BAYESIAN STATISTICS 6, pp. 000–000
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)

Bayesian Analysis of Cepheid Variable Data

WILLIAM H. JEFFERYS AND THOMAS G. BARNES, III
University of Texas, USA

SUMMARY

Cepheids are a type of variable star that play a key role in establishing the astronomical distance scale. These objects undergo regular pulsations, shrinking and expanding as their luminosities and colors vary in synchronism. It has long been known that the period of the pulsation is related to the luminosity of the star (the Period-Luminosity relationship), and therefore these objects become "standard candles" that allow us to estimate the distances of objects once we measure their periods and the flux from the stars received on Earth. Calibrating the zero-point of the Period-Luminosity relationship therefore becomes a critical measurement that has repercussions throughout astronomy. These stars are very luminous, and therefore can be seen to rather large distances, which makes them very valuable as standard candles; however, most of them are out of reach of the most direct distance measurements, and even the nearest can only be measured with rather large errors. We have been investigating another approach to determining the distance to these objects that depends upon a detailed study of both photometry and radial velocity information. Analysis of such data has up until now been rather *ad hoc*; we have been exploring Bayesian methods in order to systematize the technique. Among the features with which a fully Bayesian analysis of these data must contend are: Model selection (the light curves and velocity profiles need to be modeled accurately yet parsimoniously), and errors-in-variables. We will describe several approaches, one due to Gull, and another using the program BUGS by Spiegelhalter *et. al.* We are particularly interested in learning ideas that may assist us in carrying out our program.

Keywords: CEPHEID VARIABLES; MODEL SELECTION; ERRORS-IN-VARIABLES.

1. STATEMENT OF THE PROBLEM

The classical method of measuring the distance to a star is to estimate the angle ϕ subtended by the Earth's orbit as the ratio R/d of the known radius R of the Earth's orbit to the distance to the star. Thus, $\phi \approx tan\, \phi = R/d$. ϕ can be measured directly as a displacement in the angular position of the star due to the motion of the Earth around the Sun. Unfortunately, this method works only for the nearest stars, because the angle ϕ is extremely small; it is less than one arcsecond even for the nearest star from the Sun, and with the best current techniques, the method can reliably be used (with errors of $< 10\%$) only for stars within about 100 parsecs of the Sun (about 325 light years). Since the distance to the center of our own galaxy is estimated to be about 8,000 parsecs, the distance to the nearest galaxy other than our own at about 50,000 parsecs, and the size of the observable universe of order 5×10^9 parsecs, it is easily seen that this method is wholly inadequate for determining the distances to most objects in the universe.

Another technique is to turn the relationship $\phi \approx R/d$ around, interpreting R as the radius of a star and ϕ as its angular radius. If we have some sort of estimate of R, then by measuring ϕ we can infer d. This method was first used by Tycho Brahe in the 16*th* century, who estimated

the angular radii of the brightest stars at about $1'$ of arc, and believing that they were comparable in linear size to the Sun, which has an angular radius of about $15'$, was able to put a lower bound on the distance to these stars at 15 times the distance from the Earth to the Sun. That this was a gross underestimate is in hindsight no surprise; the actual diameters of the largest stars are now known to be hundreds of times larger than the Sun, and their angular sizes are thousands of times smaller than Tycho's estimate. But Tycho's basic idea was sound.

Fortunately, there is a very good way to estimate ϕ. From the Stefan-Boltzmann law, the luminosity L of a spherical star is proportional to $R^2T_e^4$, where R is the radius of the star and T_e is the effective temperature at the surface of the star. Since the observed flux F from the star is inversely proportional to d^2, it follows that $F \propto (R/d)^2T_e^4 = \phi^2T_e^4$. A similar proportionality holds for the flux F_B observed in any wavelength bandpass B. In practice, we use an empirical relationship due to Barnes and Evans to infer F_B from the observed color index of the star, which is strongly correlated with T_e (Barnes and Evans 1976; Barnes *et. al.* 1976).

However, only under special circumstances is it possible to get a similar handle on the actual radius R of a star. One of those circumstances occurs with Cepheid variables, since they expand and contract. This means that the spectral lines in the star's spectrum are shifted in wavelength due to the Doppler shift; as the near surface of the star expands towards us, the wavelengths are shifted towards the blue, and as it contracts, towards the red. This means that we can observe the radial velocity $V_r = \bar{V_r} + \Delta\ V_r$ of the surface of the star as a function of time, and by integration can determine the radius $R(t) = R_0 + \Delta\ R = R_0 - \int\ \Delta\ V_r dt$ (the negative sign coming in because the star expands as the near surface moves towards us).

This suggests that by measuring the changing flux from the star (as well as T_e, which also changes) we may be able to infer both $\phi(t)$ and $R(t)$; from the *amplitude* of these oscillations, we can infer the distance to the star from

$$\phi(t) = R(t)/d = R_0/d - (1/d)\int\ \Delta\ V_r dt = \phi_0 + \Delta\ R/d \tag{1}$$

.

2. THE CURRENT APPROACH

The current approach (e.g., Barnes, *et. al.* 1977, Gieren *et. al.* 1990, Gieren *et. al.* 1993) has been to do just as outlined above. That is, we observe the velocity curve, smooth it either by eye or by fitting a function, integrate the smoothed curve to obtain the displacement $\Delta R(t) = R(t) - R_0$, and then perform a least-squares fit of Eq.(1) to the photometric data, using the observed flux and color index to predict ϕ.

This is clearly inadequate on statistical grounds, for the following reasons: First, the fitting of the velocities is *ad hoc*. There is no clear way to determine how well the velocities have been fitted, if an "eyeball" method is used; and if a fit to a Fourier series is employed, one still must contend with the question of how many terms to take in the series. This faces us with a *model selection* problem.

Second, the fitting of Eq. (1) has been done with error in the independent variable $\int\ \Delta\ V_r\ dt$. This makes the problem into an errors-in-variables problem, and brings the worrying problem that the quantity $1/d$, which we are trying to estimate, may be estimated with avoidable bias. This point has been raised by Laney and Stobie (1995), who have advocated a maximum likelihood approach.

Third, since the error in $\Delta R(t)$ has not been taken into account, the variance of d will be underestimated. This is clearly undesirable.

3. GAUSSFIT—AN APPROXIMATE BAYESIAN APPROACH

For these reasons, we decided to investigate a Bayesian approach to this problem. We had available to us a software package, GaussFit (Jefferys *et. al* 1988), developed as a tool to reduce Hubble Space Telescope astrometry data, which implements a maximum likelihood algorithm due to Jefferys (1990). GaussFit includes a maximum likelihood analysis of the errors-in-variables problem. Of course, maximum likelihood estimation is not Bayesian. However, it can be considered as an approximation to a Bayesian maximum *a posteriori* (MAP) estimator with a flat prior. Indeed, GaussFit also provides the possibility of introducing a normal prior for any parameter by the obvious technique of regarding the parameter as an observation with a specified mean and variance, although we have not used that capability in our problem up to this point.

Since GaussFit solves the second and third problems by allowing us to solve the errors-in-variables problem exactly as a maximum likelihood problem, we decided to concentrate on the first of the points in Section 2: selecting which model of the velocity curve was the most satisfactory. To do this, we have employed a suggestion of Gull (1988). This idea is ideally suited to the GaussFit software, since it uses normal approximations and the numbers needed to apply it can be calculated from the maximum likelihood solution.

Gull begins by considering a linear model

$$x = A\theta + \epsilon \tag{2}$$

where x is the N-vector of observations, A an $(N \times M)$ design matrix, θ the M-vector of parameters, and ϵ the N-vector of errors, which are assumed *iid* normal: $\epsilon_j \sim \mathcal{N}(0, \sigma)$. Observing that by neglecting ϵ we would have, approximately,

$$x'x \approx \theta' A'A \theta$$

he suggests a maximum-entropy prior $p(\theta \mid M)$ determined with the constraint

$$\mathcal{E}(\theta' A'A\theta) = x'x$$

This yields the prior $p(\theta \,|\, M, \beta) \propto \beta^{-M/2} exp(-\theta' A'A\theta/2\beta)$. The likelihood is

$$\mathcal{L} \propto \sigma^{-N/2} exp(-(x - A\theta)'(x - A\theta)/2\sigma).$$

With Jeffreys priors on β and σ, and with the assumption that the data are very good compared to the unmodeled data so that $|\beta| \gg |\sigma|$, which is the case for our data, the posterior distribution splits into two integrable parts. Integrating over everything except M, Gull arrives at a very simple expression for the posterior probability of M:

$$p(M \,|\, x) \propto \left(\frac{V(0)}{V(M)}\right)^{\frac{N-M}{2}} \Gamma\left(\frac{N-M}{2}\right)\Gamma\left(\frac{M}{2}\right) \tag{3}$$

where $V(0) = x'x$ is the "sum-of-squared-residuals" for the *unmodeled* data and $V(M) = (x - A\hat{\theta})'(x - A\hat{\theta})$ is the "sum-of-squared-residuals" of the data modeled with the M-vector $\hat{\theta}$, which is the ordinary least-squares estimator for θ.

For simplicity, we have used this formula to evaluate the modeling of the velocity data by Fourier series. The results were very satisfactory, and agree well with what we intuitively see in the data. For the star T Monocerotis (T Mon), which is a somewhat difficult case since the velocity curve contains a strange, and physically real "wiggle," the fifth-order Fourier

polynomial owns 80% of the posterior probability, and the sixth-order polynomial owns 20%. Only negligible amounts of posterior probability are owned by the other choices. This agrees with what our eyeballs tell us (it passes Savage's "interocular traumatic test"). This can be seen in Figures (1-3), which show the result of the fit for fourth through sixth order. The fourth-order fit is clearly inadequate as it fails to fit a physically real "glitch;" It is difficult to choose between fifth and sixth order, but there is some evidence of overfitting in the sixth-order picture.

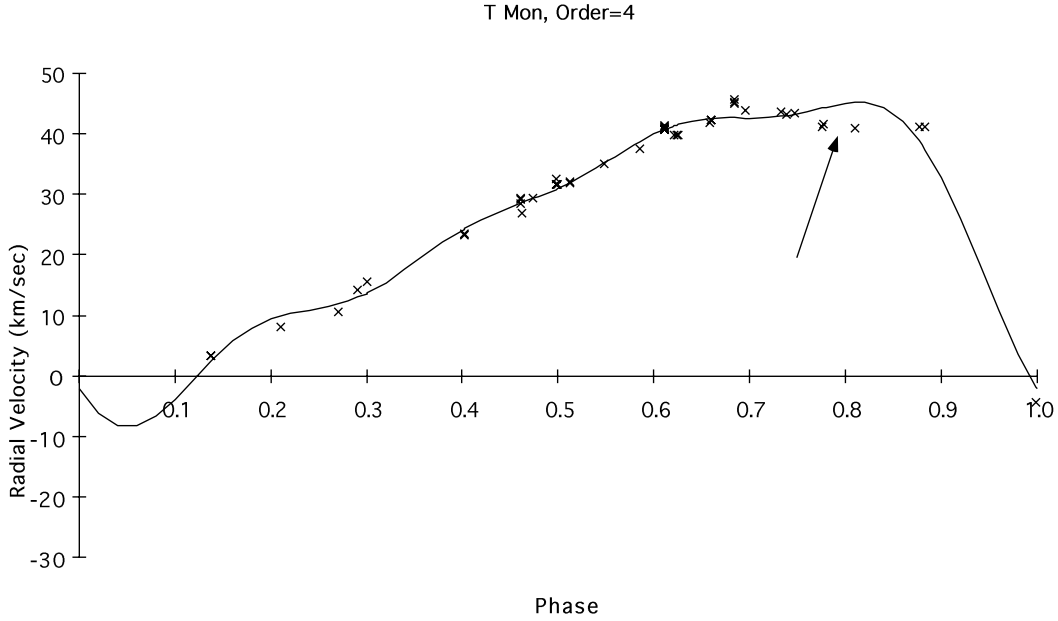

Figure 1. *The radial velocity data for T Mon fitted with a fourth-order trigonometric polynomial. The arrow points to a physically real "glitch" in the velocity. This fit is clearly inadequate.*

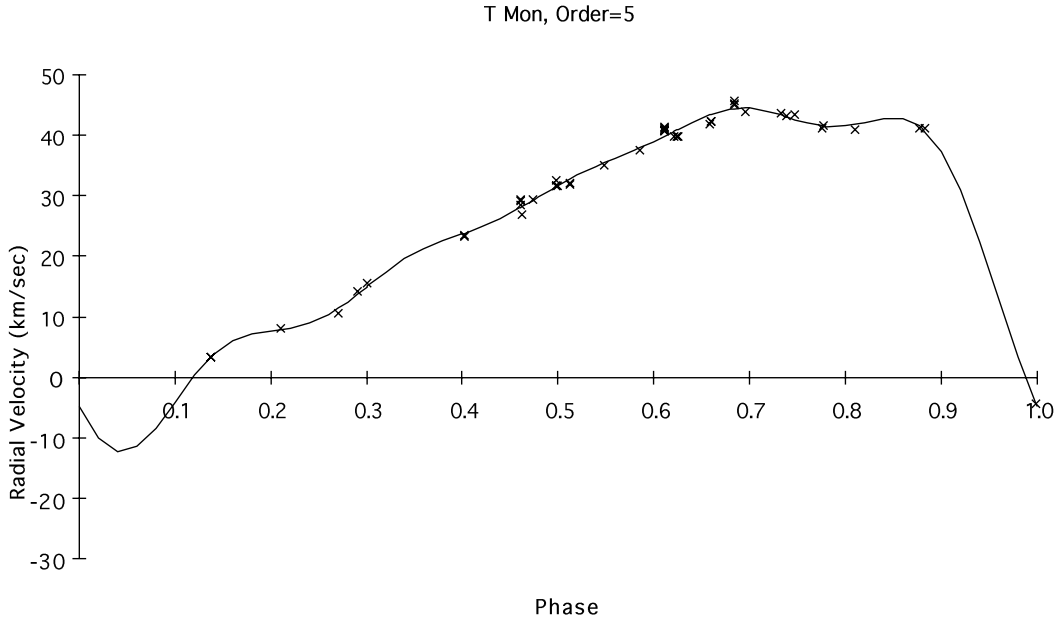

Figure 2. *The radial velocity data for T Mon fitted with a fifth-order trigonometric polynomial. This fit seems quite adequate to the data, including the fit to the "glitch" of Figure 1.*

We have so far analyzed the data from three stars: T Mon, T Vulpeculae and SZ Tauri. All gave similar results, with T Mon being the most difficult case because of its unusual velocity

curve. In each case, the order of trigonometric polynomial picked out by Gull's rule (Eq. (3)) agrees well with what the eye sees. So, we believe that Gull's rule is a simple and effective criterion for choosing how far to take the expansion, at least for our data.

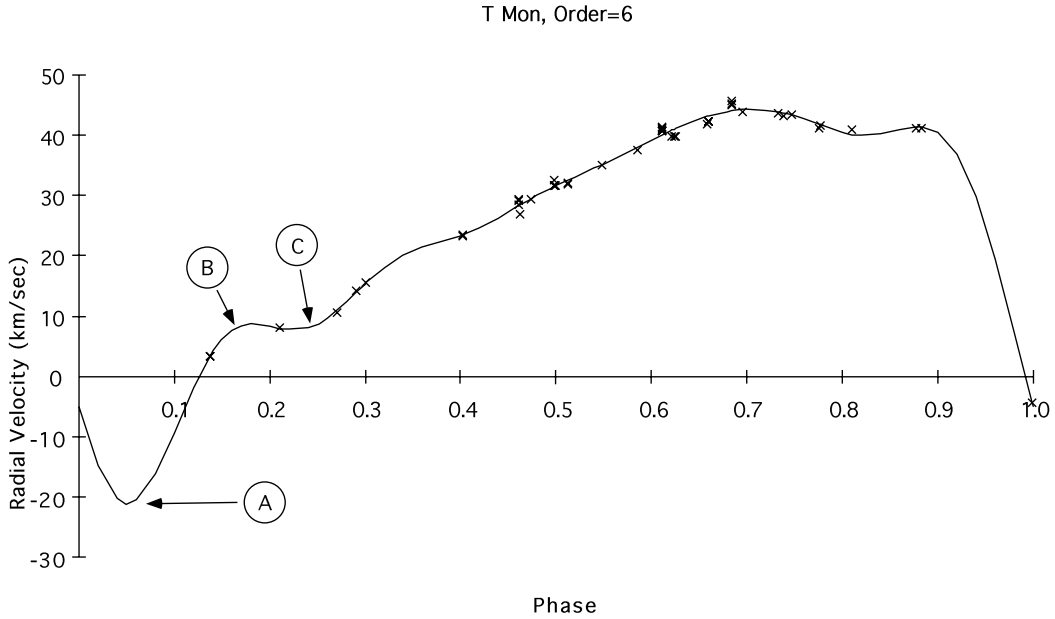

Figure 3. *The radial velocity data for T Mon fitted with a sixth-order trigonometric polynomial. This fit is not clearly better than the fit of Figure 2, and shows some evidence of overfitting, as indicated by the arrows $A - C$; these bumps are not supported by any data (c.f. Figure 2). Bump A, in particular, is much larger than in the lower order fit; Bumps B and C are probably a consequence of the program forcing the curve nearly through the adjacent points*

4. BUGS—A FULLY BAYESIAN APPROACH

Nonetheless, it is our aim to analyze our data from a fully Bayesian point of view, and to do this we turned to BUGS (Spiegelhalter *et. al.* 1996). Like GaussFit, BUGS provides a fully-featured computer language that is designed specifically for expressing inference problems. However, the inference is Bayesian, not maximum likelihood, and BUGS provides a variety of tools to express Bayesian problems in an economical and transparent way.

A significant oversight in BUGS, from our point of view, is the lack of the trignometric functions needed to handle the Fourier polynomials that we need. This is apparently a legacy of the biostatistics environment in which BUGS was developed—no one thought that anyone would find sines and cosines useful! But, BUGS is extremely well designed for the problems it is meant to solve, and (as we have found) it is bound to be found useful in the physical sciences as well. So we hope and expect that this and other deficiencies will be remedied in due course, and we encourage the authors of BUGS to add these capabilities to their program.

We were able to overcome the lack of sines and cosines by precomputing the design matrix A and providing it as input data to BUGS; this strategy worked well, although it will not be sufficient in the future, because some of our stars will also require determining the zero-point of the phase as an additional nuisance parameter. We can see work-arounds for this problem, but they are not pretty and it would be better if the trigonometric functions were available in the BUGS language itself.

Our ultimate aim is to produce a unified program that will allow us to analyze the photometric and radial velocity data from a star as a whole. It would determine all of the relevant coefficients

and parameters, and simultaneously solve the model-selection problem by considering each order of trigonometric polynomial, indexed by M, in the complete model. BUGS would determine which models are best favored by the data by computing the posterior probability as a function of M (integrating over all the other parameters). Since BUGS can automatically produce the marginal distributions of any desired parameter by monitoring the parameter's value after the initial "burn-in" period, we would then be able to determine the marginal distribution of the distance d to the star, averaged over all orders of model.

Thus far we have not been able to carry out this program to its conclusion. This is partly because we embarked on this aspect of our enterprise relatively recently, and are still learning how to use BUGS. We have successfully built a BUGS model that will perform a complete errors-in-variables solution for a *given* choice of M, and calculate the marginal distribution of d given M. However, we have not yet achieved a model that can choose amongst the different orders M. Our first attempts to do this work well with a simplified problem having only a small number of parameters, but for our actual problem, involving 15-20 parameters (mostly coefficients of the Fourier polynomials), we have not succeded in getting BUGS to sample all of the choices of M. Instead, the Gibbs sampler is getting "stuck" on one model (not necessarily even the "best" model, given our experience with the work reported in Section 3) and ignoring the others. We think that this is due to our inexperience with BUGS, and that more work will enable us to overcome this problem.

We are encouraged by our experience so far with BUGS. BUGS was not designed for problems like ours, yet what we have learned about it so far, despite the deficiencies we have found, indicates that BUGS can be used as an effective tool for Bayesian inference in problems very different from those for which it was designed. The authors of BUGS are to be congratulated for what they have accomplished, and we encourage them, with the cooperation of workers in other fields, to expand the capabilities of this very powerful tool.

We hope that our experience with BUGS will encourage others in the physical sciences to consider using it for their problems, and that they will communicate their needs to the authors of the program. We have had excellent feedback from Professor Spiegelhalter on the use of BUGS. We think that BUGS has great promise and look forward to its future development in directions that will make it more useful in the physical sciences.

5. WHAT'S NEXT?

The first thing on our agenda, of course, is to get our BUGS model working correctly with regard to model selection. Only when this is accomplished will we be able to move forward.

When this has been done, we want to experiment with a reasonable range of priors. Other than the prior on θ, which is crucial for the model selection problem, the prior that is most critical is perhaps the one on d, the distance to the star. The galaxy is flat, and we can expect the number of stars to go as $p(d) \propto d$, up to some cutoff representing the size of the galaxy. But there are all sorts of issues here, for example, extinction due to dust in the galactic plane tends to limit how far we can see in complicated ways. It is therefore necessary to see how the choice of prior may affect the quantities we are trying to estimate. If we were to find that the results depend sensitively on the prior (over a reasonable range of priors), then this would indicate that our results are not as well-determined as the formal Bayesian credible intervals might indicate. On the other hand, if the results are robust with respect to the priors, then we can have a reasonable degree of confidence in them.

We will then want to compare the results we get from the fully Bayesian BUGS analysis with the ones we get from our approximately-Bayes method using GaussFit. Since BUGS uses Markov chain Monte Carlo methods, it is relatively inefficient in its use of computer resources,

and the question of convergence is always an issue. If it were to turn out that BUGS and GaussFit gave similar results, we would probably decide to use GaussFit for the entire database, consisting of something over 100 stars. On the other hand, it is quite possible that the BUGS model will ultimately turn out to be the model of choice.

Ultimately we will want to compare our best solutions with the alternative approaches of Laney & Stobie (1995), and Gieren *et. al.* (1993), as well as with various other approaches that appear elsewhere in the astronomical literature. Do some of these methods build in unforseen biases? If so, how can they be dealt with? Is it possible to estimate the biases or must the entire reduction be done again *de novo*? If the latter, we may have a problem in even obtaining the data!

Finally, we note that Fourier polynomials appear frequently in the astronomical literature as approximations to periodic processes. An effective way to determine how many terms to take in these polynomials is of great interest in astronomy, and probably many other fields as well. So, our work on this problem has implications that are much wider than just the problem of the zero-point of the Cepheid period-luminosity relationship, important as that problem is.

REFERENCES

Barnes, T. G., Dominy, J. F., Evans, D. S., Kelton, P. W., Parsons, S. B. and Stover, R. J. (1977). The distances of Cepheid variables. *Monthly Notices of the Royal Astronomical Society.* **178**, 661–674.

Barnes, T. G. and Evans, D. S. (1976). Stellar angular diameters and visual surface brightness I. Late spectral types. *Monthly Notices of the Royal Astronomical Society.* **174**, 489–502.

Barnes, T. G., Evans, D. S. and Parsons, S. B. (1976). Stellar angular diameters and visual surface brightness II. Early and intermediate spectral types. *Monthly Notices of the Royal Astronomical Society.* **174**, 503–512.

Gieren, W. P., Moffett, T. J., Barnes, T. G., Matthews, J. M. and Frueh, M. L. (1990). The short-period Cepheid EU Tau. II. Physical properties of the star. *Astronomical Journal* **99**, 1196–1206.

Gieren, W. P., Barnes, T. G. and Moffett, T. J. (1993). The Cepheid period-luminosity relation from independent-distances of 100 galactic variables. *Astrophysical Journal* **418**, 135–146.

Gull, S. F (1988). Bayesian inductive inference and maxiumum entropy. *Maximum-Entropy and Bayesian Methods in Science and Engineering*. (G. J. Erickson and C. R. Smith, eds.) Dordrecht: Kluwer, 153–74.

Jefferys, W. H., Fitzpatrick, M. J. and McArthur, B. E. (1988). GaussFit—A system for least squares and robust estimation. *Celestial Mechanics* **41**, 39–49.

Jefferys, W. H. (1990). Robust estimation when more than one variable per equation of condition has error. *Biometrika*, **77** 597–607.

Laney, C. D. and Stobie, R. S. (1995). The radii of galactic Cepheids. *Monthly Notices of the Royal Astronomical Society.* **274**, 337–360.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996). BUGS 0.5: Bayesian inference Using Gibbs Sampling Manual (version *ii*). *Tech. Rep.*, MRC Biostatistics Unit, Institute of Public Health, Cambridge, England.

}