

Fourier decomposition of variable stars using regularized regression

Daniel Wysocki

Rochester Institute of Technology

AST Graduate Seminar – December 7th, 2015



Variable Stars

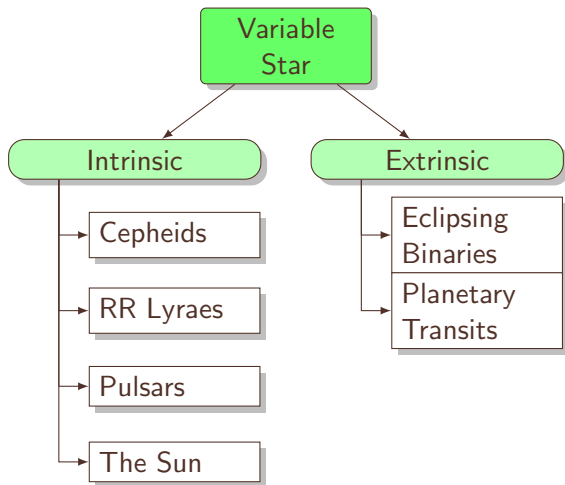


Overview

- in general, any star whose brightness changes on short timescales is a variable star
- many different types exist



Some classes of variable stars



Pulsating periodic intrinsic variables

For the remainder of this talk:

variable star \equiv pulsating periodic intrinsic variable star.

- not in hydrostatic equilibrium
 - typically in the instability strip
- periodic oscillation
 - predictable
- stellar pulsation
 - κ -mechanism



Henrietta Swan Leavitt



Public domain in USA

- worked as a “computer” at Harvard in the early 20th century
- discovered a relation between the period and luminosity of Cepheids
 - Leavitt’s law
 - standard candles
- enabled Edwin Hubble to measure the expansion of the Universe

Henrietta Swan Leavitt



Light Curves



Overview

- repeated photometric measurements of an object over time
- plotting brightness versus time gives us a light curve



Light Curve of a Cepheid variable star

Visualization of OGLE-LMC-CEP-0002

© Daniel Wysocki



Fourier decomposition



Public domain in USA

Joseph Fourier

- any continuous, periodic function can be represented as an infinite Fourier series

$$f(t) = A_0 + \sum_{k=1}^{\infty} A_k \cos(k\omega t + \Phi_k)$$

- characterized by the angular frequency ω , the mean A_0 , the amplitudes A_k , and the phase shifts Φ_k

Fourier (1808)



Fourier decomposition of periodic light curves

$$m(t) = A_0 + \sum_{k=1}^n A_k \cos(k\omega t + \Phi_k)$$

- Cepheid-like light curves well described by n th order Fourier Series
- physically they are close to harmonic oscillators



Solving for series parameters

$$\begin{aligned}m(t) &= A_0 + \sum_{k=1}^n A_k \cos(k\omega t + \Phi_k) \\&= A_0 + \sum_{k=1}^n [a_k \sin(k\omega t) + b_k \cos(k\omega t)]\end{aligned}$$

- Fourier series are non-linear
- simultaneously finding the optimal n , ω , A_k , and Φ_k is not easy



Period finding

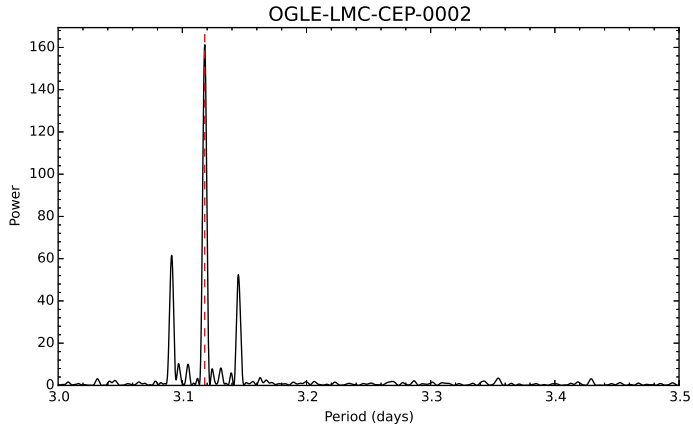
- the most important parameter is the period

$$\omega = 2\pi/P$$

- we can approximate this by itself using a periodogram
 - Lomb-Scargle



Lomb-Scargle periodogram



Periodogram of star with 3.11804 day period

It's linear!

$$m(t) = A_0 + \sum_{k=1}^n [a_k \sin(k\omega t) + b_k \cos(k\omega t)]$$

can be written in the form

$$\mathbf{X}\vec{\beta} = \vec{y}$$

which can be approximated using ordinary linear regression



System of equations

$$\vec{y} \rightarrow (m_1 \quad m_2 \quad \dots \quad m_N)$$

$$\vec{\beta} \rightarrow (A_0 \quad a_1 \quad b_1 \quad \dots \quad a_n \quad b_n)$$

$$\mathbf{X} \rightarrow \begin{pmatrix} 1 & \sin(1\omega t_1) & \cos(1\omega t_1) & \dots & \sin(n\omega t_1) & \cos(n\omega t_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \sin(1\omega t_N) & \cos(1\omega t_N) & \dots & \sin(n\omega t_N) & \cos(n\omega t_N) \end{pmatrix}$$

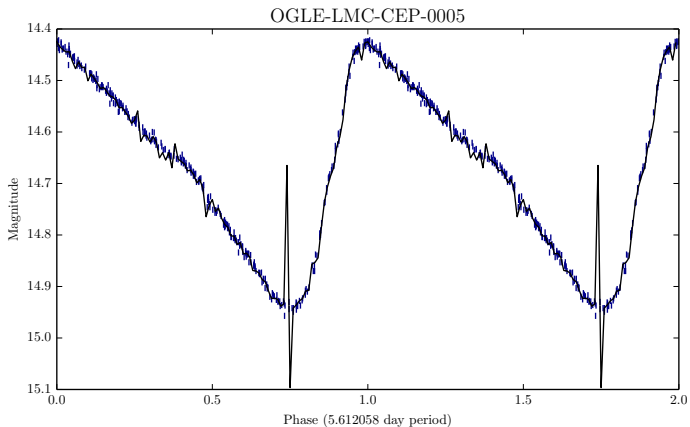


How many terms?

- wait, we never decided on the order of the fit, n
- it's just a truncated series expansion
 - more terms means better, right?
 - let's try 100 terms...



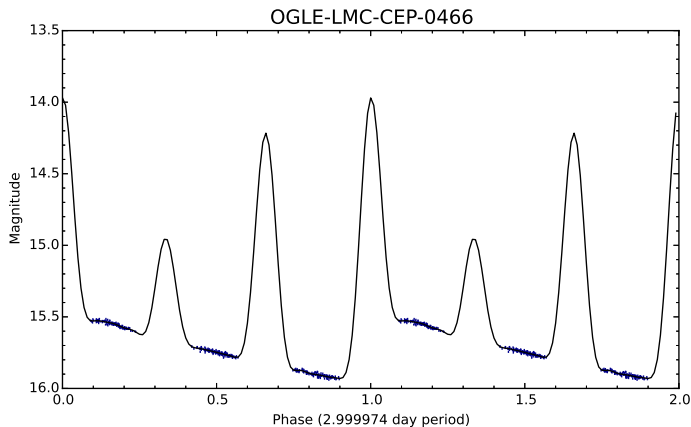
Overfitting



100th order fit

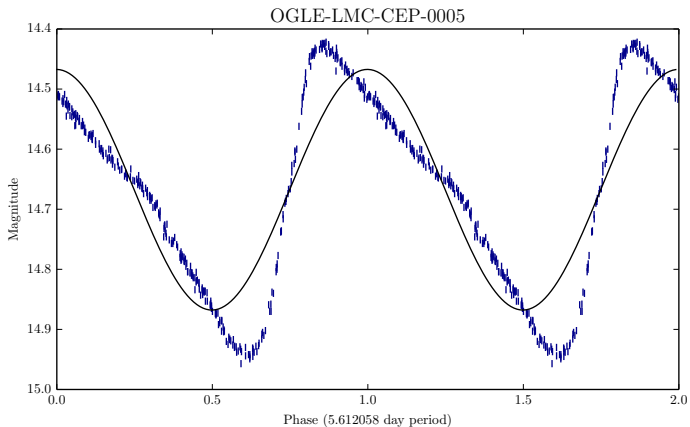


Overfitting (again)



12th order fit

Underfitting



1st order fit

Choosing n

- need some criteria to decide the order of the fit
- Baart's criteria is often used for this
 - iterative approach, increasing n until diminishing returns
 - good at avoiding underfitting
 - bad at avoiding overfitting



Taking a step back

- take photometric measurements
- find the period
- approximate coefficients with OLS
- find the best order of fit using Baart's criteria

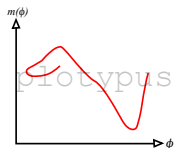


Taking a step back

- take photometric measurements
- **periodogram**
- **regression**
- **model selection**



Plotypus



- tool for modeling and plotting light curves
- free and open source
- version controlled and documented
- generated the light curve plots in this presentation
- astrowego.github.io/plotypus/
- download today!

Bellinger, Wysocki, and Kanbur (2015b)

Unconstrained regression

$$\mathbf{X}\vec{\beta} = \vec{y}$$

$$\begin{aligned} (A_0, a_k, b_k) &= \operatorname{argmin}_{\beta} \left\| \mathbf{X}\vec{\beta} - \vec{y} \right\|_2^2 \\ &= \operatorname{argmin}_{(A_0, a_k, b_k)} \sum_{i=1}^N \left(A_0 + \sum_{k=1}^n \begin{bmatrix} a_k \sin(k\omega t_i) \\ +b_k \cos(k\omega t_i) \end{bmatrix} - m_i \right)^2 \end{aligned}$$

Find coefficients which minimize magnitude of residual vector



ℓ_0 regularization

$$(A_0, a_k, b_k) = \underset{\beta}{\operatorname{argmin}} \left\{ \left\| \mathbf{X}\vec{\beta} - \vec{y} \right\|_2^2 + \lambda \left\| \vec{\beta} \right\|_0 \right\}$$

- $\left\| \vec{\beta} \right\|_0$ is equal to the number of non-zero terms in $\vec{\beta}$
- adds a penalty on the number of parameters, weighted by λ
- this is computationally expensive



ℓ_1 regularization (LASSO)

$$\begin{aligned}
 (A_0, a_k, b_k) &= \underset{\beta}{\operatorname{argmin}} \left\{ \left\| \mathbf{X}\vec{\beta} - \vec{y} \right\|_2^2 + \left\| \vec{\beta} \right\|_1 \right\} \\
 &= \underset{(A_0, a_k, b_k)}{\operatorname{argmin}} \left\{ \sum_{i=1}^N \left(A_0 + \sum_{k=1}^n \left[\begin{array}{l} a_k \sin(k\omega t_i) \\ + b_k \cos(k\omega t_i) \end{array} \right] - m_i \right)^2 \right. \\
 &\quad \left. + \lambda \sum_{k=0}^n (|a_k| + |b_k|) \right\}
 \end{aligned}$$

- least absolute shrinkage and selection operator (LASSO)
- adds a penalty on the sum of the amplitudes, weighted by λ
- automatically zeroes out non-contributing terms



Model selection with grid search

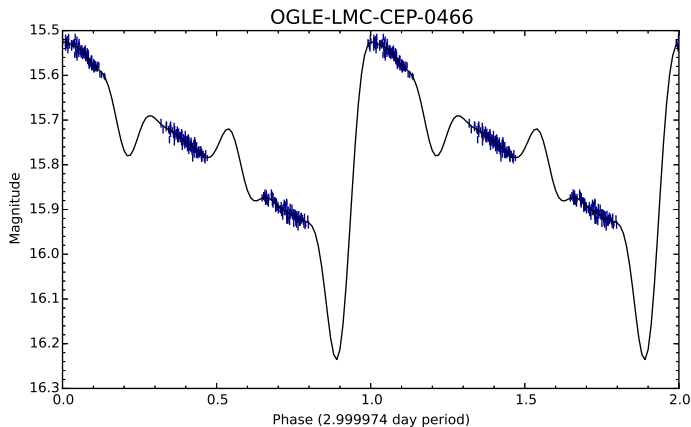
- use grid search with cross-validation
 - search over the order of fit n
- cross-validation helps fit underlying function, not just the data



Results

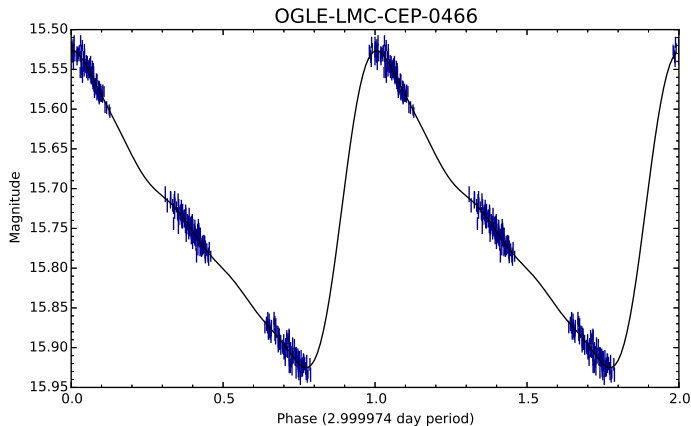


OLS/Baart light curve



10th order fit using OLS/Baart.

LASSO light curve



10th order fit using Lasso/Grid Search.



Performance of LASSO/grid search versus OLS/Baart

Galaxy	Type	Stars	N (SD)	LASSO R^2 (MAD)	Baart R^2 (MAD)	Significance
(all)	(all)	52844	643.1 (462.0)	0.8594 (0.1741)	0.8492 (0.1864)	$p < .0001$
(all)	CEP	7999	740.1 (298.4)	0.9816 (0.0191)	0.9810 (0.0198)	$p < .0001$
(all)	T2CEP	596	747.6 (612.0)	0.9145 (0.1159)	0.9009 (0.1328)	$p < .0001$
(all)	ACEP	89	497.3 (225.0)	0.9700 (0.0245)	0.9689 (0.0267)	$p < .0001$
(all)	RRLYR	44160	624.4 (481.6)	0.8316 (0.1816)	0.8197 (0.1926)	$p < .0001$
LMC	(all)	28491	522.3 (227.7)	0.7812 (0.1695)	0.7723 (0.1779)	$p < .0001$
LMC	CEP	3342	536.8 (219.7)	0.9840 (0.0172)	0.9833 (0.0180)	$p < .0001$
LMC	T2CEP	201	538.3 (232.6)	0.8672 (0.1569)	0.8599 (0.1653)	$p < .0001$
LMC	ACEP	83	477.3 (214.7)	0.9704 (0.0233)	0.9701 (0.0245)	$p < .0001$
LMC	RRLYR	24865	520.3 (228.6)	0.7544 (0.1667)	0.7452 (0.1755)	$p < .0001$
SMC	(all)	7146	851.4 (256.7)	0.9109 (0.1241)	0.9091 (0.1266)	$p < .0001$
SMC	CEP	4625	886.5 (256.2)	0.9800 (0.0195)	0.9796 (0.0200)	$p < .0001$
SMC	T2CEP	42	891.2 (241.4)	0.7965 (0.2235)	0.7888 (0.2379)	$p < .0001$
SMC	ACEP	6	774.3 (190.2)	0.9277 (0.0709)	0.9272 (0.0706)	$p = 0.2188$
SMC	RRLYR	2473	785.2 (244.8)	0.6299 (0.1915)	0.6203 (0.1962)	$p < .0001$
BLG	(all)	17207	756.8 (698.1)	0.9579 (0.0445)	0.9527 (0.0514)	$p < .0001$
BLG	CEP	32	824.2 (569.0)	0.9742 (0.0342)	0.9703 (0.0396)	$p < .0001$
BLG	T2CEP	353	849.7 (746.8)	0.9525 (0.0643)	0.9457 (0.0747)	$p < .0001$
BLG	RRLYR	16822	754.7 (697.2)	0.9581 (0.0440)	0.9528 (0.0509)	$p < .0001$

Median coefficients of determination (R^2) and median absolute deviations (MAD) for models selected by cross-validated LASSO and Baart's ordinary least squares on OGLE *I*-band photometry. P-values obtained by paired Mann-Whitney *U* tests.



Missing harmonics

- LASSO makes no distinction between higher and lower order terms
 - if it doesn't contribute, it goes to zero
- this can result in $A_i = 0$, when $A_j \neq 0$, $j > i$
 - contrary to pulsation models, which say amplitude decreases with order

$$A_1 > A_2 > \dots > A_n$$

- explanations:
 - harmonics absent from observations
 - e.g. we observe only near zero-crossing
 - interference pattern in pulsation (gets political)
 - others? (please tell me)



Multifrequency variable stars

$$m(t) = A_0 + \sum_{k_1=-n}^n \dots \sum_{k_p=-n}^n A_{\mathbf{k}} \cos((\mathbf{k} \cdot \boldsymbol{\omega})t + \Phi_{\mathbf{k}})$$

$$\mathbf{k} \rightarrow (k_1 \quad \dots \quad k_p) \quad \boldsymbol{\omega} \rightarrow (\omega_1 \quad \dots \quad \omega_p)$$




- some variable stars oscillate with multiple (p) periods
- OLS fails to accurately fit these light curves
 - tools exist to manually fix certain amplitudes to zero
- LASSO successful in automatically zeroing out amplitudes

Bellinger, Wysocki, and Kanbur (2015a)



References



-  Earl Bellinger, Daniel Wysocki, and Shashi Kanbur. Measuring amplitudes of harmonics and combination frequencies in variable stars. *Arxiv e-prints*, November 2015. arXiv: 1512.00004 [astro-ph.SR].
-  Earl Bellinger, Daniel Wysocki, and Shashi Kanbur. Plotypus: light curve analysis in Python. [Online; accessed 2015-04-04]. 2015. URL: <https://astrowego.github.io/plotypus/>.
-  Joseph Fourier. Mémoire sur la propagation de la chaleur dans les corps solides. *Nouveau bulletin des sciences de la société philomathique de paris*, 6:112–116, 1808.