ISBA 2000, Proceedings, pp. 000–000

Model Selection for Cepheid Star Oscillations

WILLIAM H. JEFFERYS, THOMAS G. BARNES AND RAQUEL RODRIGUES
University of Texas at Austin

JAMES O. BERGER AND PETER MÜLLER
Duke University

SUMMARY

Cepheid variables are a class of pulsating variable stars with the useful property that their periods of variability are strongly correlated with their absolute luminosity. Once this relationship has been calibrated, knowledge of the period gives knowledge of the luminosity. This makes these stars useful as "standard candles" for estimating distances in the universe. This paper updates and expands work reported by Jefferys and Barnes (1999). We consider fully Bayesian inference using reversible-jump MCMC simulation that takes as data photometric and velocity information and gives as output posterior inference for useful physical information such as the absolute luminosity of the star, its distance, its radius, and other parameters. We model the photometry and velocities as Fourier polynomials with an unknown or selectable number of terms; the photometry and velocities are connected by nonlinear relations involving the physical parameters of interest. From amongst these models with varying numbers of Fourier coefficients we select models with the highest posterior probability, and obtain information on the physical parameters averaged over the models, with weights proportional to the posterior probabilities of the models. We discuss issues concerning priors, effectiveness of the sampling, and the practical results of our research program. We briefly discuss alternative photometric and velocity models using wavelets instead of Fourier polynomials, and alternative approaches to the priors.

Keywords: MODEL SELECTION; MODEL AVERAGING; REVERSIBLE JUMP MCMC; CEPHEID VARIABLE STARS.

1. A LITTLE ASTRONOMY—WHAT'S THE GOAL?

Cepheid variable stars pulsate, regularly varying their luminosity (light output) and size. We can measure both the velocity of the surface of the star as it moves in and out, and the variable luminosity and color of the star. There is a mathematical relationship between these quantities that enables us to infer the distance to the star. The period of pulsation is related to the luminosity through the *period-luminosity relationship*, according to which the log of the luminosity is a linear function of the log of the period. Determining the zero-point (intercept) of the period-luminosity relationship is a fundamental problem in astrophysics.

These stars are "standard candles" for determining the distances to the galaxies in which they are found. Cepheid variables are thus fundamental to understanding the cosmological distance scale, i.e., the size and age of the universe.

2. MATHEMATICAL MODEL AND LIKELIHOOD FUNCTION

We have unequally-spaced observations of velocity data $U_i, i = 1, \ldots, m$, and photometry data consisting of magnitude $V_i, i = 1, \ldots, n$ and color index $C_i, i = 1, \ldots, n$. We are given standard deviations σ_{U_i}, σ_{V_i}, σ_{C_i}. However, we are not very confident of these numbers and take the variances of the data to be given by $\sigma^2_{U_i}/\tau_U$, $\sigma^2_{V_i}/\tau_V$, $\sigma^2_{C_i}/\tau_C$, where the *numbers* τ_U, τ_V, τ_C,are to be estimated. Let u_i, v_i , and c_i denote unknown mean velocity, magnitude, and color index, respectively. Conditional on u_i, v_i and c_i we assume independent normal distributions

$$\begin{aligned} U_i &\sim N(u_i, \sigma^2_{U_i}/\tau_U) \\ V_i &\sim N(v_i, \sigma^2_{V_i}/\tau_V) \\ C_i &\sim N(c_i, \sigma^2_{C_i}/\tau_C) \end{aligned} \tag{1}$$

The velocities u and photometry (v, c) are periodic functions of the time, and so are functions of observed phases θ_i where $0 \leq \theta_i < 1$. An obvious strategy is to represent them as Fourier polynomials of some unknown or selectable order, resulting in a model selection/averaging problem. We need to do this only for u and v, since the colors c are mathematically related to u and v through (Eq. 2) below. M and N are the unknown order of the Fourier polynomials for the **U** and **V** data, respectively. The polynomials contain $2M + 1$ and $2N + 1$ terms, respectively, including the leading constant terms. Thus we write

$$\begin{aligned} \mathbf{u} &= u_0 + \mathbf{X}_u \mathbf{a}_u \\ \mathbf{v} &= v_0 + \mathbf{X}_v \mathbf{a}_v \end{aligned}$$

where u_0 and v_0 are the mean velocity and luminosity, $\mathbf{X}_u$ and $\mathbf{X}_v$ are $(m \times 2M)$ and $(n \times 2N)$ design matrices consisting of sines and cosines of multiple angles, evaluated at the phases of the data, and $\mathbf{a}_u$ and $\mathbf{a}_v$ are vectors of Fourier coefficients. Note that the velocity data and photometry data are taken independently, and the phases in general are different. Because the velocity and photometry data are taken at different times, there will also be an unknown phase error $\Delta\theta$ between the two (due to imperfect knowledge of the period of the star).

Figures (1-3) show maximum likelihood fits of the velocity data for the star T Monocerotis (T Mon). As can be seen, the fourth-order model does not appear to fit the data adequately. In particular, a physically real "glitch" near phase 0.8 is not fitted well. The fifth-order model seems to do an adequate job, but the sixth-order model shows evidence of overfitting (A–C). It will be interesting to see how these results compare to the results of our Bayesian analysis.

We also have the nonlinear relationship

$$c_i = \frac{1}{\beta}\left(0.1 v_i - \alpha + 0.5 \log\left(\phi_0 + \Delta R_i/s\right)\right) \tag{2}$$

where (α, β) are known constants, ϕ_0 and s are the angular size and distance of the star (to be estimated), and ΔR_i is calculated from the $\mathbf{a}_u$ by integrating the velocity term-by-term with respect to the phase. This allows us to write down the likelihood function directly from Eqs. (1). Some of the parameters appear in the resulting likelihood function in awkward and nonlinear ways that will make straightforward Gibbs sampling impossible. A suitably informed Metropolis scheme will be needed.

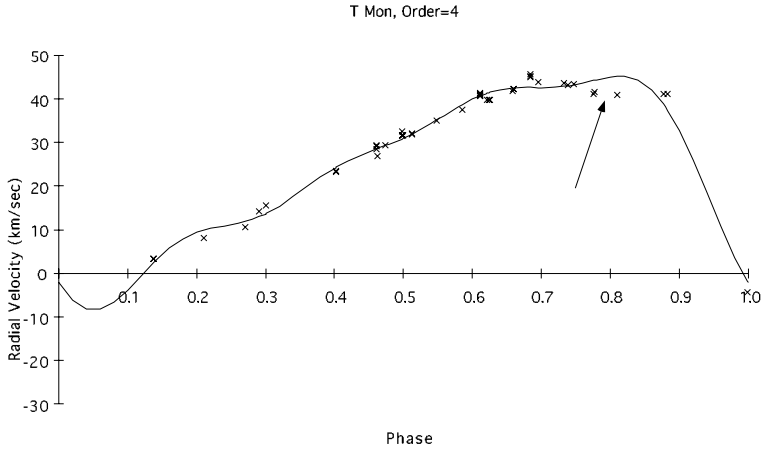

Figure 1. *The radial velocity data for T Mon fitted with a fourth-order trigonometric polynomial. The arrow points to a physically real "glitch" in the velocity. This fit is clearly inadequate.*

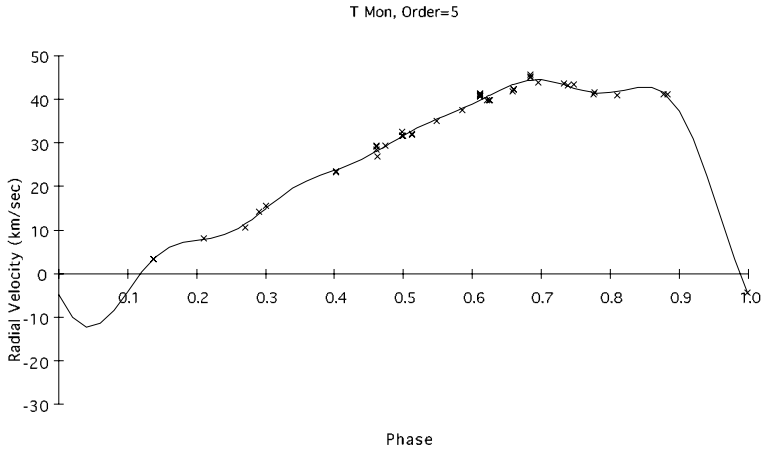

Figure 2. *The radial velocity data for T Mon fitted with a fifth-order trigonometric polynomial. This fit seems quite adequate to the data, including the fit to the "glitch" of Figure 1.*

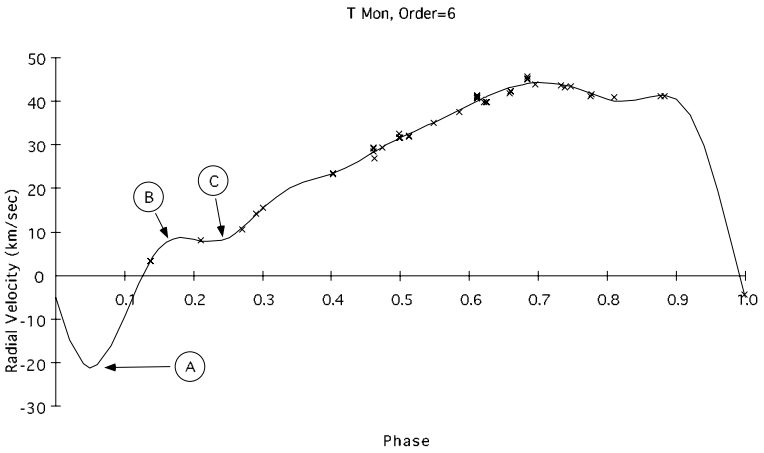

Figure 3. *The radial velocity data for T Mon fitted with a sixth-order trigonometric polynomial. This fit is not clearly better than the fit of Figure 2, and shows some evidence of overfitting, as indicated by the arrows $A - C$; these bumps are not supported by any data (*cf. *Figure 2). Bump A, in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

3. PRIORS

The posterior inference is summarized in the posterior distribution on the unknown parameters

(1) The orders of the Fourier models, M and N.

(2) The precision parameters τ_U, τ_V, τ_C.

(3) The angular diameter ϕ_0 and the unknown phase error $\Delta\theta$.

(4) The distance s.

(5) The intercepts u_0 and v_0 and the Fourier coefficients $\mathbf{a}_u$, $\mathbf{a}_v$.

Indirectly the posterior distribution is also a function of the hyper parameters which define the probability model.

We expect the order of the models to be modest; we choose a uniform prior on the models (M, N) up to some cut-off, and zero beyond.

The precision parameters τ_U, τ_V, τ_C are given standard Jeffreys priors. Probably we could give them more informative priors but it didn't seem necessary in this case. So

$$p(\tau_U) \propto 1/\tau_U,\ p(\tau_V) \propto 1/\tau_V,\ p(\tau_C) \propto 1/\tau_C$$

We take the priors on $\Delta\theta$ and ϕ_0 to be flat. They are well-determined by the data and we have no real prior information that would override the data.

Failure to take the spatial distribution of the stars into account would result in the so-called $Lutz\text{-}Kelker\ bias$, which is a bias in the estimated distance. Astronomers realized this only fairly recently (but it is obvious to a Bayesian since it is a consequence of a clearly inappropriate (flat) prior on the distance). The spatial distribution of Cepheid variables is known to be flattened with respect to the galactic plane. We choose a spatial distribution of stars that is exponentially stratified as we go away from the galactic plane. We adopted a scale height $z_0 = 97 \pm 7$ parsecs (1 parsec=3×10^{13} km), and sampled z_0. Our prior on the distance looks like

$$p(s) \propto \rho(s) s^2 ds,$$

where $\rho(s)$ is the spatial density of stars:

$$\rho(s) \propto \exp\left(-|z|/z_0\right)$$

with $z = s \sin\phi$, and ϕ is the galactic latitude of the star (its angle above the galactic plane).

The constant terms u_0 and v_0 get a flat prior. Unlike the terms in sines and cosines, which represent the physics of the pulsations, they are just intercepts reflecting an arbitrary choice of coordinates. The priors on the periodic Fourier coefficients $\mathbf{a}_u$ and $\mathbf{a}_v$ must be chosen carefully. If our prior is too vague, significant terms may be rejected, but if it is too sharp, overfitting may result. For our models we have used a Zellner G-prior, which is equivalent to a maximum entropy prior (Gull 1988), of the form

$$p(\mathbf{a}) \propto \exp\left(-\frac{\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a}}{2}\tau\right)$$

where $\mathbf{a}$ is the vector of Fourier coefficients, $\mathbf{X}$ is the design matrix of sines and cosines for the problem, and τ is a hyperparameter.

The hyperparameters τ also need priors. Since they are scale parameters, one might naively put a Jeffreys prior on these; however, the resulting posterior distribution would be improper (Gull 1988), so a slight adjustment is required. Thus, we pick a prior on τ of the form

$$p(\tau_a) \propto \frac{1}{\tau^{3/2}}$$

4. SAMPLING STRATEGY

We employ a reversible-jump MCMC algorithm to generate posterior distributions and estimates. We use ideas outlined by Dellaportas *et al.* (1997) in their excellent tutorial paper on the subject.

Fortunately, the full conditional distributions for the precision parameters and the hyperparameters are standard χ^2 distributions and so the sampling for these parameters can be accomplished with straightforward Gibbs sampling, that is, these parameters are updated by draws from the respective complete conditional posterior distributions.

For $\Delta\theta$, ϕ_0 and s, we use a random-walk Metropolis algorithm, using as our proposal a multinormal distribution centered on the currently imputed parameter values, with a variance-covariance matrix that is proportional to the variance-covariance matrix for the

linearized least-squares problem for just these three parameters. This means linearizing the logarithm in the expression for c_i (Eq. 2). The idea behind this strategy is that we'll take longer steps in directions with larger variances and shorter steps in directions with smaller variances, while obtaining good sampling in directions that are not parallel to the axes defined by the parameters. This turns out to have very good acceptance-rejection probabilities and good sampling of the parameter space for these parameters.

The sampling for $\mathbf{a}_u$ and $\mathbf{a}_v$ is more direct. We base our proposal for a Metropolis step on the solution of the linear least squares problems generated by

$$\mathbf{U} \sim \mathbf{N}(u_0 + \mathbf{X}_u \mathbf{a}_u, \sigma_U^2/\tau_U)$$
$$\mathbf{V} \sim \mathbf{N}(v_0 + \mathbf{X}_v \mathbf{a}_v, \sigma_V^2/\tau_V)$$

This results in a near-Gibbs sampler for these parameters. It isn't quite Gibbs because of the nonlinear way in which $\mathbf{a}_u$ and $\mathbf{a}_v$ appear in the full likelihood. However, it is very close; the acceptance probabilities for these proposals are over 90%, and the sampling of the Fourier parameter space is very effective.

The steps in $\mathbf{a}_u$ and $\mathbf{a}_v$ are included within a step that proposes a jump between models. Thus, if the current model has a certain number of parameters, we propose a jump to a model with a (possibly different) number of parameters, and simultaneously propose new values for all the Fourier coefficients. To make the sampling efficient, during the burn-in phase we also estimate the posterior probabilites of the individual models. We use this as the basis for the proposal probabilities of new models during the computation phase of the calculation. Thus we will propose models of higher posterior probability with greater frequency.

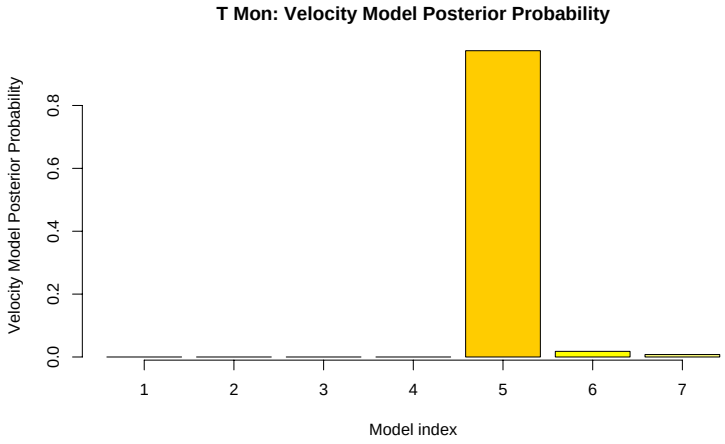

Figure 4. *Posterior marginal distribution of velocity models for T Mon.*

5. RESULTS

As shown in Figure 4, the Bayesian analysis does indeed agree with what our eyes already told us about the velocity model. The fifth-order model is overwhelmingly preferred. The photometry model (Figure 5) shows that the third and fourth-order models are about equally good, so our sampler will average over these two models. Figure 6 shows the posterior distribution of the parallax $\varpi = 1/s$ for T Mon, and Figure 7 shows the simulation history of the parallax, demonstrating good acceptance rates for the proposals.

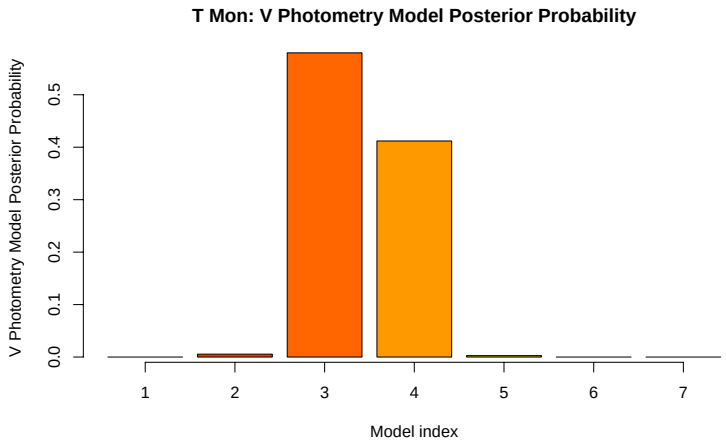

Figure 5. *Posterior marginal distribution of photometry models for T Mon.*

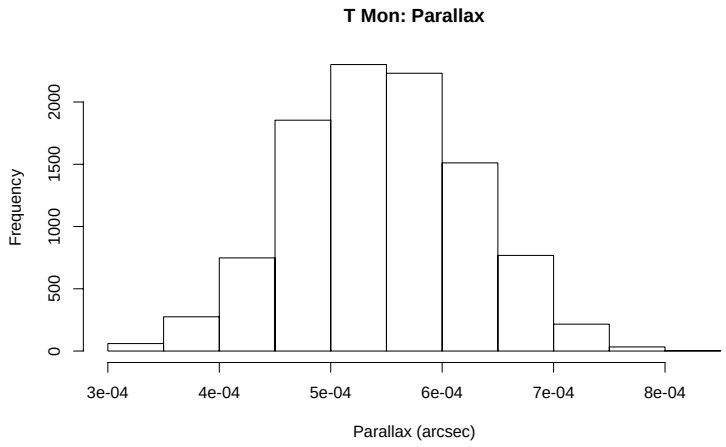

Figure 6. *Posterior marginal distribution of the parallax of T Mon.*

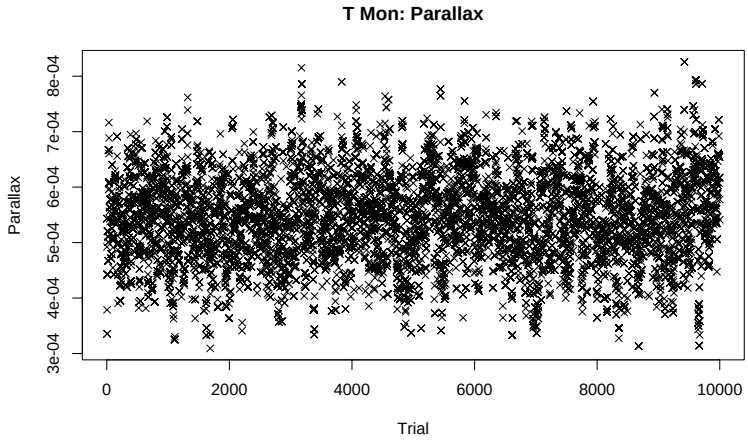

Figure 7. *Simulation history of the parallax of T Mon.*

6. ALTERNATIVE APPROACHES

We are investigating other functional forms to represent the velocity and photometry data. In particular, the velocity curve suffers a steep drop-off near $\theta = 1$, which results in overshoot and "ringing" due to the global nature of Fourier polynomials combined with the Gibbs phenomenon; a more local representation using wavelets looks very promising. We are also investigating other priors on the Fourier coefficients that have been used successfully in other contexts. One promising approach is the "Expected Posterior Prior" developed by Pérez (1998).

Of these approaches, the wavelet analysis is the furthest advanced. We have adopted a suggestion of Vannucci and Corradi (1999). So far we have applied it only to the problem of fitting the velocity curve; application to the full astronomical problem of determining distances is in the future.

Their idea is to put our prior on the *functions* $f(\theta)$ rather than on the wavelet coefficients. We let $f_i = f(\theta_i)$ be defined on an equally spaced grid, and consider the Gaussian process

$$d_i = f_i - f_{i-1}$$

where

$$d = (d_1, d_2, \ldots, d_n) \sim \mathbf{N}(\mathbf{0}, \Delta)$$
$$\Delta_{ij} = \lambda \exp(-\rho|i-j|)$$

This leads to a prior on the f's of the form

$$p(f_0, f_1, \ldots, f_{n-1} \mid f_0 = f_n) = \mathbf{N}(\mathbf{0}, \lambda V)$$

where V has a prescribed form. This in turn induces a multivariate normal prior on the wavelet coefficients with variance-covariance matrix that can be computed from V either explicitly or preferably (as Vannucci and Corradi do) as a bivariate wavelet decomposition of V.

As can be seen from Figure 8, our preliminary results show better behavior around the "overshoot" region near phase 0.05 (*cf.* Figures 2 and 3), and also adequately address the "glitch" near phase 0.8. So our wavelet approach appears promising.

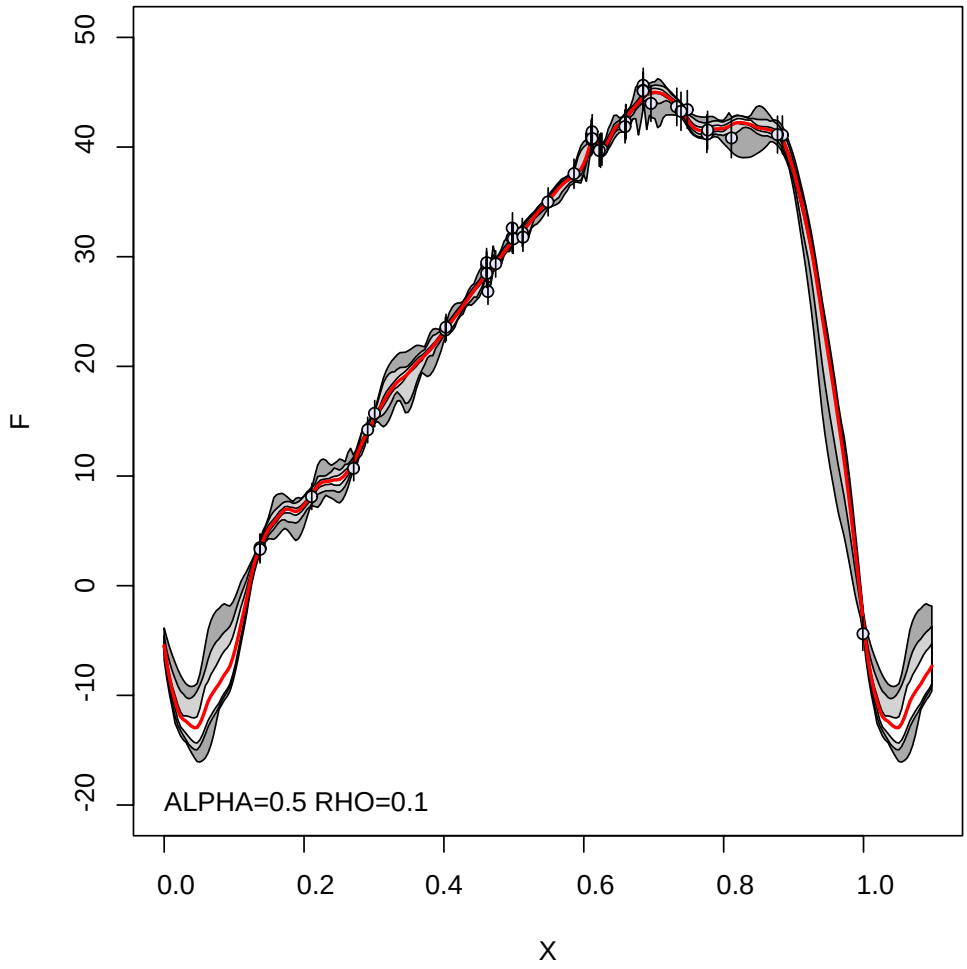

Figure 8. *Wavelet fit of velocity data for T Mon. Shown are contour lines for the posterior distribution of the velocity as a function of phase; the thick smooth line in the center is the posterior mean curve. The grey shaded margins show central 50% (light grey) and central 90% (dark grey) intervals. The points are the observed data points, with little error bars showing 2 standard deviations for the measurement error.*

REFERENCES

Dellaportas, P., Forster, J. and Ntzoufras, I. (1997). On Bayesian model and variable selection using MCMC. Private communication. Available as `http://www.stat-athens.aueb.gr/~ptd/gvs.ps`

Gull, S. (1988). Bayesian inductive inference and maximum entropy. *Maximum-Entropy and Bayesian Methods in Science and Engineering, Volume 1: Foundations.* (G. J. Erickson and C. R. Smith, eds.). Dordrecht: Kluwer, 53–74.

Jefferys, W. H. and Barnes, T. G. (1999). Bayesian analysis of Cepheid variable data. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 777–783.

Pérez, J. M. (1998). *Develoment of Expected Posterior Prior Distributions for Model Comparisons*. Dissertation, Duke University.

Vannucci, M. and Corradi, F. (1999). Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *J. Roy. Statist. Soc. B* **61**, 971–986.