just the mean of the signal. This is obtained by setting $\omega = 0$ and $\eta = \overline{y_j}$ in Eq. (B.5). The signal is therefore expected to be just Gaussian noise fluctuations about a constant. The total likelihood for the no-model, L^{NM}, is the product of these event likelihoods (Eq. (8)). This is a useful baseline model against which to compare the likelihood of other models. As this model has no adjustable parameters, this likelihood is equal to both the evidence and the K-fold CV likelihood.

B.3. Both TSMod2 bypass and negligible time uncertainties

If, in addition to a purely deterministic time series model, we also have negligible uncertainties on time, then the time part of the measurement model in Eq. (B.1) is a delta function, $\delta(t_j - s_j)$. The likelihood then involves no integration. If the signal part of the measurement model is a Gaussian, the likelihood is

$$P(D_j|\sigma_{y_j}, \theta, M) = P(s_j|\theta_3, M) \frac{1}{\sqrt{2\pi}\sigma_{y_j}} \mathrm{e}^{-(y_j - \eta[s_j;\theta_1])^2/2\sigma_{y_j}^2}. \quad (B.6)$$

We also reach this result if we set $\omega = 0$ in Eq. (B.5).

References

Bailer-Jones, C. A. L. 2004, A&A, 419, 703
Bailer-Jones, C. A. L. 2008, MNRAS, 384, 1145
Bailer-Jones, C. A. L. 2009, Int. J. Astrobiol., 8, 213
Bailer-Jones, C. A. L. 2011, MNRAS, 416, 1163; Erratum: MNRAS, 418, 2111 (CBJ11)
Bailer-Jones, C. A. L., & Mundt, R. 2001, A&A, 367, 218 (BJM)
Berger, J. O., & Sellke, T. 1987, J. Am. Stat. Associat., 82, 112
Berliner, L. M. 1996, in Fundamental Theories of Physics, 79, Maximum Entropy & Bayesian Methods: Proceedings of the 15th International Workshop, eds. K. Hanson, & R. N. Silver, Santa Fe, NM, 1995
Brockwell, P. J., & Davis, R. A. 2002, Introduction to time series and forecasting, 2nd edn. (Springer)
Chatfield, C. 1996, The analysis of time series, 5th edn. (London: Chapman & Hall)
Christensen, R. 2005, Amer. Statist., 59, 121
Gillespie, D. T. 1996a, Phys. Rev. E, 54, 2084
Gillespie, D. T. 1996b, Am. J. Phys., 64, 225
Jaynes, E. T. 2003, Probability theory. The logic of science (Cambridge University Press)
Jones, R. H. 1986, J. Appl. Probab., 23, 89
Kadane, J. B., & Lazar, N. A. 2004, J. Am. Stat. Assoc., 99, 465
Kass, R., & Raftery, A. 1995, J. Am. Stat. Assoc., 90, 773
Kelly, B. C., et al. 2009, ApJ, 698, 895
Kozlowski, S., et al. 2010, ApJ, 708, 927
MacKay, D. J. C. 2003, Information Theory, Inference and Learning Algorithms (Cambridge University Press)
Morales-Calderón, M., Stauffer, J. R., Kirkpatrick, J. D., et al. 2006, ApJ, 653, 1454
Radigan, J., Jayawardhana, R., Lafrenière, D., et al. 2012, ApJ, 750, 105
Reiners, A., & Basri, G. 2008, ApJ, 684, 1390
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. 2002, J. R. Statist. Soc. B, 64, 583
Uhlenbeck, G. E., & Ornstein, L. S. 1930, Phys. Rev., 36, 823
Vehtari, A., & Lampinen, J. 2002, Neural Comput., 14, 2439
von Toussaint, U. 2011, Rev. Mod. Phys., 83, 943