in preparation

Fitting the Undetected: Light-curve Feature Estimation in the Presence of Non-detections

Some wild & crazy guys

ABSTRACT

In multi-epoch imaging surveys, faint variable sources will not be detectable at all epochs. In principle, the best approach is to perform photometric measurements anyway, so that there is a datum recorded at at every epoch. In practice, most surveys only provide detection upper limits at some epochs; in some not even upper limits but just the information that the star was not observed. Here we demonstrate with real data on periodic variables from the All-Sky Automated Survey (ASAS) that these non-detections are nonetheless useful in fitting and parameter estimation; their inclusion improves both precision and accuracy. One novel aspect of this work is that we do not require the data source to include accurate—or even any—information about the uncertainty variances on detections or the values of the upper limits; we find that we can model these as latent variables, even when they are different at every epoch. Using realistic simulations of Mira variable stars, we obtain amplitude estimates that are **XXX%** more accurate by incorporating the non-detections into the analysis. For Mira and RR Lyrae variables observed by ASAS, our method obtains a tighter period - amplitude relationship than the standard technique.

1. Introduction

Surveys working in real-time tend to analyze each image with an eye to populating databases quickly. Unless a particular region of the sky gets special treatment (such as it is a known place of interest), discovery of variability happens by comparison of flux values at the database level. Analysis on such databases of transient/variables thus deals with censored data where there is often less information available that there should be. Even then, we might be able to infer what the thresholds were based on S/N measurements of objects neighboring the source of interest.

♣ **joey:** What people actually do, what problems that could cause, and what people should do ♣

A photometric survey (think of it as operating in a single photometric bandpass for now) scans over the celestial position of a particular variable star a large number of times t_i. At each of these times, the imaging data are analyzed by software, which treats each scan as a *completely independent* survey. That is, when analyzing the data from time t_i, none of the information from any of the other times is used in any way.

At each time t_i, the star is either detected ($q_i = 1$) by the software or not ($q_i = 0$). When it is detected, the software returns a (possibly bad) flux value f_i and (likely bad) uncertainty variance s_i^2. When the source is *not* detected ($q_i = 0$), nothing is reported—in this case, because the software is not awesome, it doesn't see any reason to report anything at all, since on the non-detect passes, it has no inkling that this patch of the sky is interesting in any way. Welcome to the desert of the real.

Beyond this, there are two additional issues. The first is that each night of imaging is different in an unreported and unknown way. That is, there is different transparency, sky brightness, and point-spread function. So the detection limit, or completeness level, or censoring of the catalog is different every night. The second issue is that either because we are getting the data from a non-generous source, or because the conditions change rapidly and unpredictably, we can't analyze all the sources in a finite patch simultaneously. For each variable star of interest, we *only* get the data on that star itself.

♣ **joey:** RRL end at 120 kpc in Sesar et al. paper ♣

♣ **joey:** discuss in context of: parameter estimation (amplitude / period), discovery (we're gonna find a lot more variable sources if we take the non-detections into account), and classification; lossiness ♣

♣ **joey:** let's make a point to stress variable discovery ♣

♣ **joey:** Numbers of LCs affected: Stripe 82, ASAS ♣

2. Formulating the Model

As a reminder, we are modeling light curve data in the presence of non-detections, which are epochs of observation in which no detection of the source of interest was made.

2.1. Preliminaries

For each astronomical object, a photometric survey measures a multi-epoch light curve over N (typically unevenly spaced) epochs. At each epoch i, with associated time t_i, the survey takes an exposure at the location of the object and either (a) detects the object and records an estimate of its photon flux, f_i and the variance of the statistical uncertainty in that estimate, s_i^2, or (b) fails to detect the object. In the latter case, most modern surveys either record a reference value to signify that no detection was made (as is done in ASAS) or an estimate of the upper detection limit, b_i, which is the brightest the object could have been given that it was not detected at a significant level by the software.

There are many reasons that a source might not show up in a catalog. These include:

- low S/N of the source, due either to a higher noise level or a fainter signal,
- the source falling outside of the detection window (e.g., near chip edge),
- occulting of the source by an artifact of the detector (e.g., hot pixels, masked out), or
- the source was out-shone by an intervening object (e.g., asteroid, comet, variable star, airplane, etc.).

Here, we present a statistical model that can be used to detect variable sources and model their variability using multi-epoch light curves containing epochs of non-detection. Previous efforts to detect and model variability using multi-epoch photometry have typically ignored non-detections (REFs) or used them in an ad-hoc manner lacking statistical rigor (REFs). An exception is Lang et al. (2009), who measure proper motions of sources in SDSS falling below the detection limit.

In this paper, we will assume that f_i and s_i^2 take on real-valued numbers in epochs for which the source was detected and receive the reference value `NA` in epochs where no detection was made. For notational convenience we assemble all the light curve data for one source into a data set D given by

$$D \equiv \{D_i\}_{i=1}^{N} \tag{1}$$

$$D_i \equiv (f_i, s_i^2) \quad , \tag{2}$$

where (f_i, s_i^2) are the light curve measurements at t_i. The goal, for each astronomical object, is to construct the likelihood for the data D, given a set of model parameters that describe the variability of the object and the characteristics of the observations, and then to maximize the likelihood with respect to the model parameters.

2.2. The Light Curve Model

To model the mean brightness of the light curve as a function of time, we use a multiple-harmonic Fourier model with angular oscillation frequency ω,

$$\mu_i \quad = \quad A_0 + \sum_{k=1}^{K} A_k \, \sin(t_i \, \omega k) + B_k \, \cos(t_i \, \omega k) \quad , \tag{3}$$

where μ_i is the flux at time t_i, $\sqrt{A_k^2 + B_k^2}$ is the amplitude of the $k^{\rm th}$ harmonic of the frequency ω and $\tan^{-1}(B_k, A_k)$ is the relative phase offset of harmonic k. The number of harmonics, K, can either be fixed or treated as a free parameter over which to optimize the likelihood. In addition to the expected brightness in Equation 3, we assume that there is uncertainty or inappropriateness in the model, leading to a *model variance* s_μ^2 at each point (assumed constant but easily generalized).

Additionally, we need to instantiate a latent variable, b_i, to represent the detection limit, in units of flux, at epoch i. The b_i parameter is essential because it constrains the possible values of the mean brightness, μ_i, of the light curve when there is a non-detection. By employing a hierarchical model for the distribution of b_i, we can fully utilize all of the information encoded in both the detections and non-detections when computing the data likelihood. At each epoch, we connect the observed flux, f_i, to the latent variable, f_i^*, signifying the true observable flux of the object, via the detection limit, b_i by

$$f_i \quad = \quad \begin{cases} f_i^* & \text{if } f_i^* \geq b_i \\ \texttt{NA} & \text{if } f_i^* < b_i \end{cases} \tag{4}$$

so that the observed flux is NA only when the true observable flux, f_i^*, is below the detection limit, b_i.

Also, because we do not necessarily believe the reported measurement uncertainties, s_i^2, we choose to introduce a parameter, σ_i^2, to represent the true uncertainty variance for the brightness measurement at epoch i.

Hence, our initial model consists of the parameter vectors $\{f_1^*, ..., f_N^*\}$, $\{b_1, ..., b_N\}$ and $\{\sigma_1^2, ..., \sigma_N^2\}$, and the model parameters

$$\theta \quad \equiv \quad (\omega, A_0, \{A_k, B_k\}_{k=1}^K, s_\mu^2, \cdots) \quad , \tag{5}$$

along with prior information about observation times, $\{t_1, ..., t_N\}$ and other prior assumptions ♣ **joey:** we should be more explicit here ♣ , which we make explicit by creating the prior information set

$$I \quad \equiv \quad (\{t_i\}_{i=1}^N, \text{assumptions}) \quad . \tag{6}$$

Our goal is to write down the form of the likelihood of the data, D, given θ and I. Then, for each light curve we can search for the vector θ that maximizes the data likelihood, and use those maximum likelihood estimates for downstream astrophysical inference.

2.3. Statistical Model for Light Curves with Non-Detections

We model the observed flux, f_i, as a Gaussian distribution with variance that has both measurement (σ_i^2) and model (s_μ) contributions. In the case that a detection is made, we require that the observed brightness be greater than the brightness of the detection limit ($f_i = f_i^* \geq b_i$) while in the case that no detection is made, we require that the brightness (if it could be measured) be less than the detection limit ($f_i^* < b_i$). To derive the likelihood of the observed f_i, given σ_i^2, θ and I, we must integrate over the prior distribution of the unknown b_i,

$$\begin{aligned} p(f_i|\sigma_i^2,\theta,I) &= \begin{cases} N(f_i^*|\mu_i,\sigma_i^2+s_\mu^2)\,\int_0^{f_i^*} p(b_i|\theta)\,\mathrm{d}b_i & \text{if } f_i \neq \mathtt{NA} \\ \int_0^\infty \int_0^{b_i} N(f_i^*|\mu_i,\sigma_i^2+s_\mu^2)\,p(b_i|\theta)\,\mathrm{d}f_i^*\,\mathrm{d}b_i & \text{if } f_i = \mathtt{NA} \end{cases} && (7) \\ p(b_i|\theta) &= N(b_i|B,V_B) && (8) \\ \theta &\equiv (\omega, A_0, \{A_k,B_k\}_{k=1}^K, s_\mu^2, B, V_B, \cdots) \quad , && (9) \end{aligned}$$

where we have introduced the hyperparameters B and V_B for the Gaussian prior distribution of b_i. In Equation 7, in the epochs for which a detection was made ($f_i \neq \mathtt{NA}$), we marginalize over the unknown b_i from 0 (low brightness limit) to f_i^*, enforcing that the detection limit be fainter than the observed brightness. Likewise, in the epochs for which no detection was made ($f_i = \mathtt{NA}$), we integrate the joint (f_i^*, b_i) likelihood over all possible values of b_i and over the unknown f_i^* from 0 to b_i, ensuring that the brightness (if it were able to be observed) be fainter than the detection limit.

In the above, we have assumed no extra information on each of the b_i values besides the knowledge of whether a detection was made on that epoch. Hence, we draw, in Equation 8, each b_i value from a global prior distribution which is the same at all epochs. If instead, we are given an estimate of b_i plus its error distribution for each epoch (which, in principle can be inferred from the raw telescope images), we can replace Equation 8 with a different distribution per epoch. In the case that the b_i are assumed to be completely known (without error), the data likelihood of f_i becomes

$$p(f_i|\sigma_i^2,\theta,I,b_i) \quad = \quad \begin{cases} N(f_i^*|\mu_i,\sigma_i^2+s_\mu^2)\,I(f_i^* \geq b_i) & \text{if } f_i \neq \mathtt{NA} \\ \int_0^{b_i} N(f_i^*|\mu_i,\sigma_i^2+s_\mu^2)\,\mathrm{d}f_i^* & \text{if } f_i = \mathtt{NA} \end{cases} \qquad (10)$$

where the boolean indictor function, $I(f_i^* \geq b_i)$, is 1 if $f_i^* \geq b_i$ and 0 otherwise.

Next, we model the reported variance, s_i^2, on the uncertainty of the magnitude measurement. Instead of assuming that s_i^2 is a perfect, error-free measurement, we probabilistically connect it to the true uncertainty variance, σ_i^2 through a Gamma likelihood, which is defined to take on non-negative values. Our likelihood of observed s_i^2, given σ_i^2, θ and I, is

$$p(s_i^2|f_i,\sigma_i^2,\theta,I) \quad = \quad \begin{cases} \Gamma(s_i^2|\sigma_i^2,V_\sigma) & \text{if } f_i \neq \texttt{NA} \\ 1 & \text{if } f_i = \texttt{NA} \end{cases} \tag{11}$$

where $\Gamma(x|m,V)$ is the standard Gamma distribution with mean m and variance V, and we have added a model parameter V_σ to represent the variance in the distribution of reported uncertainties given the true uncertainty. (Typically, the Gamma distribution is parameterized by parameters (α,β), where the mean is $\alpha\beta$ and the variance is $\alpha\beta^2$.) The purpose of including the likelihood in Equation 11 to the model, in practice, to keep the σ_i^2 from drifting very far away from the s_i^2, as set by the hyperparameter V_σ.

Putting it all together, we can write down the likelihood of the data, D_i, for the parameters θ, on a single epoch, as

$$p(D_i|\theta,I) \quad = \quad \int_0^\infty p(f_i|\sigma_i^2,\theta,I)\,p(s_i^2|f_i,\sigma_i^2,\theta,I)\,p(\sigma_i^2|\theta,I)\,\mathrm{d}\sigma_i^2 \tag{12}$$

$$p(\sigma_i^2|\theta,I) \quad \propto \quad \sigma_i^{-1} \tag{13}$$

$$\theta \quad \equiv \quad (\omega, A_0, \{A_k,B_k\}_{k=1}^K, s_\mu^2, B, V_B, V_\sigma) \quad , \tag{14}$$

where we have inserted the expressions from Equations (7) and (11), and integrated out the nuisance parameter, σ_i^2. We have assumed a Jeffrey's prior on σ_i^2, which is non-informative and invariant to reparametrization of the variance. We could alternatively use an inverse-gamma prior, which takes two hyperparameters (and is conjugate in the case of a normal likelihood with unknown variance).

Finally if we assume that the data collected at each epoch are independent given the model parameters, we have that

$$p(D|\theta,I) \quad = \quad \prod_i p(D_i|\theta,I) \quad . \tag{15}$$

This is the likelihood for the entire data set (all the measurements and non-detections of this star from all the epochs, as delivered by the survey) given the $2K+5$ parameter vector θ. This model "correctly" or at least "justifiably" uses all of the information available, without making strong assumptions about the survey or its veracity.

2.4. Implementation

♣ **joey:** Describe Python implementation here ♣

3. Experiments

♣ **joey:** Note: ASAS provides periods and amplitude for all of the objects in ACVS! We can and should use this as a comparison set ♣

♣ **joey:** From a brief reading of a few Mira papers, it seems that some papers throw away Miras for which no trough of the LC was observed. This seems silly. I have an idea of the P-A plots from Hipparcos. Should be a good comparison set. Also, we can consider applying this to Hipparcos, which has publicly available data. ♣

3.1. Simulating Faint Mira Variables

In this first experiment, we begin with a well-observed Mira variable from the ASAS Catalog of Variable Stars (ACVS, Pojmanski et al. 2005) , ASAS 235627-4947.2. This star has a pulsation period of 266.6286 days and a Lomb-Scargle amplitude of 2.38 mag (V band). Note there are no non-detections in the ASAS light curve.

To simulate faint Mira stars from ASAS 235627-4947.2, we use the following procedure:

1. Convert the observed magnitudes, m_i, and errors, $s_{m,i}$ to fluxes, f_i and flux errors, $s_{f,i}$ (we assume a V-band zero-point of 3.67×10^{-9} erg/s/cm^2/Å)

2. For a given flux dimming parameter, d, sample the new fluxes, $\tilde{f}_i$, from a Gaussian distribution centered around f_i/d with standard deviation sampled from the empirical ASAS distribution of $s_{f,i}|f_i$. ♣ **joey:** I found that relationship to be linear with a slope ≈ 1, which is probably BS since the ASAS mag errors are all crap. So I added a reference value to each flux error to ensure a mag limit ≈ 14.5 mag. ♣

3. Fluxes that are not at least 5σ above zero are denoted as non-detections and their flux estimates (and errors) are censored.

Folded light curves of ASAS 235627-4947.2, dimmed by four different values of d, are plotted in Figure 1.

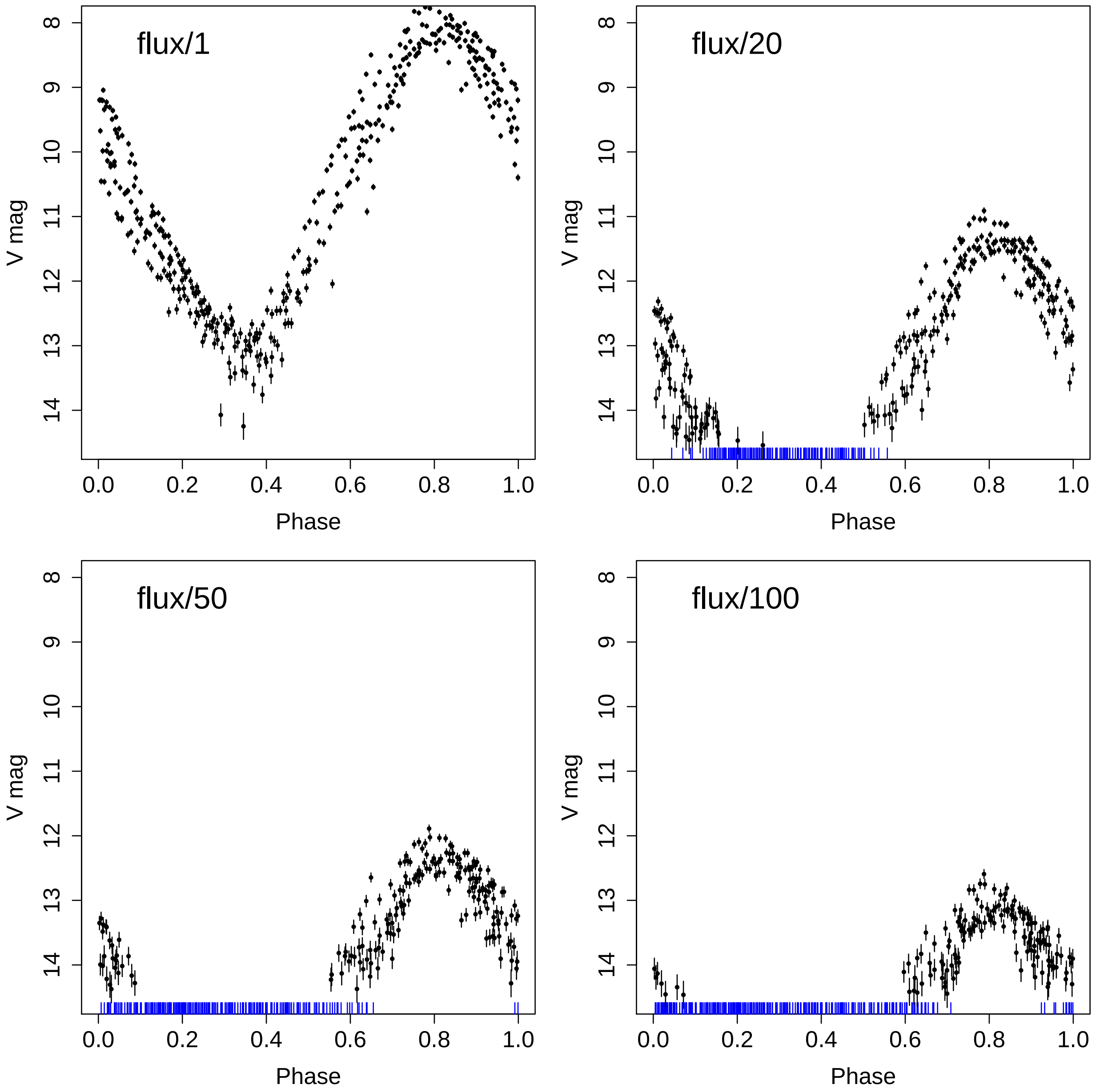

Fig. 1.— Folded light curves for data simulated from the Mira variable ASAS 235627-4947.2 using the prescription in §3.1. Blue tick marks along the bottom axis denote phases where there were non-detections. As the original flux measurements are dimmed by higher factors, the troughs of the sinusoidal light curve are censored, resulting in an incomplete view of the data.

3.2. Results of Fitting Miras

4. Results: ASAS Light Curves

We use the methodology of Richards et al. (2011) to select the top ASAS Mira and RR Lyrae, Fundamental Mode candidates. Using a posterior probability threshold of 0.8 gives us 1720 Mira and 1029 RR Lyrae candidates.

4.1. Mira Variables

Show P - A relationship before and after using the method

4.2. RR Lyrae Variables

Show P - A relationship before and after using the method

There is an RRL P - A relationship (strong linear anti-correlation)

5. Discussion

What people can do, starting from the data-taking procedure.

limitations

REFERENCES

Lang, D., Hogg, D. W., Jester, S., & Rix, H.-W. 2009, AJ, 137, 4400

Pojmanski, G., Pilecki, B., & Szczygiel, D. 2005, Acta Astronomica, 55, 275

Richards, J. W., et al. 2011, ApJ, 743

This preprint was prepared with the AAS LaTeX macros v5.2.