Period Analysis of Variable Stars: Temporal Dependence and Local Optima

Peter Hooper
Department of Mathematical and Statistical Sciences
University of Alberta

hooper@stat.ualberta.ca
www.stat.ualberta.ca/~hooper

November, 2006
Outline

• Introduction

• Methodology:
  – Adaptive logistic basis (ALB) regression models
  – Finding and comparing local optima
  – Comparing harmonics
  – Bootstrap confidence intervals for light curves

• Examples

• Conclusion
• Variable stars (or binary systems, etc.): brightness varies over time.
  – Stellar magnitude = \(-2.5 \log(\text{flux density})\)
  – Informally define “brightness” as negative R-magnitude
  – Periodic variable stars: change is periodic.
  – Examples: eclipsing binary and RR Lyrae.

• Data and model.
  – Observe brightness \(y_i\) at time \(t_i\) (in days) for \(i = 1 \ldots n\).
  – Times are irregularly spaced.
  – Given a period \(\theta\), define phase values \(u_i = t_i/\theta \mod 1\).
  – Model: \(y_i = g(t_i/\theta) + e_i = g(u_i) + e_i\)
  – Light curve \(g\) is a smooth periodic function with period 1.

• Problem: estimate period \(\theta\) and light curve \(g\).
• Previous approaches.

• Simple cosine models (Deeming, 1975; Lomb, 1976; Scargle, 1982).

• Minimize $\sum (y_{i+1}^* - y_i^*)^2$ where $(u_i^*, y_i^*)$ are ordered phase values and corresponding brightness (Lafler and Kinman, 1965).

• Minimize sum of distances between points $(y_i^*, u_i^*)$ and $(y_{i+1}^*, u_{i+1}^*)$ (Dwortesky 1983).

• ANOVA method that implicitly approximates $g$ by a piecewise constant function (Stellingwerf 1978).

• Fit $g$ using SuperSmooother and select $\theta$ by minimizing an absolute error criterion. (Riemann 1994).
• Use penalized smoothing splines for $g$ with a robust cross-validation (RCV) criterion to estimate both $\theta$ and a smoothing parameter which depends on $\theta$. The RCV criterion is a computationally efficient approximation of a leave-one-out cross-validated estimate of expected Huber loss. (Oh, Nychka, Brown, and Charbonneau, 2004).

• Oh et. al. used data derived from the project ‘Stellar astrophysics and research on exoplanets’ (Charbonneau, Brown, Latham, and Mayor, 2000). These data (also used here) consist of
  – coincident measurements of R-magnitude for seven stars,
  – $n = 351$ observations for each star,
  – the observations were made on 13 nights over a 44-night interval, and
  – the number of observations per night varies from 7 to 42.

• Previous methods assume independent errors, but plots of residuals against time show clear patterns (missed in earlier analysis).
Figure 1: Plots illustrating patterns among the residuals from a light curve estimate for star 0306 with period 0.8763. Two outliers are omitted from both plots. Two sets of points are highlighted: ○ for nights 776 and 777, and △ for nights 804, 805, and 807.
• The patterns suggest a need for dependent errors; e.g.,

\[ y_i = g(t_i/\theta) + h(t_i) + e_i \]

where \( h \) is a random function of time, perhaps reflecting changes in observing conditions, and the \( e_i \) are independent.

• Patterns related to \( h \) can

  – distort the estimate \( \hat{g} \),
  – complicate choice of goodness-of-fit criterion used to select \( \theta \),
  – increase the difficulty in quantifying uncertainty.

• Distortion in \( \hat{g} \) can be large when large gaps are present in the ordered phase values.
Figure 2: Line plots showing the three largest gap sizes as a function of period: largest gap, sum of two largest, and sum of three largest. Gap sizes are defined as \((u_i^* - u_{i-1}^*)\) for \(i = 2, \ldots, n\), and \((u_1^* - u_n^* + 1)\), where \(u_1^* \leq \ldots \leq u_n^*\) are the ordered phase values determined by period \(\theta\) and times \(t_1, \ldots, t_n\).
• Proposed strategy:

• Select a goodness-of-fit criterion; e.g., squared error, absolute error, Huber loss. Focus on absolute error here.

• Initially, ignoring $h$, obtain rough estimates $\hat{g}$ for a large number of $\theta$ values. Use these to obtain a rough estimate of a risk function $R(\theta)$. Identify a much smaller number of candidate $\theta$ values.

• For each candidate $\theta$, apply a backfitting algorithm to estimate $g$ and $h$.

• Use an ALB model for $g$ and a BIC-type penalty to control overfitting.

• Use a piecewise constant model for $h$; i.e., a night-effect model.
Adaptive Logistic Basis (ALB) regression (Hooper, 2001, CJS).

Represent periodic functions $g$ by functions $f$ defined on $\mathbb{R}^2$ by mapping time points $t_i$ to points $x_i$ on the unit circle:

$$g(t/\theta) = f(x) = f(\cos(2\pi t/\theta), \sin(2\pi t/\theta)).$$

ALB models approximate $f$ by a linear combination of logistic basis functions, shown here with a “reference point” parameterization:

$$f(x) \approx f_K(x) = \sum_{k=1}^{K} \delta_k \phi_k(x),$$

$$\phi_k(x) = \frac{\exp\left(\gamma_k - \tau^{-2}\|x - \xi_k\|^2\right)}{\sum_{m=1}^{K} \exp\left(\gamma_m - \tau^{-2}\|x - \xi_m\|^2\right)}.$$
No constant term required since $\sum \phi_k(x) = 1$ for all $x$.

$f_K$ is determined by scalar parameters $\tau, \gamma_k, \delta_k$ and vectors $\xi_k \in \mathbb{R}^2$. The model is over-parameterized, with $\tau$ and one pair $(\gamma_k, \xi_k)$ redundant, so the effective number of parameters determining $f_K$ is $4K - 3$.

The functions $\phi_k$ and $f_K$ are defined on $\mathbb{R}^2$, but only their values on the unit circle are relevant. These values can be plotted as functions of phase; e.g., the $K = 4$ basis functions used for $\hat{g}$ in Figure 1(a).
Estimate $g$ by minimizing a risk estimate over $K$ and other parameters.

Given $\theta$ and $\hat{h}$, define

$$R(\theta) = \left(\frac{\tilde{n}}{\tilde{n} - 4K + 3}\right)^q \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{h}(t_i) - \hat{f}_K(x_i)|^q$$

Setting $q = 2$ and $q = 1$ gives squared error and absolute error.

The multiplicative penalty is motivated by GCV (Craven & Wahba, 1979).

If $\tilde{n} = n$ then the criterion is similar to AIC when $4K/n$ is small.

If $\tilde{n} = 2n/\log(n)$ then the criterion is similar to BIC.

AIC tends to overfit so I prefer BIC here.
• Finding and comparing local optima.

• Initial scan.
  – Fit ALB models with $\hat{h} = 0$, fixed $K = 12$, using a streamlined stochastic gradient algorithm.
  – Search over grid, $\log(\theta)$ equally spaced for $0.01 \leq \theta \leq 10$, about 56000 $\theta$ values.
  – Identify best 500 candidates, cluster, find local minima.

• Final selection.
  – Estimate $K$ and ALB parameters using backfitting and penalized risk.
  – Estimate $h(t_i)$ as median of $y_j - \hat{g}(t_j/\theta)$ for $t_j$ on same night.
  – For each $\theta$, record four risk values: $R_1$ after initial $\hat{g}$ with $\hat{h} = 0$. $R_2$ after first $\hat{h}$. $R_3$ after second $\hat{g}$. $R_4$ after second $\hat{h}$.
  – Use $R_3$ to locate and compare local minima.
Figure 3: Smoothed estimates of absolute-error risk for star 4699, based on ALB fits with $K = 12$ and a highly streamlined training algorithm. Risk is plotted against $\log(\theta)$ for $0.2 < \theta < 10$. The four best local minima are highlighted.
Compare harmonics using approximate LRT.

Test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$.

If $\theta_1 = 2\theta_0$ then the null hypothesis is nested in alternative.

Assume fixed night effects $h$, appropriate distribution for $e_i$, fixed $K_0$ and $K_1$.

$$LRT \approx 2n \log \left\{ \frac{\tilde{n} - 4K_0 + 3}{\tilde{n} - 4K_1 + 3} \left( \frac{R_3(\theta_0)}{R_3(\theta_1)} \right)^{1/q} \right\}$$

Approximate null distribution by $\chi^2_\nu$ with $\nu = 4(K_1 - K_0)$.

The BIC-type penalty $R_3$ typically selects $\theta_0$ unless the test indicates strong evidence against $\theta_0$. 

15
Confidence intervals for light curves given fixed period.

Approximate pointwise $100(1 - \alpha)\%$ confidence intervals based on the golden-section (or wild) bootstrap (Härdle and Marron, 1991).

Given $\hat{\theta}$ and $\hat{g}$, put $r_i = y_i - \hat{g}(t_i/\hat{\theta})$ and $y_i^* = \hat{g}(t_i/\hat{\theta}) + w_i r_i$.

Weights $w_i$ are equal for $t_i$ on same night, otherwise independent:

$$w_i = \begin{cases} (1 - \sqrt{5})/2 & \text{with probability } p = (5 + \sqrt{5})/10 \\ (1 + \sqrt{5})/2 & \text{with probability } 1 - p. \end{cases}$$

For phase values $u \in [0, 1]$, let $\hat{g}_1(u)$ and $\hat{g}_2(u)$ be the $\alpha/2$ and $1 - \alpha/2$ sample quantiles of bootstrap estimates $\{g^*_m(u), m = 1, \ldots, 1000\}$.

Proposed interval: $\hat{g}(u) \pm \{\hat{g}_2(u) - \hat{g}_1(u)\}/2$. 
• Oh et al. (2004) used data for seven stars to compare their RCV method with seven alternative methods.

• Their results for all eight methods were similar for six of the seven stars, yielding period estimates consistent with one of the ALB local minima.

• For the exception (star 4699), the estimates were split between two neighbouring local minima.

• The consistency in estimates might be due to a restricted range of search.

• \( \hat{\theta}_{\text{ALB}} \) denoted by †

• \( \hat{\theta}_{\text{RCV}} \) is close to period denoted by *
### Table

<table>
<thead>
<tr>
<th>star</th>
<th>period</th>
<th>$R_1$</th>
<th>$R_2$</th>
<th>$R_3$</th>
<th>$R_4$</th>
<th>$K$</th>
<th>#lm</th>
<th>symm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1164</td>
<td>1.4633*</td>
<td>1.956</td>
<td>1.691</td>
<td>1.661</td>
<td>1.648</td>
<td>5</td>
<td>2</td>
<td>0.900</td>
</tr>
<tr>
<td></td>
<td>2.9433†</td>
<td>1.625</td>
<td>1.271</td>
<td>1.239</td>
<td>1.234</td>
<td>11</td>
<td>3</td>
<td>0.888</td>
</tr>
</tbody>
</table>

### Figures

(c) Star 1164

Phase for period 1.4633

(d) Star 1164

Phase for period 2.9433
\begin{tabular}{cccccccc}
\hline
star & period & $R_1$ & $R_2$ & $R_3$ & $R_4$ & $K$ & #lm & symm \\
4699 & 3.2257$^{*\dagger}$ & 1.230 & 0.952 & 0.950 & 0.948 & 4 & 2 & 0.972 \\
3.2905 & 1.201 & 0.968 & 0.967 & 0.964 & 4 & 2 & 0.850 \\
\hline
\end{tabular}

$\hat{\theta}_{RCV} \approx \hat{\theta}_{ALB} = 3.2257$ but six other methods have $\hat{\theta} \approx 3.2905$. 
<table>
<thead>
<tr>
<th>star</th>
<th>period</th>
<th>$R_1$</th>
<th>$R_2$</th>
<th>$R_3$</th>
<th>$R_4$</th>
<th>$K$</th>
<th>#lm</th>
<th>symm</th>
</tr>
</thead>
<tbody>
<tr>
<td>4865</td>
<td>1.1629†</td>
<td>1.344</td>
<td>0.999</td>
<td>0.970</td>
<td>0.968</td>
<td>3</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>2.3256*</td>
<td>1.275</td>
<td>1.158</td>
<td>1.142</td>
<td>1.120</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0.952</td>
</tr>
</tbody>
</table>
Approximate LRT has $P$-value $< 0.00001$. 
Phase for period 0.4903
Brightness

Phase for period 0.9806
Brightness

Phase for period 0.9806
Time

<table>
<thead>
<tr>
<th>star</th>
<th>period</th>
<th>$R_1$</th>
<th>$R_2$</th>
<th>$R_3$</th>
<th>$R_4$</th>
<th>$K$</th>
<th>#lm</th>
<th>symm</th>
</tr>
</thead>
<tbody>
<tr>
<td>5954</td>
<td>0.4903</td>
<td>0.768</td>
<td>0.629</td>
<td>0.622</td>
<td>0.621</td>
<td>2</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>0.9806</td>
<td>0.800</td>
<td>0.664</td>
<td>0.655</td>
<td>0.653</td>
<td>4</td>
<td>2</td>
<td>0.996</td>
</tr>
</tbody>
</table>

Approximate LRT has $P$-value = 0.074.
• Concluding remarks:

• Fit models for multiple periodicity by backfitting:

\[ y_i = \sum_{m=1}^{M} g_m(t_i/\theta_m) + h(t_i) + e_i \]

• Theory/simulations supporting LRT and bootstrap confidence intervals.

• Evaluate uncertainty about \( \theta \), local optima.

• Use expert knowledge about known categories of variable stars; e.g., relationships among period, brightness, number of local minima, and other characteristics of shape. Combine estimation with classification.