The Astrophysical Journal, 756:67 (12pp), 2012 September 1 doi:10.1088/0004-637X/756/1/67

NONPARAMETRIC BAYESIAN ESTIMATION OF PERIODIC LIGHT CURVES

Yuyang Wang[1], Roni Khardon[1], and Pavlos Protopapas[2,3]
[1] Department of Computer Science, Tufts University, Medford, MA, USA
[2] Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA
[3] Institute for Applied Computational Science, Harvard University, Cambridge, MA, USA
Received 2012 February 27; accepted 2012 June 29; published 2012 August 16

ABSTRACT

Many astronomical phenomena exhibit patterns that have periodic behavior. An important step when analyzing data from such processes is the problem of identifying the period: estimating the period of a periodic function based on noisy observations made at irregularly spaced time points. This problem is still a difficult challenge despite extensive study in different disciplines. This paper makes several contributions toward solving this problem. First, we present a nonparametric Bayesian model for period finding, based on Gaussian Processes (GPs), that does not make assumptions on the shape of the periodic function. As our experiments demonstrate, the new model leads to significantly better results in period estimation especially when the light curve does not exhibit sinusoidal shape. Second, we develop a new algorithm for parameter optimization for GP which is useful when the likelihood function is very sensitive to the parameters with numerous local minima, as in the case of period estimation. The algorithm combines gradient optimization with grid search and incorporates several mechanisms to overcome the high computational complexity of GP. Third, we develop a novel approach for using domain knowledge, in the form of a probabilistic generative model, and incorporate it into the period estimation algorithm. Experimental results validate our approach showing significant improvement over existing methods.

Key words: methods: data analysis – methods: statistical – stars: variables: general

Online-only material: color figures

1. INTRODUCTION

Many astronomical phenomena exhibit periodic behavior. Discovering their period and the periodic pattern they exhibit is an important task toward understanding their behavior. A significant effort has been devoted to the analysis of light curves from periodic variable stars. For example, the top part of Figure 1 shows the magnitude of a light source over time. The periodicity of the light source is not obvious before we fold it. However, as the bottom part illustrates, once folded with the right period we get convincing evidence of periodicity. The object in this figure is classified as an eclipsing binary (EB). Other sources show periodic variability due to processes internal to the star (Petit 1987).

The problem of period estimation from noisy and irregularly sampled observations has been studied before in several disciplines. Most approaches identify the period by some form of grid search. That is, the problem is solved by evaluating a criterion Φ at a set of trial periods $\{p\}$ and selecting the period p that yields the best value for $\Phi(p)$. The commonly used techniques vary in the form and parameterization of Φ, the evaluation of the fit quality between model and data, the set of trial periods searched, and the complexity of the resulting procedures. Two methods we use as baselines in our study are the Lomb–Scargle (LS) periodogram (Scargle 1982; Reimann 1994) and the phase dispersion minimization (PDM; Stellingwerf 1978), both known for their success in empirical studies. The LS method is relatively fast and is equivalent to maximum likelihood estimation under the assumption that the function has a sinusoidal shape. It therefore makes a strong assumption on the shape of the underlying function. On the other hand, PDM makes no such assumptions and is more generally applicable, but it is slower and is less often used in practice. A more extensive discussion of related work is given in Section 5.

The paper makes several contributions toward solving the period estimation problem. First, we present a new model for period finding, based on Gaussian Processes (GPs), that does not make strong assumptions on the shape of the periodic function. In this context, the period is a hyperparameter of the covariance function of the GP and accordingly the period estimation is cast as a model selection problem for the GP. As our experiments demonstrate, the new model leads to significantly better results compared to LS when the target function is non-sinusoidal. The model also significantly outperforms PDM when the sample size is small.

Second, we develop a new algorithm for period estimation within the GP model. In the case of period estimation the likelihood function is not a smooth function of the period parameter. This results in a difficult estimation problem which is not well explored in the GP literature (Rasmussen & Williams 2005). Our algorithm combines gradient optimization with grid search and incorporates several mechanisms to improve the complexity over the naive approach.

In particular we propose and evaluate: an approximation using a two-level grid search, approximation using limited cyclic optimization, a method using sub-sampling and averaging, and a method using low-rank Cholesky approximations. An extensive experimental evaluation using artificial data identifies the most useful approximations and yields a robust algorithm for period finding.

Third, we develop a novel approach for using astrophysics knowledge, in the form of a probabilistic generative model, and incorporate it into the period estimation algorithm. In particular, we propose to employ the generative model to bias the selection of periods by using it as a prior over periods or as a post-processing selection criterion choosing among periods ranked highly by the GP. The resulting algorithm is applied and evaluated on astrophysics data showing significantly improved performance over previous work.

The next section provides some technical background and defines the period estimation problem as GP inference. The following three sections present our algorithm, report on

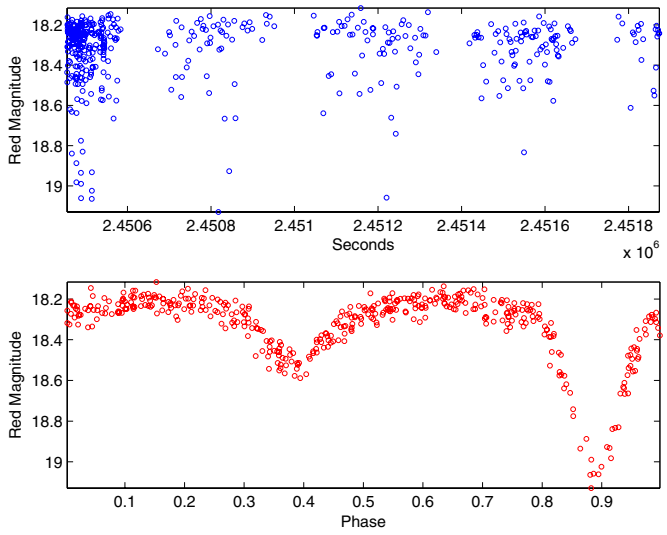

Figure 1. Top: brightness of an eclipsing binary star over time; bottom: brightness vs. phase.

(A color version of this figure is available in the online journal.)

experiments evaluating it and applying it to astrophysics data, and discuss related work. The final section concludes with a summary and directions for future work.

2. PRELIMINARIES: GP FOR PERIOD FINDING

This section provides technical background on GPs and their optimization procedures and defines the period finding problem in this context.

Throughout the paper, scalars are denoted using italics, as in $x, y \in \mathbb{R}$; vectors and matrices use lowercase and capital bold typeface, as in $\boldsymbol{x}$, $\boldsymbol{y}$, $\mathbf{K}$, $\mathbf{A}$; and x_i denotes the ith entry of $\boldsymbol{x}$. For a vector $\boldsymbol{x}$ and real-valued function $f : \mathbb{R} \to \mathbb{R}$, we extend the notation for f to vectors so that $f(\boldsymbol{x}) = [f(x_1), \ldots, f(x_n)]^{\mathrm{T}}$, where the superscript T stands for transposition. $\mathbb{I}$ is the identity matrix.

2.1. Gaussian Processes

This section gives a brief review of GP regression. A more extensive introduction can be found in Rasmussen & Williams (2005) and Bishop (2006).

We start with the following regression model:

$$y = f_{\boldsymbol{w}}(\boldsymbol{x}) + \epsilon, \tag{1}$$

where $f_{\boldsymbol{w}}(x)$ is the regression function with parameter $\boldsymbol{w}$ and ϵ is iid Gaussian noise. For example, in linear regression $f_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ and therefore $y \sim N(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}, 1/\sigma^2)$. Given the data $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}, i = 1, \ldots, N$, one wishes to infer $\boldsymbol{w}$ and the basic approach is to maximize the likelihood $\mathcal{L}(\boldsymbol{w}, \mathcal{D}) = \Pr(\mathcal{D}|\boldsymbol{w})$.

In Bayesian statistics, the parameter $\boldsymbol{w}$ is assumed to have a prior probability $\Pr(\boldsymbol{w})$ which encodes the prior belief on the parameter. The inference task becomes calculating the posterior distribution over $\boldsymbol{w}$, which, using the Bayesian formula, is given as

$$\Pr(\boldsymbol{w}|\mathcal{D}) \propto \Pr(\mathcal{D}|\boldsymbol{w}) \Pr(\boldsymbol{w}). \tag{2}$$

The predictive distribution for a new observation $\boldsymbol{x}^*$ is given by

$$\Pr(f(\boldsymbol{x}^*)|\mathcal{D}) = \int \Pr(f(\boldsymbol{x}^*)|\boldsymbol{w}) \Pr(\boldsymbol{w}|\mathcal{D}) d\boldsymbol{w}. \tag{3}$$

Returning to linear regression, the common model assumes that the prior for $\boldsymbol{w}$ is a zero-mean multivariate Gaussian distribution, and the posterior turns out to be multivariate Gaussian as well. In contrast with many Bayesian formulations, the use of GP often allows for simple inference or calculation of desired quantities because of properties of multivariate Gaussian distributions and corresponding facts from linear algebra.

This approach can be made more general using a nonparametric Bayesian model. In this case, we replace the parametric latent function $f_{\boldsymbol{w}}$ by a stochastic process f where f's prior is given by a GP. A GP is specified by a mean function (assumed to be zero in this paper) and covariance function $\mathcal{K}(\cdot, \cdot)$. This allows us to specify a prior over functions f such that the distribution induced by the GP over any finite sample is normally distributed. More precisely, the GP regression model with zero mean and covariance function $\mathcal{K}(\cdot, \cdot)$ is as follows. Given sample points $[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^{\mathrm{T}}$ let $\mathbf{K} = (\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j}$. The induced distribution on the values of the function at the sampling points is

$$\mathbf{f} \triangleq [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)]^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \tag{4}$$

where $\mathcal{N}$ denotes the multivariate normal distribution. Now assuming that y_i is generated from $f(\boldsymbol{x}_i)$, using iid noise as in Equation (1), and denoting $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ we get that $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K} + \sigma^2\mathbb{I})$ and the joint distribution is given by

$$\begin{bmatrix} \mathbf{f} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{K} \\ \mathbf{K} & \mathbf{K} + \sigma^2\mathbb{I} \end{bmatrix} \right). \tag{5}$$

Using properties of multivariate Gaussians we can see that the posterior distribution $\mathbf{f}|\boldsymbol{y}$ is given by

$$\Pr(\mathbf{f}|\mathcal{D}) = \mathcal{N}(\mathbf{K}(\sigma^2\mathbb{I} + \mathbf{K})^{-1}\boldsymbol{y}, \quad \sigma^2(\sigma^2\mathbb{I} + \mathbf{K})^{-1}\mathbf{K}). \tag{6}$$

Similarly, the predictive distribution for some test point $\boldsymbol{x}_*$ distinct from the training examples is given by

$$\begin{aligned} \Pr(f(\boldsymbol{x}_*)|\boldsymbol{x}_*, \mathcal{D}) &= \int \Pr(f(\boldsymbol{x}_*)|\boldsymbol{x}_*, f) \Pr(f|\mathcal{D}) df \\ &= \mathcal{N}(\mathbf{k}(\boldsymbol{x}_*)^{\mathrm{T}}(\sigma^2\mathbb{I} + \mathbf{K})^{-1}\boldsymbol{y}, \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) \\ &\quad - \mathbf{k}(\boldsymbol{x}_*)^{\mathrm{T}}(\sigma^2\mathbb{I} + \mathbf{K})^{-1}\mathbf{k}(\boldsymbol{x}_*)), \end{aligned} \tag{7}$$

where $\mathbf{k}(\boldsymbol{x}_*) = [\mathcal{K}(\boldsymbol{x}_1, \boldsymbol{x}_*), \ldots, \mathcal{K}(\boldsymbol{x}_N, \boldsymbol{x}_*)]^{\mathrm{T}}$.

Figure 2 illustrates GP regression, by showing how a finite sample induces a posterior over functions and their values for new sample points.

2.2. Problem Definition

In the case of period estimation the sample points $\boldsymbol{x}_i$ are scalars x_i representing the corresponding time points, and we denote $\boldsymbol{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$. The underlying function $f(\cdot)$ is periodic with unknown period p and corresponding frequency $w = 1/p$. To model the periodic aspect we use a GP with a periodic covariance function

$$\mathcal{K}_{\boldsymbol{\theta}}(x_i, x_j) = \beta \exp\left\{ -\frac{2\sin^2(w\pi(x_i - x_j))}{\ell^2} \right\}, \tag{8}$$

where the set of hyperparameters[4] of the covariance function is given by $\boldsymbol{\theta} = \{\beta, w, \ell\}$. It can be easily seen that any f generated

[4] Typically, in a hierarchical model, the parameters of the top level (e.g., parameters of the prior) that affect the next level are called hyperparameters. In GP regression, the parameter is the regression function f and the hyperparameters are the parameters of the covariance function.

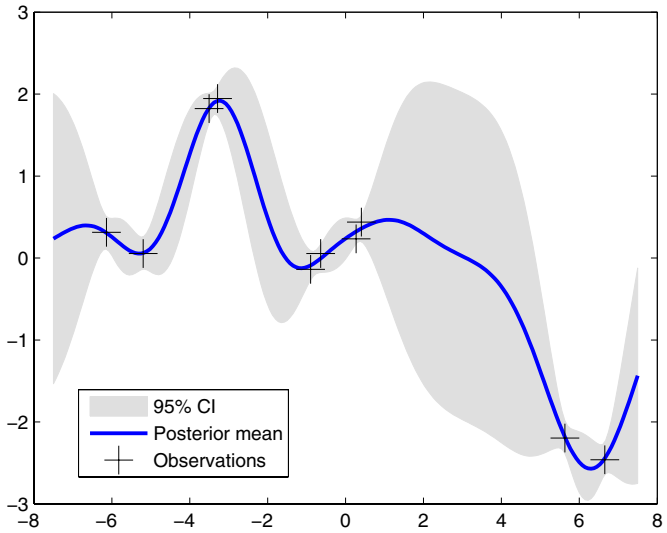

Figure 2. Illustration of prediction with GP regression. The data points $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}$ are given by the crosses. The shaded area represents the pointwise 95% confidence region of the predictive distribution. As can be seen from Equation (7), GP regression can be seen to perform a variant of kernel regression where $f(\boldsymbol{x}_*)$ is a weighted average of all the measurements $\boldsymbol{y}$. While the values of the weights are obscured because of the inverse of the covariance matrix in that expression, one can view this roughly by an analogy to nearest neighbor regression where the mean of $f(\boldsymbol{x}^*)$ is affected more by the measurements whose sampling points are close to $\boldsymbol{x}^*$ and the variance of $f(\boldsymbol{x}^*)$ is small if $\boldsymbol{x}^*$ is surrounded by measurements. A deeper discussion of the equivalent kernel is given in Rasmussen & Williams (2005).

(A color version of this figure is available in the online journal.)

by $\mathcal{K}_{\boldsymbol{\theta}}$ is periodic with period $1/w$. Figure 3 illustrates the role of the other two hyperparameters. We can see that β controls the magnitude of the sampled functions. At the same time, ℓ which is called characteristic length determines how sharp the variation is between two points. The plots also demonstrate that the shape of the periodic functions is highly variable. If desired, other base kernels (Rasmussen & Williams 2005) can be used and made to be periodic in a similar manner, and as in other work it is easy to add a "trend" to the data to capture functions that are not purely periodic. In this paper, we focus on period finding with the purely periodic kernel and leave such extensions to future work.

In our problem each star has its own period and shape and therefore each has its own set of hyperparameters. Our model, thus, assumes that the following generative process is the one producing the data. For each time series j with arbitrary sample points $\boldsymbol{x}^j = [x_1^j, \ldots, x_{N_j}^j]^{\mathrm{T}}$, we first draw

$$f_j|\theta_j \sim \mathcal{GP}(0, \mathcal{K}_{\boldsymbol{\theta}_j}). \tag{9}$$

Then, given $\boldsymbol{x}^j$ and f_j we sample the observations

$$\boldsymbol{y}^j \sim \mathcal{N}(f_j(\boldsymbol{x}^j), \sigma^2\mathbb{I}). \tag{10}$$

Denote the complete set of parameters by $\mathcal{M} = \{\boldsymbol{\theta}, \sigma^2\}$. For each time series j, the inference task is to select the correct model for the data $\{\boldsymbol{x}^j, \boldsymbol{y}^j\}$, that is, to find $\mathcal{M}$ that best describes the data. This is the main computational problem studied in this paper. The next subsection reviews two standard approaches for this problem.

Before presenting these we clarify two methodological issues. First, notice that our model assumes homogeneous noise $\mathcal{N}(0, \sigma^2)$, i.e., the observation error for each x_i is the same. Experimental results on the OGLEII data set (not shown here) show that σ^2 estimated from the data is very close to the mean of the recorded observation errors, and therefore there is no advantage in explicitly modeling the recorded observation errors. Of course, this may be different in other surveys; incorporating observation errors can be easily done by using $\sigma_{\mathrm{obs}}^2 + \sigma^2$ in Equation (10).

Second, as defined above, our task is to find the full set of parameters $\mathcal{M}$. Therefore, our framework and induced algorithms can estimate the underlying function, f, through the posterior mean $\hat{f}$, and thus yield a solution for the regression problem—predicting the value of the function at unseen sample points. However, our main goal and interest in solving the problem is to infer the frequency w where the other parameters are less important. Therefore, a large part of the evaluation in the paper focuses on accuracy in identifying the frequency, although we also report results on prediction accuracy for the regression problem.

2.3. Model Selection

2.3.1. Marginal Likelihood

The standard Bayesian approach is to identify the hyperparameters that maximize the marginal likelihood. More precisely, we try to find $\mathcal{M}^*$ such that

$$\mathcal{M}^* = \underset{\mathcal{M}}{\operatorname{argmax}} \left[\log \left[\Pr(\boldsymbol{y}|\boldsymbol{x}; \mathcal{M})\right]\right], \tag{11}$$

where the marginal likelihood is given by

$$\begin{aligned} \log \Pr(\boldsymbol{y}|\boldsymbol{x}; \mathcal{M}) &= \log \left(\int \Pr(\boldsymbol{y}|f, \boldsymbol{x}; \mathcal{M}) \Pr(f|\boldsymbol{x}; \mathcal{M})\, df \right) \\ &= -\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{K} + \sigma^2\mathbb{I})^{-1}\boldsymbol{y} \\ &\quad - \frac{1}{2}\log|\boldsymbol{K} + \sigma^2\mathbb{I}|^{-1} - \frac{n}{2}\log 2\pi \end{aligned} \tag{12}$$

and Equation (12) holds because $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K} + \sigma^2\mathbb{I})$ (Rasmussen & Williams 2005). Typically, one can optimize the marginal likelihood by calculating the partial derivative of the marginal likelihood w.r.t. the hyperparameters and optimizing the hyperparameters using gradient-based search (Rasmussen & Williams 2005). As we show below, gradients alone cannot be used to solve our problem completely and therefore our algorithm elaborates and improves over this approach. We do, however, use the conjugate gradients optimization as a basic step in our algorithm. The partial derivative of Equation (12) w.r.t. the parameter θ_j is (Rasmussen & Williams 2005)

$$\frac{\partial}{\partial\theta_j}\log \Pr(\boldsymbol{y}|\boldsymbol{x}; \mathcal{M}) = \mathrm{Tr}\left(\left(\alpha\alpha^{\mathrm{T}} - \boldsymbol{K}_\sigma^{-1}\right)\frac{\partial \boldsymbol{K}_\sigma}{\partial\theta_j}\right), \tag{13}$$

where $\boldsymbol{K}_\sigma = \boldsymbol{K} + \sigma^2\mathbb{I}$ and $\alpha = \boldsymbol{K}_\sigma^{-1}\boldsymbol{y}$.

2.3.2. Cross-validation

An alternative approach (Rasmussen & Williams 2005) picks hyperparameter $\mathcal{M}$ by minimizing the empirical loss on a hold out set. This is typically done with a leave-one-out (LOO) formulation, which uses a single observation from the original sample as the validation data, and the remaining observations as the training data. The process is repeated such that each

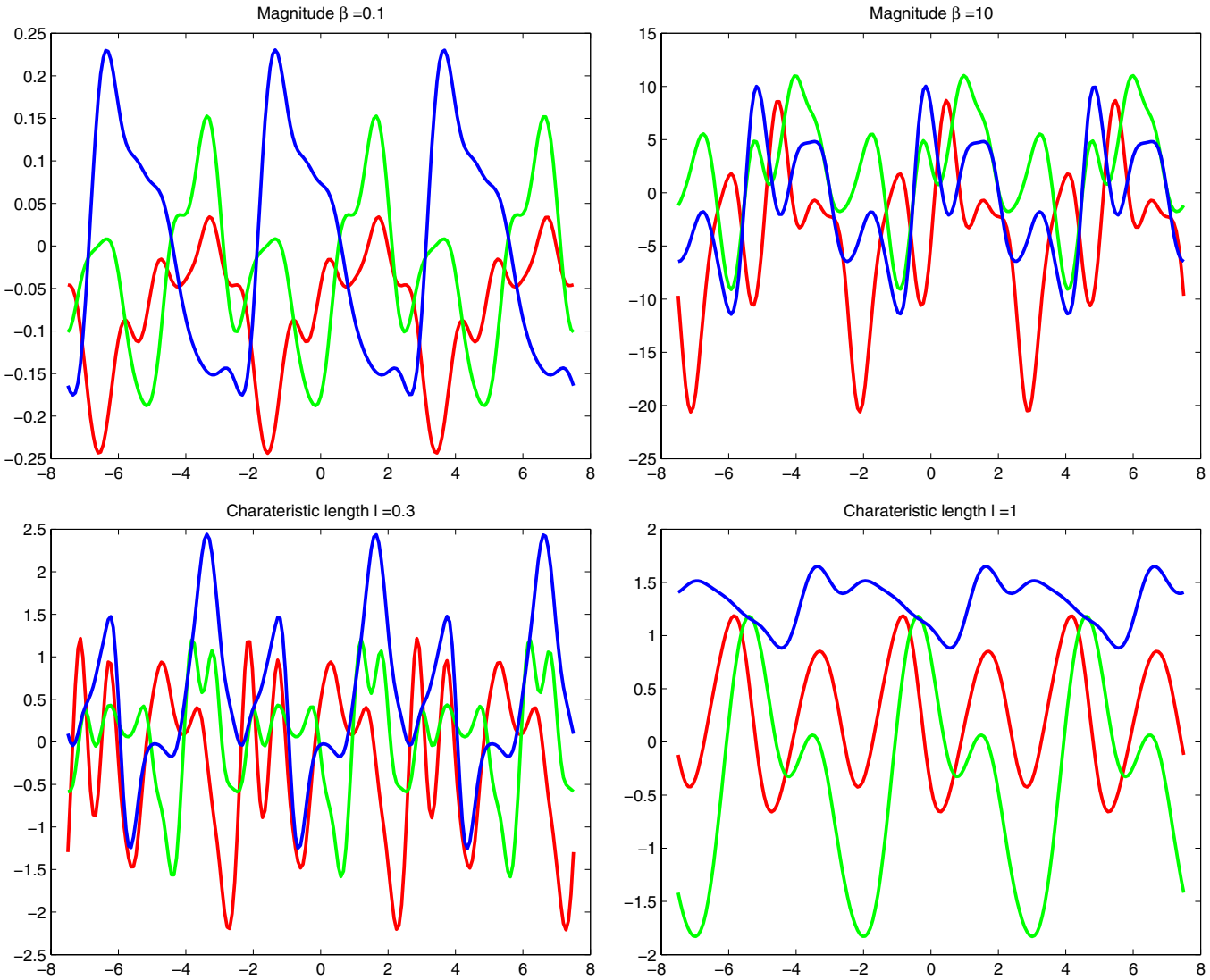

Figure 3. Sample functions from a GP with covariance function in Equation (8) where the period is fixed to be 5, i.e., $w = 0.2$. Top row: $\beta = 0.1$ vs. $\beta = 10$ while ℓ is fixed to be 0.6. Bottom row: $\ell = 0.3$ vs. $\ell = 1$ with $\beta = 0.3$.

(A color version of this figure is available in the online journal.)

observation in the sample is used once as the validation data. To be precise, we choose the hyperparameter $\mathcal{M}^*$ such that

$$\mathcal{M}^* = \underset{\mathcal{M}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{f}_{-i}(x_i))^2, \tag{14}$$

where $\hat{f}_{-i}$ is defined as the posterior mean given the data $\{\boldsymbol{x}_{-i}, \boldsymbol{y}_{-i}\}$ in which the subscript $-i$ means all but the ith sample, that is,

$$\hat{f}_{-i}(x) = \mathcal{K}(\boldsymbol{x}_{-i}, x)^{\mathrm{T}} (\boldsymbol{K}_{-i} + \sigma^2 \mathbb{I})^{-1} \boldsymbol{y}_{-i}. \tag{15}$$

It can be shown that this computation can be simplified (Rasmussen & Williams 2005) using the fact that

$$y_i - \hat{f}_{-i}(x_i) = \frac{[(\boldsymbol{K} + \sigma^2 \mathbb{I})^{-1} \boldsymbol{y}]_i}{[(\boldsymbol{K} + \sigma^2 \mathbb{I})^{-1}]_{ii}}, \tag{16}$$

where $[\cdot]_i$ is the ith entry of the vector and $[\cdot]_{ii}$ denotes the (i, i)th entry of the matrix.

3. ALGORITHM

We start by demonstrating experimentally that gradient-based methods are not sufficient for period estimation. We generate synthetic data and maximize the marginal likelihood w.r.t. $\boldsymbol{\theta} = \{\beta, w, \ell\}$ using conjugate gradients. For this experiment, 30 samples in the interval $[-10, 10]$ are generated according to the periodic covariance function in Equation (8) with $\boldsymbol{\theta} = [1, 0.25, 1]$. Fixing β, ℓ to their correct values, the marginal likelihood w.r.t. the period $1/w$ is shown in Figure 4 (left). The figure shows that the marginal likelihood has numerous local minima in the high frequency (small period) region that have no relation to the true period. Figure 4 (right) shows two functions with the learned parameters based on different starting points (initial values).

The function plotted in dark color estimates the true function correctly while the one in light color does not. This is not surprising because from Figure 4 (left), we can see that there is only a small region of initial points from which the algorithm can find the correct period. We repeated this experiment using several other periodic functions with similar results. These preliminary experiments illustrate two points.

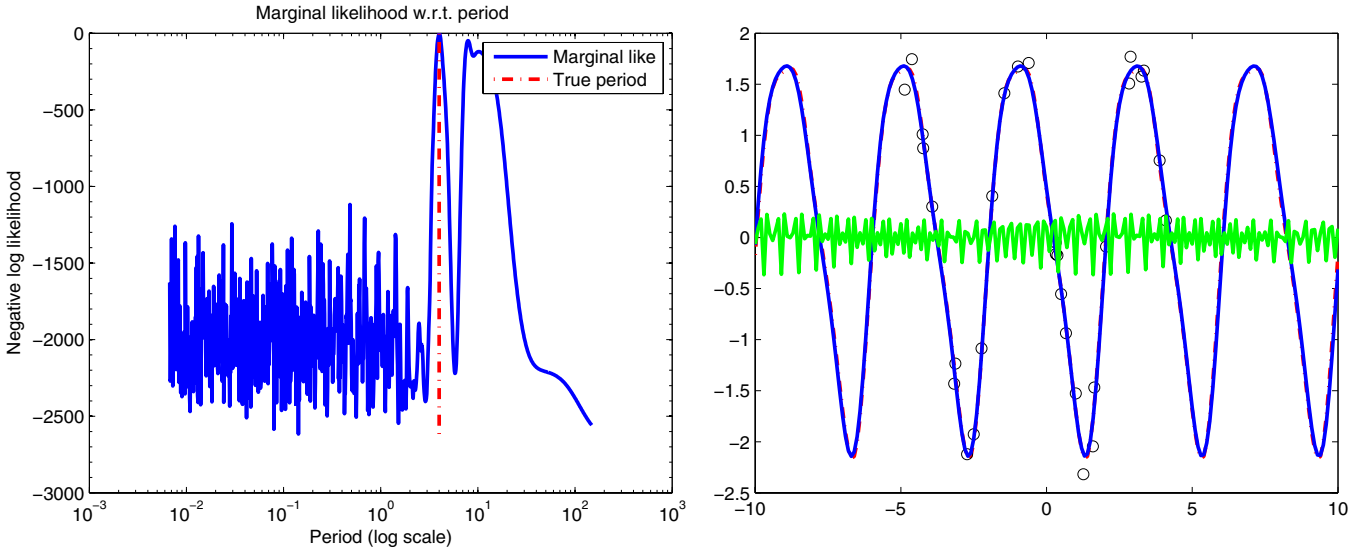

Figure 4. Illustration of sensitivity of the marginal likelihood. A light curve is generated using the GP model with parameters $\beta = 1$, $w = 0.25$, and $\ell = 1$. Left: the marginal likelihood function vs. the period, where the dotted line indicates the true period. Right: the black circles are the observations and the dotted line (covered by the dark estimated curve) is the true function. The dark line which covers the true curve and the light line are the learned regression functions given two different starting points of w.

(A color version of this figure is available in the online journal.)

1: Initialize the parameters randomly.
2: **repeat**
3: Jointly find $\tilde{w}, \beta^*, \ell^*, \sigma^*$ that maximize Equation (12) using conjugate gradients.
4: **for all** w in a coarse grid set $\mathcal{C}$ **do**
5: Calculate the marginal likelihood Equation (12) or the LOO Error Equation (14) using $\beta^*, \ell^*, \sigma^*$.
6: **end for**
7: Set w to the best value found in the **for** loop.
8: **until** Number of iterations reaches L_1 ($L_1 = 2$ by default)
9: Record the Top K ($K = 10$ by default) frequencies $\mathcal{W}^*$ found in the last run of **for** loop (lines 4-6).
10: **repeat**
11: Jointly find $\tilde{w}, \beta^*, \ell^*, \sigma^*$ that maximize Equation (12) using conjugate gradients.
12: **for all** w in a fine grid set $\mathcal{F}$ that covers $\mathcal{W}^*$ **do**
13: Calculate the marginal likelihood Equation (12) or the LOO Error Equation (14) using $\beta^*, \ell^*, \sigma^*$.
14: **end for**
15: Set w to the best value found in the **for** loop.
16: **until** Number of iterations reaches L_2 ($L_2 = 2$ by default)
17: Output the frequency w^* that maximizes the marginal likelihood or minimizes the LOO Error in the last run of **for** loop (lines 11-13).

Figure 5. Hyperparameter Optimization Algorithm.

1. At least for the simple data in this experiment, and when other parameters are known, the marginal likelihood function is maximized at the correct period. This shows that in principle we can find the correct period by optimizing the marginal likelihood. In practice, the region around the maximum may be very narrow, we have to deal with multiples of the correct period, and account for possibly very small periods and the problem is not so easy.
2. On the other hand, the plots clearly show that it is not possible to identify the period using only gradient-based search.

Therefore, as in previous work (Reimann 1994; Hall et al. 2000), our algorithm uses grid search for the frequency. The grid used for the search must be sufficiently fine to detect the correct frequency and this implies high computational complexity. We therefore follow a two-level grid search for frequency where the coarse grid must intersect the smooth region of the true maximum and the fine grid can search for the maximum itself. The two-level search significantly reduces the computational cost. Our algorithm presented in Figure 5 combines this with gradient-based optimization of the other parameters. There are several points that deserve further discussion, as follows.

1. In step 3, we can successfully maximize the marginal likelihood w.r.t. β, ℓ, and σ^2 using the conjugate gradients method, but this approach does not work for the frequency w. The reason is that the objective function is highly sensitive w.r.t. w and the gradient is not useful for finding the global maximum. This property justifies the structure of our algorithm. This issue has been observed before and grid search (in particular using two stages) is known to be the most effective solution (Reimann 1994; Hall et al. 2000).
2. Our algorithm uses cyclic optimization estimating w, σ, β, and ℓ. That is to say, we fix other parameters σ, β, ℓ and optimize w and then optimize σ, β, ℓ when w is fixed. We keep doing this iteratively but use a small number of iterations (in our experiments, the default number of iterations is 2). A more complete algorithm would iterate until convergence but this incurs a large computational cost. Our experiments demonstrate that a small number of iterations is sufficient.

3. In steps 3 and 11, we incorporate w into the joint optimization of the marginal likelihood. This yields better results than optimizing w.r.t. the other parameters with fixed w. This shows that the gradient of w sometimes still provides useful information locally, although the obtained optimal value $\tilde{w}$ is discarded.
4. We use an adaptive search in the frequency domain, where at the first stage we use a coarse grid and later a fine grid search is performed at the neighbors of the best frequencies previously found. By doing this, the computational cost is dramatically reduced while the accuracy of the algorithm is still guaranteed.
5. Two possible improvements to the algorithm that might appear useful are less effective than our algorithm. First, in the coarse grid search, optimizing β, ℓ, and σ^2 for each w separately is too expensive because each computation of the gradient requires costly inversion of the kernel matrix. Second, one might be tempted to replace the fine grid search with a gradient-based search for the optimal w. Our experiments on OGLEII (not reported here) show that this routine is inferior both in accuracy and in time complexity. This suggests that the region around the maximum is very narrow in many cases and shows that gradient search is expensive in this problem.

Two additional approximations are introduced next, specifically targeting the coarse and fine grids respectively and using observations that are appropriate in each case.

3.1. Ensemble Subsampling

The coarse grid search in lines 4–6 of the algorithm needs to compute the covariance matrix w.r.t. each frequency in $\mathcal{C}$ and invert the corresponding covariance matrix, and therefore the total time complexity is $\mathcal{O}(|\mathcal{C}|N^3)$. In addition, different stars do not share the same sampling points.[5] Therefore, the covariance matrix and its inverse cannot be cached to be used on all stars. The computational cost is too high when the coarse grid has a large cardinality. Our observation here is that it might suffice to get an approximation of the likelihood at this stage of the algorithm, because additional fine grid search is done in the next stage.

Therefore, to reduce the time complexity, we propose an ensemble approach that combines the marginal likelihood of several sub-sampled times series. The idea (Protopapas et al. 2005) is that the correct period will get a high score for all sub-samples, but wrong periods that might score well on some sub-samples (and be preferred to others due to outliers) will not score well on all of them and will thus not be chosen. For the approximation, we sub-sample the original time series such that it only contains a fraction f of the original time points, repeating the process R times. The marginal likelihood score is the average over the R repetitions. Our experiments over the synthetic data set justify using $f = 15\%$ and $R = 10$. For OGLEII we constrain this to have at least 30 points (to maintain minimal accuracy) and at most 40 points (to limit complexity). This approximation reduces the time complexity to $\mathcal{O}(|\mathcal{C}| \times R \times (fN)^3)$.

[5] When multiple time series have the same sampling points (as might be the case with a whole field in a survey) we can store the values of the kernel matrices and their inverses (per setting of w, β, and l) and reuse these. This has the potential to significantly reduce the time complexity of the algorithm.

3.2. First-order Approximation with Low-rank Approximation

Similar to the previous case, the time complexity of fine grid search is $\mathcal{O}(|\mathcal{F}|N^3)$. In this case, we can reduce the constant factor in the $\mathcal{O}(N^3)$ term. Note that in step 13, other parameters are fixed and the grid is fine so that the marginal likelihood is a smooth function of w. Suppose we have $w_0, w_1 \in \mathcal{F}$ where $\mathcal{F}$ is the fine grid and $\Delta w = |w_0 - w_1| < \epsilon$, where ϵ is a predefined threshold. Then, given $\boldsymbol{K}_{w_0}$, the covariance matrix w.r.t. w_0, we can get $\boldsymbol{K}_{w_1}$ by its Taylor expansion as

$$\boldsymbol{K}_{w_1} = \boldsymbol{K}_{w_0} + \frac{\partial \boldsymbol{K}}{\partial w}(w_0)\Delta w + o(\epsilon^2). \tag{17}$$

Denote $\widetilde{\boldsymbol{K}} = (\partial \boldsymbol{K}/\partial w)(w_0)$ where $\widetilde{\boldsymbol{K}}\Delta w$ can be seen as a small perturbation to $\boldsymbol{K}_{w_0}$. At first sight, the Sherman–Morrison–Woodbury formula (Bishop 2006) appears to be suitable for calculating the update of the inverse efficiently. Unfortunately, preliminary experiments (not shown here) indicated that this method fails due to numeric instability. Instead, we use an update for the Cholesky factors of the matrix and calculate the inverse through these. Namely, given the Cholesky decomposition of $\boldsymbol{K}_{w_0} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$ we calculate $\widetilde{\boldsymbol{L}}$ such that $\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^{\mathrm{T}} = \boldsymbol{K}_{w_0} + \Delta w\widetilde{\boldsymbol{K}} \approx \boldsymbol{K}_{w_1}$. Details of this construction are given in the Appendix.

3.3. Astrophysical Input Improvements

For some cases we may have further information on the type of periodic functions one might expect. We propose to use such information to bias the selection of periods, by using it to induce a prior over periods or as a post-processing selection criterion. The details of these steps are provided in the next section.

4. EXPERIMENTS

This section evaluates the various algorithmic ideas using synthetic and astrophysics data and then applies the algorithm to a different set of light curves. Our implementation of the algorithms makes use of the *gpml* package (Rasmussen & Nickisch 2010).[6]

4.1. Synthetic Data

In this section, we evaluate the performance of several variants of our algorithm, study the effects of its parameters, and compare it to the two most used methods in the literature: the LS periodogram (Lomb 1976) and PDM (Stellingwerf 1978).

The LS method (Lomb 1976) chooses w to maximize the periodogram defined as

$$P_{\mathrm{LS}}(\omega) = \frac{1}{2}\left\{\frac{[\sum y_j \cos(\eta_j)]^2}{\sum \cos^2(\eta_j)} + \frac{[\sum y_j \sin(\eta_j)]^2}{\sum \sin^2(\eta_j)}\right\}, \tag{18}$$

where $\eta_j = \omega(x_j - \tau)$. The phase τ (that depends on ω) is defined as the value satisfying $\tan(2\omega\tau) = \sum \sin(2\omega x_j)/\sum \cos(2\omega x_j)$. As shown by Reimann (1994), LS fits the data with a harmonic model using least squares.

In the PDM method, the period producing the least possible scatter in the derived light curve is chosen. The score for a proposed period can be calculated by folding the light curve using the proposed period, dividing the resulting observation phases into bins, and calculating the local variance within each

[6] http://www.gaussianprocess.org/gpml/code/matlab/doc/

bin, $\sigma^2 = \sum_j (y_j - \bar{y})^2/N - 1$, where $\bar{y}$ is the mean value within the bin and the bin has N samples. The total score is the sum of variances over all the bins. This method has no preference for a particular shape (e.g., sinusoidal) for the curve.

We generate two types of artificial data, referred to as harmonic data and GP data below. For the first, data are sampled from a simple harmonic function,

$$y \sim \mathcal{N}(a \sin(\omega x + \phi_1) + b \cos(\omega x + \phi_2), \sigma^2 \mathbb{I}), \quad (19)$$

where $a, b \sim \text{Uniform}(0, 5), \omega \sim \text{Uniform}(1, 4), \phi_i \sim \mathcal{N}(0, 1)$, and the noise level σ^2 is set to be 0.1. Note that this is the model assumed by LS. For the second, data are sampled from a GP with periodic covariance function in Equation (8). We generate β, ℓ uniformly in $(0, 3]$ and $(0, 3]$, respectively, and the noise level σ^2 is set to be 0.1. The period is drawn from a uniform distribution between (0.5, 2.5]. For each type we generate data under the following configuration. We randomly sampled 50 time series each having 100 time samples in the interval $[-5, 5]$. Then the comparison is performed using sub-samples with size increasing from 10 to 100. This is repeated ten times to generate means and standard deviations in the plots.

The setting of the algorithms is as follows. In our algorithm, we only use one stage grid search. For our algorithm and LS, the lowest frequency $f_{\min}$ to be examined is the inverse of the span of the input data $1/(x_{\max} - x_{\min}) = 1/T$. The highest frequency $f_{\max}$ is N/T. For the grid, the range of frequencies is broken into even segments of $1/8T$. For PDM we set the frequency range to be $[0.02, 5]$ with the frequency increments of 0.001 and the number of bins in the folded period is set to be 15.

For performance measures we consider both "accuracy" in identifying the period and the error of the regression function. For accuracy, we consider an algorithm to correctly find the period if its error is less than 1% of the true period, i.e., $|\hat{p} - p|/p \leqslant 1\%$. Further experiments (not shown here) justify this approach by showing that the accuracies reported are not sensitive to the predefined error threshold.

The results, where our algorithm does not use the sampling and low-rank approximations, are shown in Figure 6 and they support the following observations.

1. As expected, the top left plot shows that LS performs very well on the harmonic data and it outperforms both PDM and our algorithm. This means that if we know that the expected shape is sinusoidal, then LS is the best choice. This confirms the conclusion of other studies. For example, in the problem of detecting periodic genes from irregularly sampled gene expressions (Wentao et al. 2008; Glynn et al. 2006), the periodic time series of interest were exactly sine curves. In this case, studies showed that LS is the most effective of several statistical models.
2. On the other hand, the top right plot shows that our algorithm is significantly better than LS on the GP data, showing that when the curves are non-sinusoidal the new model is indeed useful.
3. The two plots in the top row together show that our algorithm performs significantly better than PDM on both types of data, especially when the number of samples is small.
4. The first two rows show the performance of the cyclic optimization procedure with 1–5 iterations. We clearly see that for these data sets there is little improvement beyond two iterations. The bottom row shows two examples of the learned regression curves using our method with different numbers of iterations. Although one iteration does find the correct period, the reconstruction curves are not accurate. However, here too, there is little improvement beyond two iterations. This shows that for the data tested here two iterations suffice for period estimation and for the regression problem.
5. The performance of marginal likelihood and cross validation is close, with marginal likelihood dominating on the harmonic data and doing slightly worse on GP data.

Table 1
Comparison of GPs: Original, Subsampling, and Subsampling Plus Low Rank Cholesky Update

	Original	Subsampling	Sub + LowR
ACC	0.831 ± 0.033	$\mathbf{0.857 \pm 0.038}$	0.849 ± 0.028
S/TS	518.52 ± 121.49	197.59 ± 14.10	$\mathbf{170.75 \pm 17.93}$

Notes. ACC denotes accuracy and S/TS denotes the running time in seconds per time series.

We next investigate the performance of the speedup techniques. For this we use GP data under the same configuration as the previous experiments. The experiment was repeated 10 times where in each round we generate 100 light curves each having 100 samples but generated from different θs. For the algorithm we used two iterations for cyclic optimization and varied the subsampling size, number of repetitions, and rank of the approximation. Table 1 shows results with our chosen parameter setting using sampling rate of 15%, 10 repetitions, approximation rank $M = \lfloor N/2 \rfloor$, and grid search threshold $\epsilon = 0.005$. We can see that the subsampling technique saves over 60% of the run time while at the same time slightly increasing the accuracy. Low-rank Cholesky approximation leads to an additional 15% decrease in run time, but gives slightly less good performance. Figure 7 plots the performance of the speedup methods under different parameter settings. The figure clearly shows that the chosen setting provides a good tradeoff in terms of performance versus run time.

4.2. Astrophysics Data

In this section, we estimate the periods of unfolded astrophysics time series from the OGLEII survey (Soszynski et al. 2003).

OGLE surveyed the sky over a number of years and has a huge number of light sources. The data we use here are a subset of OGLEII, containing a total of 14087 light curves of periodic variable stars that have previously been identified to be periodic (and thus their period is known) and to be members of one of three types: Cepheids, RR Lyrae, and EB (as illustrated in Figure 8).

We first explore, validate, and develop our algorithm using a subset of OGLEII data and then apply the algorithm to the full OGLEII data[7] except this development set. The OGLE subset is chosen to have 600 time series in total where each category is sampled according to its proportion in the full data set.

4.2.1. Evaluating the General GP Algorithm

The setting for our algorithm is as follows. The grid search ranges are chosen to be appropriate for the application using coarse grid of $[0.02, 5]$ in the frequency domain with the increments of 0.001. The fine grid is a 0.001 neighborhood of the

[7] http://www.cs.tufts.edu/r/ml/index.php?op=data_software

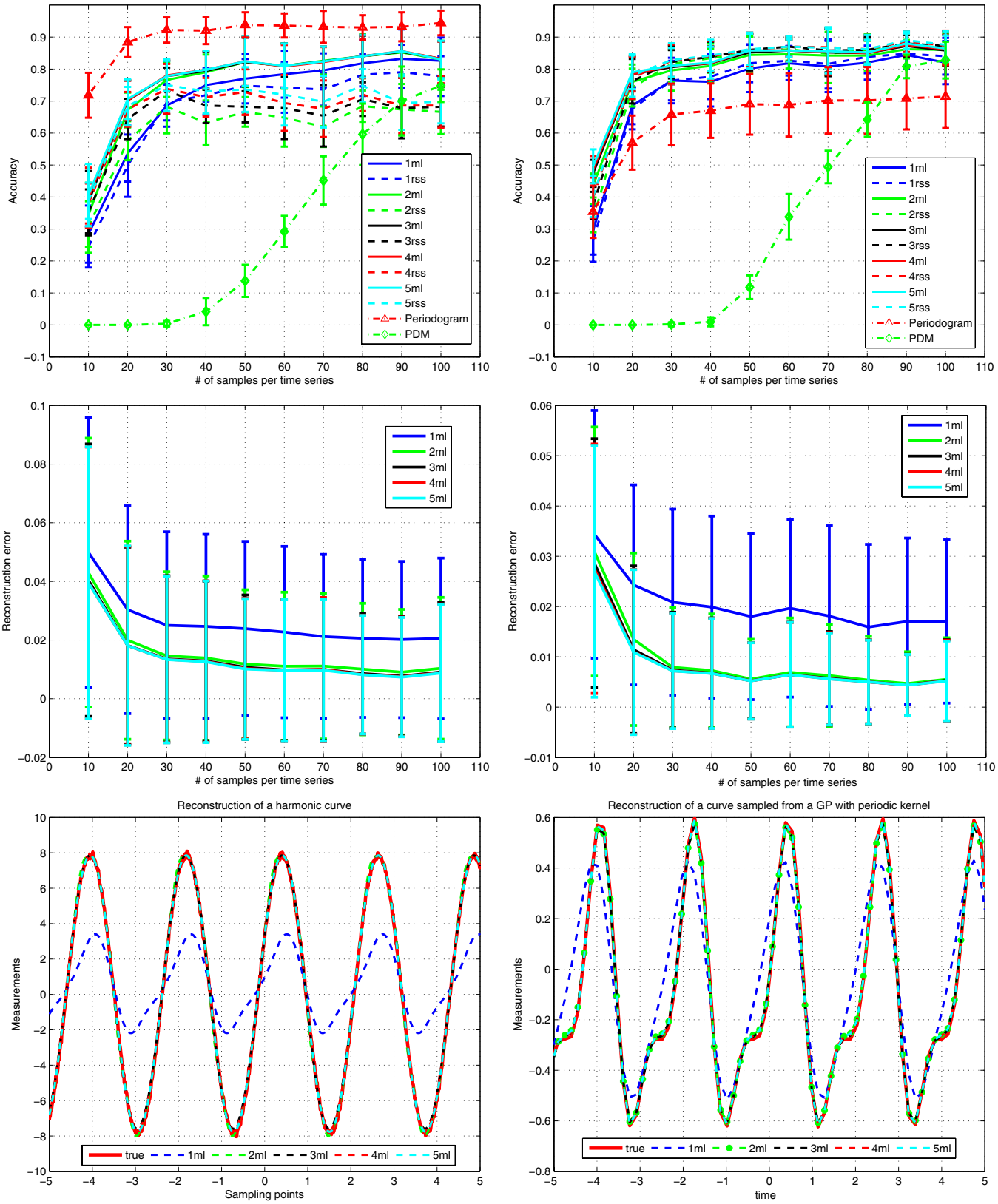

Figure 6. Results for harmonic data (left column) and GP data (right column). Top: accuracy (mean and standard deviations) vs. the number of samples, where solid lines marked with *n*ml represent GP with marginal likelihood where *n* denotes the number of iterations. The corresponding dotted lines marked *n*rss denote cross-validation results with *n* iterations. Middle: reconstruction error for the regression function vs. the number of samples. Bottom: reconstruction curve of GP in two specific runs using maximum likelihood with different numbers of iterations.

(A color version of this figure is available in the online journal.)

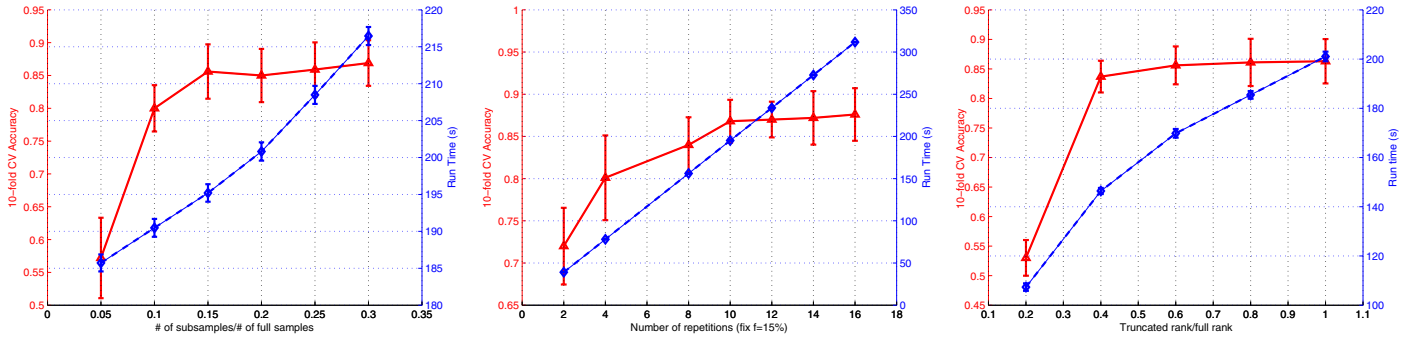

Figure 7. Accuracy (solid line) and Run time (dashed line) of approximation methods as a function of their parameters. Left: sub-sampling ratio (with $R = 10$). Middle: number of repetitions (with 15% sub-sampling). Right: rank in low-rank approximation.

(A color version of this figure is available in the online journal.)

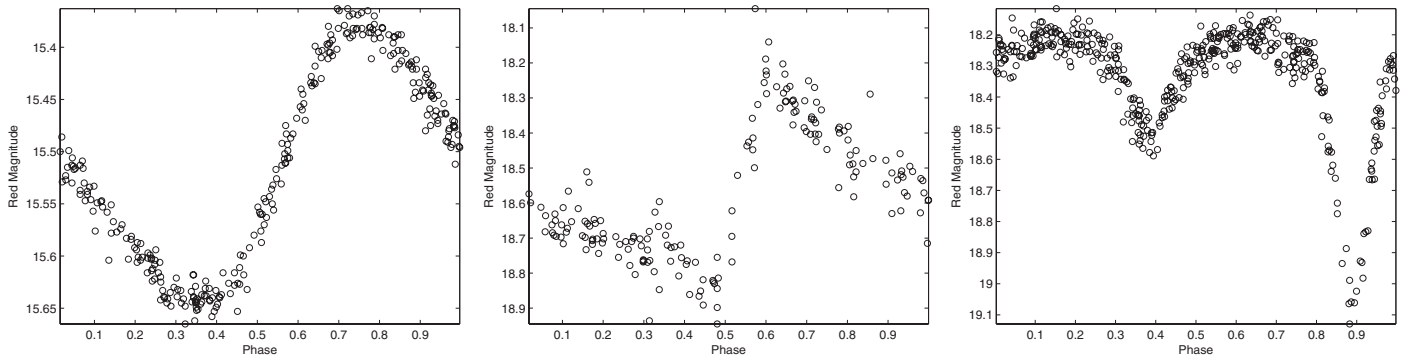

Figure 8. Examples of light curves of periodic variable stars folded according to their period to highlight the periodic shape. Left: Cepheid, middle: RR Lyrae, and right: eclipsing binary.

Table 2
Comparisons of Different GPs on OGLEII Subset

	GP-ML	GP-CV	SGP-ML	SGP-CV	LS
1ITR ACC	0.7856	0.7769	0.7874	0.7808	0.7333
2ITR ACC	0.7892	0.7805	0.7910	0.7818	...
3ITR ACC	0.7928	0.7806	0.7964	0.7845	...
4ITR ACC	0.7946	0.7812	0.7982	0.7875	...
5ITR ACC	0.7964	0.7823	0.8000	0.7906	...

Notes. GP-ML and GP-CV are GP with the ML and CV criteria. SGP-ML and SGP-CV are the corresponding subsampling versions. The first column denotes the number of iterations.

top frequencies each having 20 points with a step of 0.0001. We use $K = 20$ top frequencies in step 9 of the algorithm and vary the number of iterations in a cyclic optimization. When using sub-sampling, we use 15% of the original time series, but restrict sample size to be between 30 and 40 samples. This guarantees that we do not use too small a sample and that complexity is not too high. For LS we use the same configuration as in the synthetic experiment. Results are shown in Table 2 and they mostly confirm our conclusions from the synthetic data. In particular, ML is slightly better than CV and subsampling yields a small improvement. In contrast with the artificial data, more iterations do provide a small improvement in performances and five iterations provide the best results in this experiment. Finally, we can also see that all of the GP variants outperform LS.

Although this is an improvement over existing algorithms accuracy of 80% is still not satisfactory. As discussed by Wachman (2009), one particularly challenging task is finding the true period of EB stars. The difficulty comes from the following two aspects. First, for a symmetric EB, the true period and half of the true period are not clearly distinguishable quantitatively. Second, methods that are better able to identify the true period of EBs are prone to find periods that are integer multiples of single bump stars like RR Lyrae stars and Cepheids. On the other hand, methods that fold RR Lyrae stars and Cepheids correctly often give "half" of the true period of EBs. In particular, the low performance of LS is due to the fact that it gives a half or otherwise wrong period for most EBs.

To illustrate the results, Figure 9 shows the periods found by LS and by GP on four stars. The top row shows two cases where the GP method finds the correct period and LS finds half the period. The bottom row shows cases where LS identifies the correct period and the GP does not. In the example on the left the GP doubles the period. In the example on the right the GP identifies a different period from LS but given the spread in the correct period the period it uncovers is not unreasonable.

4.2.2. Incorporating Domain Knowledge

We next show how this issue can be alleviated and the performance can be improved significantly using a learned probabilistic generative model. The methods developed are general and can be applied whenever such a model is available. As illustrated in Figure 8, our astrophysics knowledge suggests that different types of stars have different typical shift-invariant "shapes." In addition, each class has more than one such shape and each individual star has some variation from the common shape. We use the *Shift-invariant Grouped Mixed-effect Model* (GMT; Wang et al. 2010), which captures the common "shapes" via a mixture of GPs while at the same time allowing for individual variations. This model was previously developed

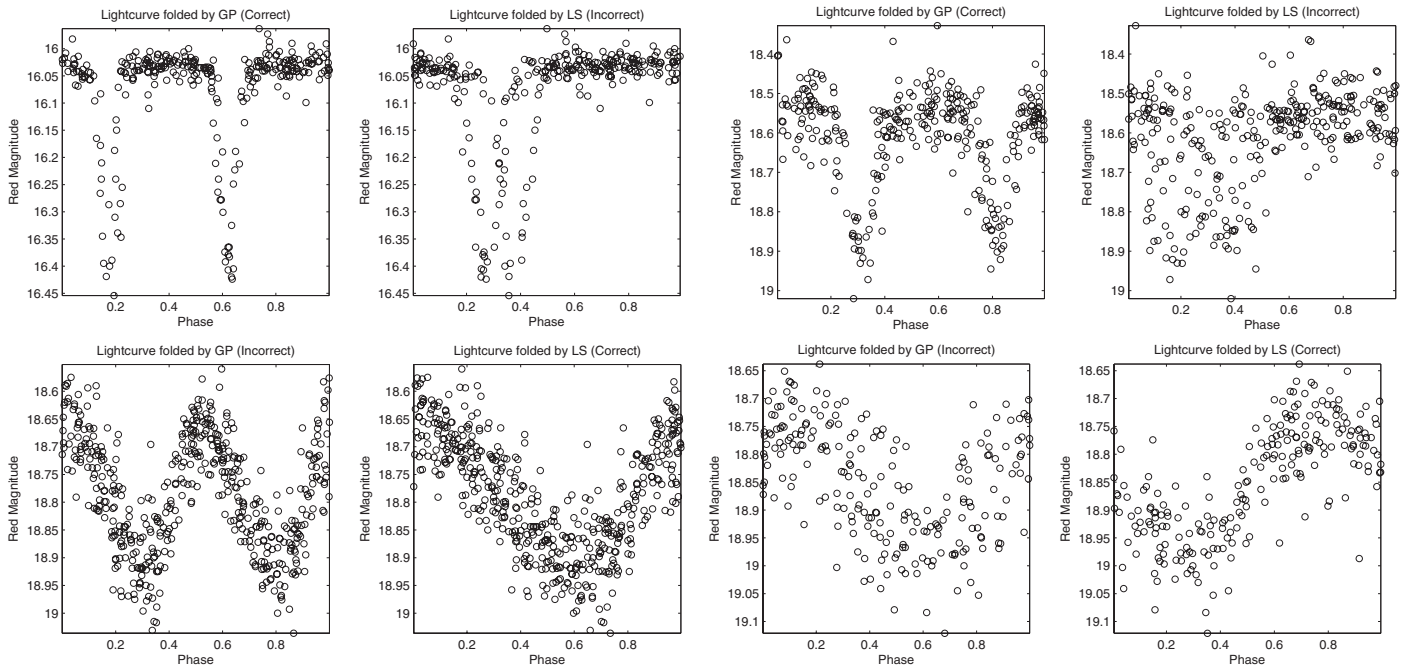

Figure 9. Examples of light curves where GP and LS identify different periods and one of them is correct. Each pair shows the time series folded by GP on the left and LS on the right. The top row shows cases where LS identifies half the period. The bottom row shows cases where GP identifies double the period or a different period.

Table 3
Comparison of Different Regularization Parameters on OGLEII Subset Using MAP

γ	0	0.1	0.3	0.5	0.7	0.9	1
ACC	0.87027	0.85946	0.81802	0.81802	0.80901	0.80721	0.8

to capture and aid in the classification of the astrophysics data. Once model parameters are learned we can calculate the likelihood of a light curve folded using a proposed period. Given the models, learned from a disjoint set of time series, for Cepheids, EBs, and RR Lyrae stars with parameter sets $\mathcal{M}_i, i = \{C, E, R\}$, there are two perspectives on how they can be used.

Model as prior. The models can be used to induce an improper prior distribution (or alternatively a penalty function) on the period p. Given period p and sample points $\boldsymbol{x}$ the prior is given by

$$\Pr(p) = \max_{i \in \{C,E,R\}} \left(\Pr(\boldsymbol{y}|\boldsymbol{x}, p; \mathcal{M}_i)\right), \tag{20}$$

where from the perspective of $\mathcal{M}_i$, $\boldsymbol{x}$ and corresponding points in $\boldsymbol{y}$ are interpreted as if they were sampled modulo p. Thus, combining this prior with the marginal likelihood, a Maximum A Posteriori (MAP) estimation can be obtained. Adding a regularization parameter γ to obtain a tradeoff between the marginal likelihood and the improper prior we get our criterion:

$$\begin{aligned}\log \Pr(p|\boldsymbol{x}, \boldsymbol{y}; \mathcal{M}) = \gamma \log \Pr(\boldsymbol{y}|\boldsymbol{x}, p; \mathcal{M}) \\ + (1-\gamma) \log \Pr(p),\end{aligned} \tag{21}$$

where $\Pr(\boldsymbol{y}|\boldsymbol{x}, p; \mathcal{M})$ is exactly as Equation (12) where the period portion of $\mathcal{M}$ is fixed to be p. When using this approach with our algorithm we use Equation (21) instead of Equation (12) as the score function in lines 5 and 13 of the algorithm. The results for different values of γ (with subsampling and five iterations) are shown in Table 3. The results show that GMT on its own ($\gamma = 0$) is a good criterion for period finding. This is as one might expect because the OGLEII data set includes only stars of the three types captured by GMT.

In this experiment, regularized versions do not improve the result of the GMT model. However, we believe that this will be the method of choice in other cases when the prior information is less strong. In particular, if the data include unknown shapes that are not covered by the generative model then the prior on its own will fail. On the other hand when using Equation (21) with enough data the prior will be dominated by the likelihood term and therefore the correct period can be detected. In contrast, the filter method discussed next does not have such functionality.

Model as filter. Our second approach uses the model as a post-processing filter and it is applicable to any method that scores different periods before picking the top scoring one as its estimate. For example, suppose we are given the top K best periods $\{p_i\}, i = 1, \ldots, K$ found by LS, then we choose the one such that

$$p^* = \operatorname*{argmax}_{i \in \{1,\ldots,K\}} \left(\max_{j \in \{C,E,R\}} \left[\log \Pr(\boldsymbol{y}|\boldsymbol{x}, p_i; \mathcal{M}_j)\right] \right). \tag{22}$$

Thus, when using the GMT as a filter, step 17 in our algorithm is changed to record the top K frequencies from the last for loop, evaluate each one using the GMT model likelihood, and output the top scoring frequency.

Heuristic for variable periodic stars. The two approaches above are general and can be used in any problem where a model is available. For the astrophysics problem we develop another heuristic that specifically addresses the half-period problem of EBs. In particular, when using the filter method, instead of choosing the top K periods, we double the selected periods, evaluate both the original and doubled periods $\{p_i, 2p_i\}$ using the GMT model, and choose the best one.

Results of experiments using the filter method with and without the domain-specific heuristic are given in Table 4, based on the five iteration versions of subsampling GP. The

Table 4
Comparisons of the Accuracy of Different Algorithms on OGLEII Subset Using the GMT as a Filter

	Original	Single Filter	Filter
LS	0.7333	0.7243	**0.9053**
GP	0.8000	0.8829	**0.9081**
LS+GP	···	0.8811	**0.9297**

Notes. Single denotes without the double period heuristic.

filter method significantly improves the performance of our algorithm showing its general applicability. The domain-specific heuristic provides an additional improvement. For LS, the general filter method does not help but the domain-specific heuristic significantly improves its performance. By analyzing the errors of both GP and LS, we found that their error regions are different. Therefore, we further propose a method that combines the two methods in the following way: pick the top K periods found by both methods and evaluate the original and doubled periods using the GMT to select the best one. As Table 4 shows, the combination gives an additional 2% improvement on the OGLEII subset.

4.2.3. Application

Finally, we apply our method using marginal likelihood with two-level grid search, sub-sampling, two iterations, and filtering on the complete OGLEII data set minus the development OGLEII subset. Note that the parameters of the algorithm, other than domain-dependent heuristics, are chosen based on our results from the artificial data. The accuracy is reported using 10-fold cross validation under the following setting: the GMT is trained using the training set and we seek to find the periods for the stars in the test set. We compare our results to the best result from Wachman (2009) that used an improvement of LS, despite the fact that they filtered out 1832 difficult stars due to insufficient sampling points and noise. The results are shown in Table 5. We can see that our approach significantly outperforms existing methods on OGLEII.

5. RELATED WORK

Period detection has been extensively studied in the literature and especially in astrophysics. The periodogram, as a tool for spectral analysis, dates back to the 19th century when Schuster applied it to the analysis of some data sets. The behavior of the periodogram in estimating frequency was discussed by Deeming (1975). The periodogram is defined as the modulus squared of its discrete Fourier transform (Deeming 1975). Lomb (1976) and Scargle (1982) introduced the so-called Lomb–Scargle (LS) periodogram that was discussed above and which rates periods based on the sum-of-squares error of a sine wave at the given period. This method has been used in astrophysics (Cumming 2004; Wachman 2009) and has also been used in bioinformatics (Glynn et al. 2006; Wentao et al. 2008). One can show that the LS periodogram is identical to the equation we would derive if we attempted to estimate the harmonic content of a data set at a specific frequency using the linear least-squares model (Scargle 1982). This technique was originally named least-squares spectral analysis method (Vaníček 1969). Many extensions of the LS periodogram exist in the literature (Bretthorst 2001). Hall & Li (2006) proposed the periodogram for non-parametric regression models and discussed its statistical properties. This was later applied to the situation where the regression model is the superposition of functions with different periods (Hall 2008).

The other main approach uses least-squares estimates, equivalent to maximum likelihood methods under Gaussian noise assumption, using different choices of periodic regression models. This approach, using finite-parameter trigonometric series of different orders, has been explored by various authors (Hartley 1949; Quinn & Thomson 1991; Quinn & Fernandes 1991; Quinn 1999; Quinn & Hannan 2001). Note that if the order of the trigonometric series is high then this is very close to nonparametric methods (Hall 2008).

Another intuition is to minimize some measure of dispersion of the data in phase space. PDM (Stellingwerf 1978), described above, performs a least-squares fit to the mean curve defined by averaging points in bins. Lafler & Kinman (1965) described a procedure which involves trial-period folding followed by a minimization of the differences between observations of adjacent phases.

Other least-squares methods use smoothing based on splines, robust splines, or variable-span smoothers. Craven & Wahba (1978) discussed the problem of smoothing periodic curves with spline functions in the regularization framework and invented the generalized cross-validation score to estimate the period of a variable star. Oh et al. (2004) extended it by substituting the smoothing splines with robust splines to alleviate the effects caused by outliers. Supersmoother, a variable-span smoother based on running linear smooths, is used for frequency estimation in McDonald (1986).

Several other approaches exist in the literature. Perhaps the most related work is Hall et al. (2000) who studied non-parametric models for frequency estimation, including the Nadaraya–Watson estimator, and discussed their statistical properties. This was extended to perform inference for multi-period functions (Hall & Yin 2003) and evolving periodic functions (Genton & Hall 2007; Hall 2008). Our work differs from Hall et al. (2000) in three aspects: (1) the GP framework presented in this paper is more general in that one can plug in different periodic covariance functions for different prior assumptions, (2) we use marginal likelihood that can be interpreted to indicate how the data agree with our prior belief, and (3) we introduce mechanisms to overcome the computational complexity of period selection.

Other approaches include entropy minimization (Huijse et al. 2011), data compensated discrete Fourier transform (Ferraz-Mello 1981), and Bayesian models (Gregory & Loredo 1996; Scargle 1998). Recently, Bayesian methods have also been applied to solve the frequency estimation problem, for example Bayesian binning for Poisson-regime (Gregory & Loredo 1996) and Bayesian blocks (Scargle 1998). Ford et al. (2011) proposed a Bayesian extension of multi-period LS that is capable of estimating periodic functions having an additional polynomial

Table 5
Comparisons of Accuracies for Full Set of OGLEII

	Method in Wachman (2009)	LS-filter	GP-filter	GP-LS-filter
ACC	0.8680	0.8975 ± 0.04	0.8963 ± 0.03	$\mathbf{0.9243 \pm 0.03}$

trend. The main difference from our work is the kernel-based formulation in our approach.

6. CONCLUSION

The paper introduces a nonparametric Bayesian approach for period estimation based on GP regression. We develop a model selection algorithm for GP regression that combines gradient-based search and grid search, and incorporates several algorithmic improvements and approximations leading to a considerable decrease in run time. The algorithm performs significantly better than existing state of the art algorithms when the data are not sinusoidal. Further, we show how domain knowledge can be incorporated into our model as a prior or post-processing filter, and apply this idea in the astrophysics domain. Our algorithm delivers significantly higher accuracy than existing state of the art in estimating the periods of variable periodic stars.

An important direction for future work is to extend our model to develop a corresponding statistical test for periodicity, that is, to determine whether a time series is periodic. This will streamline the application of our algorithm to new astrophysics catalogs such as MACHO (Alcock et al. 1993) where both periodicity testing and period estimation are needed. Another important direction is establishing the theoretical properties of our method. Hall et al. (2000) provided the first-order properties of nonparametric estimators such that under mild regularity conditions, the estimator is consistent and asymptotically normally distributed. Our method differs in two ways: we use a GP regressor instead of Nadaraya–Watson estimator and we choose the period that minimizes marginal likelihood rather than using a cross-validation estimate. Based on the well-known connection between kernel regression and GP regression, we conjecture that similar results exist for the proposed method.

This research was partly supported by NSF grant IIS-0803409. The experiments in this paper were performed on the Odyssey cluster supported by the FAS Research Computing Group at Harvard and the Tufts Linux Research Cluster supported by Tufts UIT Research Computing.

APPENDIX

LOW-RANK APPROXIMATION

In this appendix, we complete the details on how the first-order approximation with low-rank approximation can be achieved by a series of rank one updates/downdates of the Cholesky factors. As shown by Seeger (2007) each such update can be done in $\mathcal{O}(N^2)$ using a series of Givens rotations.

It can be easily seen that $\widetilde{\boldsymbol{K}}$ is a real symmetric matrix. Denote its eigendecomposition as $\widetilde{\boldsymbol{K}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}$, then it can be written as the sum of a series of rank one components,

$$\widetilde{\boldsymbol{K}} = \sum_{i=1}^{N} \mathrm{sgn}(\lambda_i)(\sqrt{|\lambda_i|}\mathbf{u}_i)(\sqrt{|\lambda_i|}\mathbf{u}_i)^{\mathrm{T}}, \quad \text{(A1)}$$

where λ_i is the ith eigenvalue and $\mathbf{u}_i$ is the corresponding eigenvector. Furthermore, we perform a low-rank approximation to $\widetilde{\boldsymbol{K}}$ such that

$$\widetilde{\boldsymbol{K}} \approx \sum_{i=1}^{M} \mathrm{sgn}(\lambda_{(i)})(\sqrt{|\lambda_{(i)}|}\mathbf{u}_{(i)})(\sqrt{|\lambda_{(i)}|}\mathbf{u}_{(i)})^{\mathrm{T}}, \quad \text{(A2)}$$

where $M < N$ is a predefined rank and $\lambda_{(i)}$ and $\mathbf{u}_{(i)}$ are the ith largest (in absolute value) eigenvalue and its corresponding eigenvector. Therefore we have,

$$\boldsymbol{K}_{w_1} \approx \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}} + \sum_{i=1}^{M} \mathrm{sgn}(\lambda_{(i)})((\Delta w)^{1/2}\boldsymbol{\ell}_i)((\Delta w)^{1/2}\boldsymbol{\ell}_i)^{\mathrm{T}}, \quad \text{(A3)}$$

where $\boldsymbol{\ell}_i = \sqrt{|\lambda_{(i)}|}\mathbf{u}_{(i)}$. We can see that the complexity for calculating the Cholesky factor of $\boldsymbol{K}_{w_1}$ becomes $\mathcal{O}(MN^2)$. Therefore, we can choose an ϵ-net $\mathcal{E}$ of the fine grid such that $\forall w \in \mathcal{F}, \sup_{v\in\mathcal{E}} |w - v| < \epsilon$, perform the exact Cholesky decomposition directly only on the ϵ-net, and use the approximation on the other frequencies. In this way, we reduce the complexity from $\mathcal{O}(|\mathcal{F}|N^3)$ to $\mathcal{O}(|\mathcal{E}|N^3 + |\mathcal{F}|MN^2)$.

REFERENCES

Alcock, C., Allsman, R. A., Axelrod, T. S., et al. 1993, in ASP Conf. Ser. 43, Sky Surveys. Protostars to Protogalaxies, ed. B. T. Soifer (San Francisco, CA: ASP), 291
Bishop, C. 2006, Pattern Recognition and Machine Learning, Vol. 4 (New York: Springer)
Bretthorst, G. L. 2001, in AIP Conf. Proc. 568, Bayesian Inference and Maximum Entropy Methods in Science and Engineering, ed. A. Mohammad-Djafari (Melville, NY: AIP), 241
Craven, P., & Wahba, G. 1978, Numer. Math., 31, 377
Cumming, A. 2004, MNRAS, 354, 1165
Deeming, T. 1975, Ap&SS, 36, 137
Ferraz-Mello, S. 1981, AJ, 86, 619
Ford, E., Moorhead, A., & Veras, D. 2011, Bayesian Anal., 6, 475
Genton, M., & Hall, P. 2007, J. R. Stat. Soc. B (Stat. Methodol.), 69, 643
Glynn, E., Chen, J., & Mushegian, A. 2006, Bioinformatics, 22, 310
Gregory, P., & Loredo, T. 1996, ApJ, 473, 1059
Hall, P. 2008, COMPSTAT, 2008, 3
Hall, P., & Li, M. 2006, Biometrika, 93, 411
Hall, P., Reimann, J., & Rice, J. 2000, Biometrika, 87, 545
Hall, P., & Yin, J. 2003, J. R. Stat. Soc. B (Stat. Methodol.), 65, 869
Hartley, H. 1949, Biometrika, 36, 194
Huijse, P., Estevez, P. A., Zegers, P., Principe, J. C., & Protopapas, P. 2011, IEEE Signal Process. Lett., 18, 371
Lafler, J., & Kinman, T. 1965, ApJS, 11, 216
Lomb, N. 1976, Ap&SS, 39, 447
McDonald, J. 1986, SIAM J. Sci. Stat. Comput., 7, 665
Oh, H.-S., Nychka, D., Brown, T., & Charbonneau, P. 2004, J. R. Stat. Soc., 53, 15
Petit, M. 1987, Variable Stars (Chichester, UK: Wiley)
Protopapas, P., Jimenez, R., & Alcock, C. 2005, MNRAS, 362, 460
Quinn, B. 1999, Biometrika, 86, 213
Quinn, B., & Fernandes, J. 1991, Biometrika, 78, 489
Quinn, B., & Hannan, E. 2001, The Estimation and Tracking of Frequency (Cambridge: Cambridge Univ. Press)
Quinn, B., & Thomson, P. 1991, Biometrika, 78, 65
Rasmussen, C., & Nickisch, H. 2010, J. Mach. Learn. Res., 11, 3011
Rasmussen, C., & Williams, C. 2005, Gaussian Process. Mach. Learn. (Cambridge, MA: MIT Press)
Reimann, J. 1994, PhD thesis, UC Berkeley
Scargle, J. 1982, ApJ, 263, 835
Scargle, J. 1998, ApJ, 504, 405
Seeger, M. 2007, Technical Report, Univ. California at Berkeley
Soszynski, I., Udalski, A., & Szymanski, M. 2003, Acta Astron., 53, 93
Stellingwerf, R. 1978, ApJ, 224, 953
Vaníček, P. 1969, Ap&SS, 4, 387
Wachman, G. 2009, PhD thesis, Tufts Univ.
Wang, Y., Khardon, R., & Protopapas, P. 2010, Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science 6323 (Berlin: Springer), 418
Wentao, Z., Kwadwo, A., Erchin, S., et al. 2008, EURASIP J. Bioinform. Syst. Biol., 2008: 769293