Period Estimation, Modeling, and Finding Structure in the Milky Way Halo

James Long

September 21, 2016

Period Uncertainty Quantification

Multiband Period Estimation

Data Based Models, Finding Structure in Milky Way Halo

Outline

Period Uncertainty Quantification

Multiband Period Estimation

Data Based Models, Finding Structure in Milky Way Halo

Frequentist Asymptotics

Model:

$$y_i = \beta_0 + a\sin\big(t_i\omega + \phi\big) + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$.

Maximum likelihood frequency estimate:

$$\widehat{\omega} = \underset{\omega}{\operatorname{argmin}} \ \min_{\beta_0, a, \phi} \sum_{i=1}^{n} (y_i - \beta_0 - a\sin(\omega t_i + \phi))^2$$

Question: How do we assess uncertainty on $\widehat{\omega}$?

Asymptotics of Maximum Likelihood Estimators

Under regularity conditions on model $f(x|\theta)$

$$\widehat{\theta}_{ML} \rightarrow_P \theta$$

and

$$\sqrt{n}(\widehat{\theta}_{ML} - \theta) \rightarrow_d N(0, I(\theta)^{-1})$$

where

$$I(\theta) = \mathbb{E}\left[-\frac{d^2}{d\theta^2}\log f(x|\theta)\right]$$

Usually report $\widehat{\theta}_{ML}$ and $\widehat{I}(\widehat{\theta}_{ML})$.

Simple Periodic Model

Consider the simple model

$$y_i = \sin(\omega t_i) + \epsilon_i$$

where $\epsilon_i \sim N(0, 1)$.

The data is $x_i = (t_i, y_i)$ iid where

$$f(x_i|\omega) = f(y_i|\omega, t_i) f_T(t_i)$$
$$f(y_i|\omega, t_i) = N(y|\sin(\omega t_i), 1).$$

$$\widehat{\omega} = \underset{\omega}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \sin(\omega t_i))^2$$

Fisher Information

$$
\begin{aligned}
I(\omega) &= \mathbb{E}[-(d^2/d\omega^2)\log f(x|\omega)] \\
&= \mathbb{E}[(d^2/d\omega^2)\frac{(y-\sin(\omega t))^2}{2}] \\
&= \mathbb{E}[(d/d\omega) - (y-\sin(\omega t))(\cos(\omega t))t] \\
&= \mathbb{E}[cos^2(\omega t)t^2 + (y-\sin(\omega t))\sin(\omega t)t^2] \\
&= \mathbb{E}[cos^2(\omega t)t^2]
\end{aligned}
$$

So our variance estimate is

$$
\widehat{\mathrm{Var}}(\widehat{\omega}) = \frac{\widehat{I}(\widehat{\omega})^{-1}}{n} = \frac{1}{\sum_{i=1}^{n} t_i^2 \cos^2(\widehat{\omega} t_i)}
$$

Code Example: `https://try.jupyter.org/`
`http://stat.tamu.edu/~jlong/astrostat/period_uncertainty.ipynb`

Full Model: Same Idea, Many More Details

The estimator is

$$\widehat{\theta} = \underset{(\beta_0, a, \phi, \omega, \sigma^2)}{\text{argmin}} \; n \log \sigma + \sum_{i=1}^{n} \frac{(y_i - \beta_0 - a \sin(\omega t_i + \phi))^2}{\sigma^2}$$

The fisher matrix $I(\theta)$ is a 5×5 matrix, so several partial derivatives to compute.

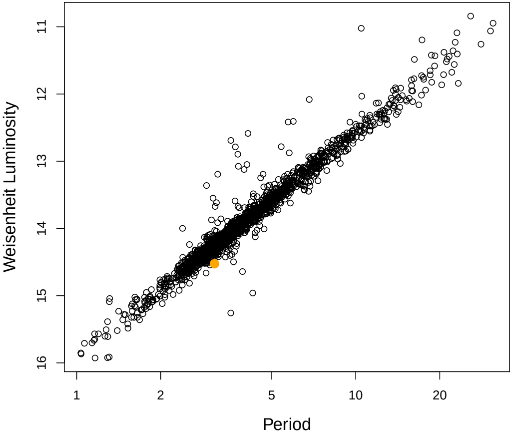
Report

$$\widehat{\theta}_4 = \widehat{\omega} \qquad \widehat{\text{Var}}\,(\widehat{\omega}) = n^{-1} \widehat{I}(\widehat{\theta})^{-1}_{(4,4)}$$

Issues:

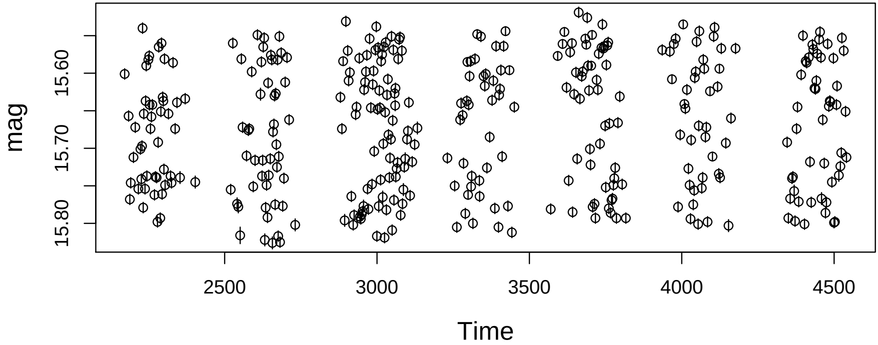
- sometimes heteroskedasticity present with σ_i known
- is the model correct? sandwich estimators?
- **infill** versus **outfill** asymptotics

Period–Luminosity Relation for Cepheids in LMC

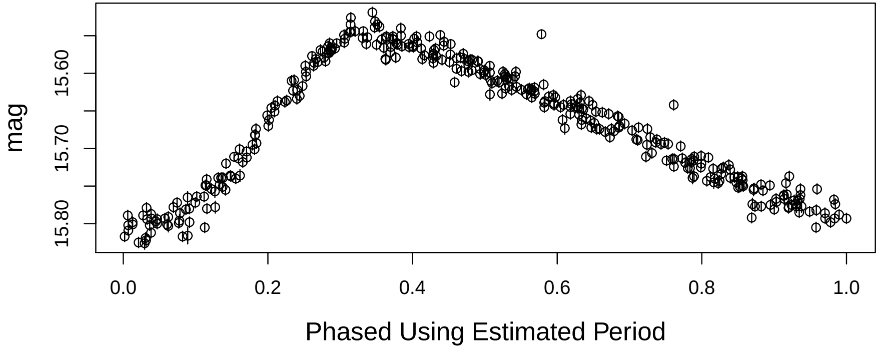


Is uncertainty in periods causing additional scatter in the relationship, biasing slope towards 0?

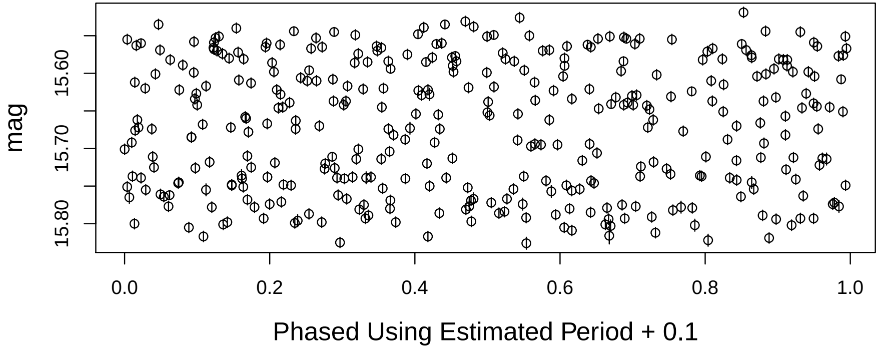
Orange Dot Light Curve from Last Slide



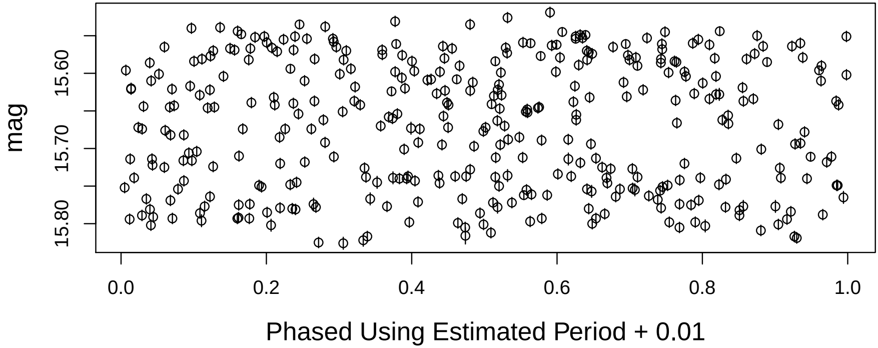
Folded on Estimated Period



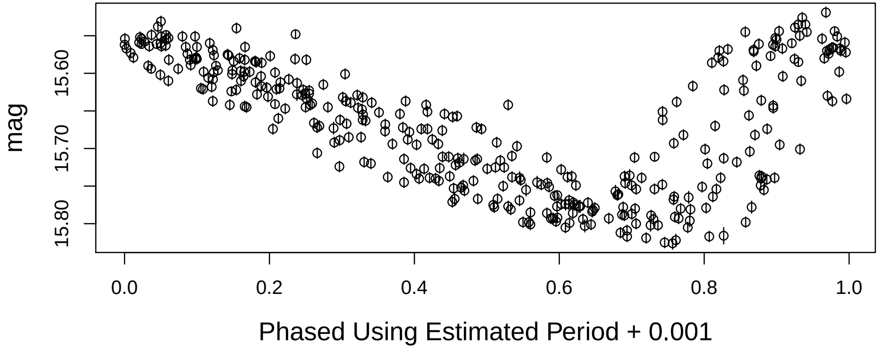
Folded on Estimated Period + 0.1



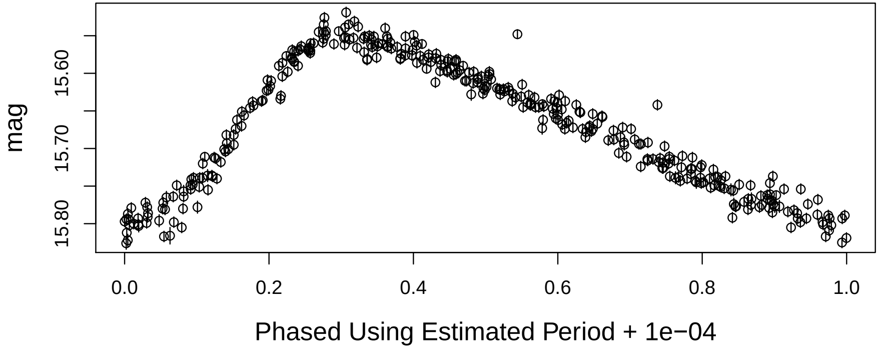
Folded on Estimated Period + 0.01



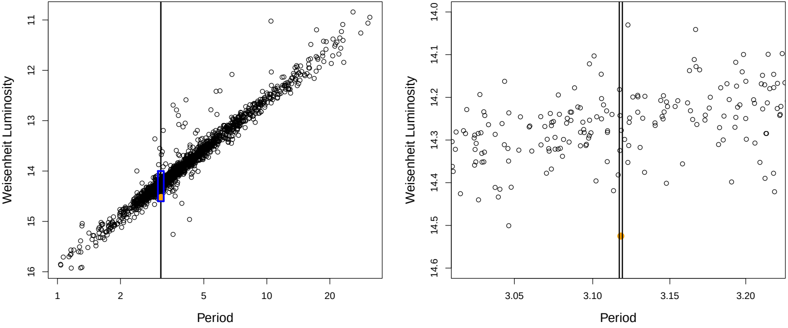
Folded on Estimated Period + 0.001



Folded on Estimated Period + 0.0001



Zoom in Around Orange Light Curve



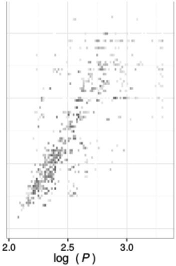
The black vertical lines are 0.001 days on either side of the point estimate. It does not appear that uncertainty in periods is causing scatter in PL relationship (at least for this light curve).

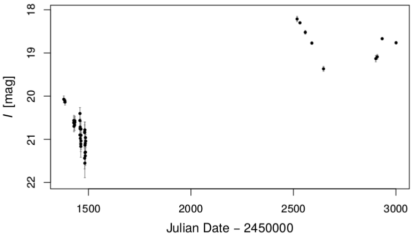
Summary

- visual inspection: period estimate accurate to at least 0.001 days $= 86$ seconds
- are uncertainties in periods affecting estimate of PL relation?
  - not in this case
  - focus on propagating uncertainty in luminosity, addressing potential non–linearity in PL, outliers
  - do not have to address an "errors–in–variables" type problem

"It is inappropriate to be concerned about mice when there are tigers abroad." - George Box

M33 Mira PL Relation: Period Uncertainty Matters





Message: Need a hierarchical model which connects uncertainty on individual period estimates to uncertainty on slope, y–intercept of PL relation. Tom Loredo's talk on October 19.

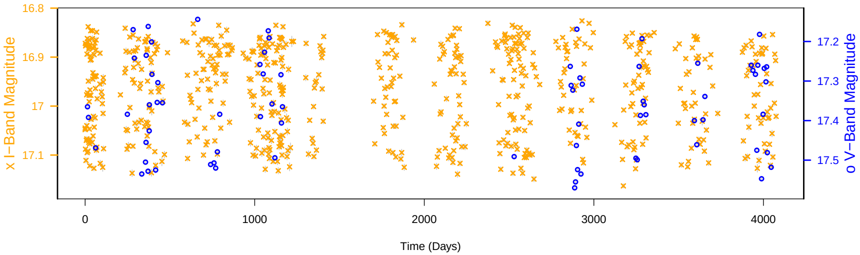
Outline

Period Uncertainty Quantification

Multiband Period Estimation

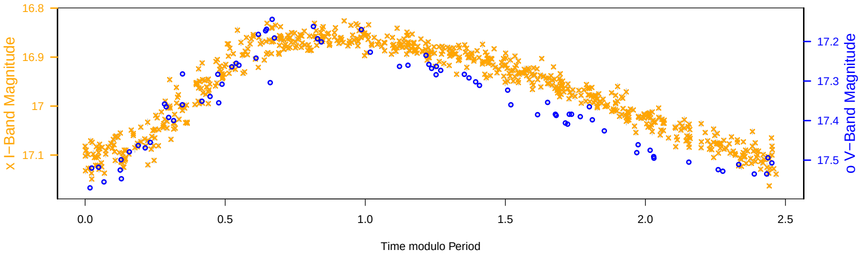
Data Based Models, Finding Structure in Milky Way Halo

Unfolded Light Curve



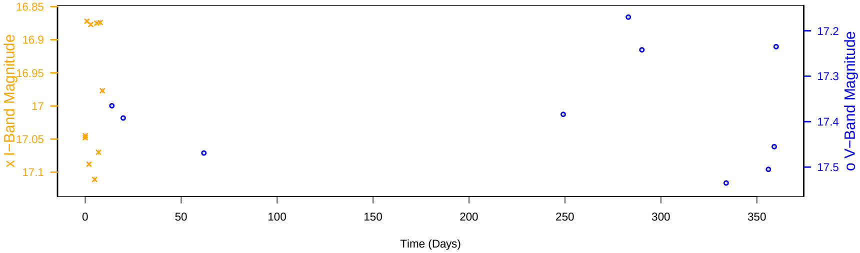
Well sampled light curve.

Folded Light Curve



Period estimated using I band of data.

Poorly Sampled Light Curve



Poorly sampled light curve. Here need period estimation method which uses all bands of data

Multiband Lomb Scargle

Fit a separate sinusoid to each band with a common frequency ω.

Model:

$$m_{bi} = \beta_b + a_b \sin(\omega t_{bi} + \rho_b) + \epsilon_{bi}$$
$$\epsilon_{bi} \sim N(0, \sigma_{bi}^2) \text{ independent}$$

Maximum Likelihood Frequency Fit:

Let: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_B)$, $\mathbf{a} = (a_1, \ldots, a_B)$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_B)$.

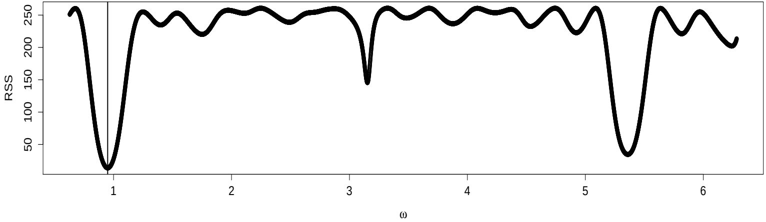
$$RSS(\omega, \boldsymbol{\beta}, \mathbf{a}, \boldsymbol{\rho}) = \sum_{b=1}^{B} \sum_{i=1}^{n_b} \left(\frac{m_{bi} - a_b \sin(\omega t_{bi} + \rho_b) - \beta_b}{\sigma_{bi}} \right)^2$$

Frequency estimate is $\widehat{\omega} = \underset{\omega}{\text{argmin}} \; \min_{\mathbf{a},\boldsymbol{\rho},\boldsymbol{\beta}} RSS(\omega, \boldsymbol{\beta}, \mathbf{a}, \boldsymbol{\rho})$.

Computing the ML Estimate $\widehat{\omega}$

- Grid search across frequency.
- For fixed frequency ω, weighted least squares on band b to compute (β_b, a_b, ρ_b):

$$\begin{aligned} m_{bi} &= \beta_b + a_b \sin(\omega t_{bi} + \rho_b) + \epsilon_{bi} \\ &= \beta_b + \underbrace{a_b \cos(\rho_b)}_{\equiv \beta_{b1}} \sin(\omega t_{bi}) + \underbrace{a_b \sin(\rho_b)}_{\equiv \beta_{b2}} \cos(\omega t_{bi}) + \epsilon_{bi} \end{aligned}$$



Algorithm developed by Lomb and Scargle for single band data [4, 8].

Notes

- $3B + 1$ total parameters
- simple multiband methods sufficient in certain regimes
  - ≈ 20 observations / band in 5 bands
- fails with very sparsely sampled data eg PanStarrs, DES
  - $5 - 10$ obs / band in 5 bands

Idea: Constrain the number of free parameters to less than $3B + 1$.

Experiment with OGLE Data

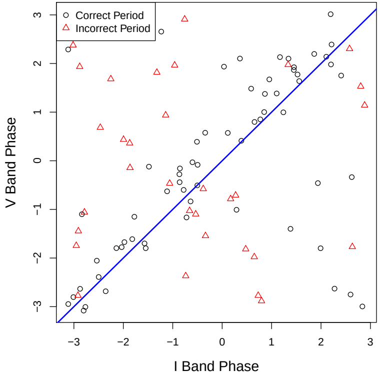
- downsample ~ 100 Cepheid periodic variables to 10 observations in I and V band.
- estimate periods using sine function

Result: 34% of period estimates incorrect.

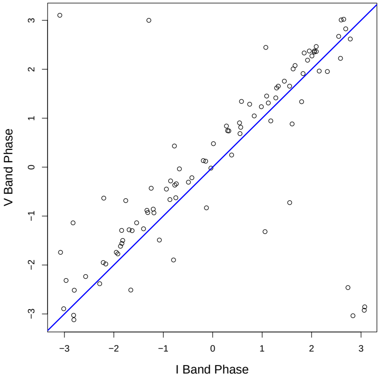
Question: What are the best fit phases and amplitudes?

Phase Estimates

10 observations / band



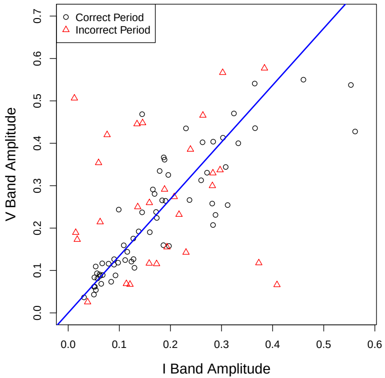
Full light curves



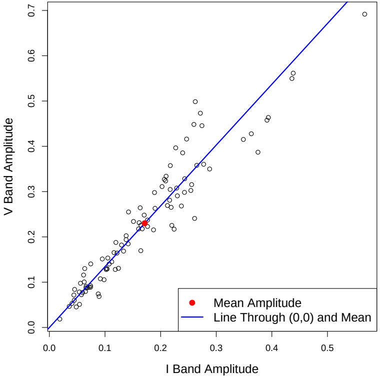
Message: Many phase estimates for the poorly sampled light curves (left) are unrealistic. An algorithm which imposes phase constraints may improve period estimation accuracy.

Amplitude Estimates

10 observations / band



Full light curves



Message: Many amplitude estimates for the poorly sampled light curves (left) are unrealistic. An algorithm which imposes amplitude constraints may improve period estimation accuracy.

Penalized Negative Log Likelihood (PNLL)

Modify likelihood so unrealistic $\boldsymbol{\rho}$ and $\mathbf{a}$ are penalized:

$$
\begin{aligned}
&PNLL(\omega, \boldsymbol{\beta}, \mathbf{a}, \boldsymbol{\rho}) \\
&= \underbrace{\frac{1}{2} \sum_{b=1}^{B} \sum_{i=1}^{n_b} \left(\frac{m_{bi} - a_b \sin(\omega t_{bi} + \rho_b) - \beta_b}{\sigma_{bi}} \right)^2}_{=RSS(\omega,\boldsymbol{\beta},\mathbf{a},\boldsymbol{\rho})} + \underbrace{J(\boldsymbol{\rho}) + J(\mathbf{a})}_{\text{penalty terms}}
\end{aligned}
$$

Choose best frequency:

$$
\widehat{\omega} = \underset{\omega}{\operatorname{argmin}} \; \min_{\boldsymbol{\beta},\mathbf{a},\boldsymbol{\rho}} PNLL(\omega, \boldsymbol{\beta}, \mathbf{a}, \boldsymbol{\rho})
$$

Choosing $J(\boldsymbol{\rho})$

Define:

$$\bar{\rho} = \frac{1}{B} \sum_{b=1}^{B} \rho_b$$

Let:

$$\begin{aligned} J(\boldsymbol{\rho}) &= \frac{\gamma_1}{2} \sum_{b=1}^{B} (\rho_b - \bar{\rho})^2 \\ &= \frac{\gamma_1}{2} \left|\left| \boldsymbol{\rho} - \frac{\mathbf{1}^T \boldsymbol{\rho}}{\mathbf{1}^T \mathbf{1}} \mathbf{1} \right|\right|_2^2 \\ &= \frac{\gamma_1}{2} \boldsymbol{\rho}^T \left[\mathbf{I} - \frac{1}{B} \mathbf{1} \mathbf{1}^T \right] \boldsymbol{\rho} \end{aligned}$$

Choosing $J(\mathbf{a})$

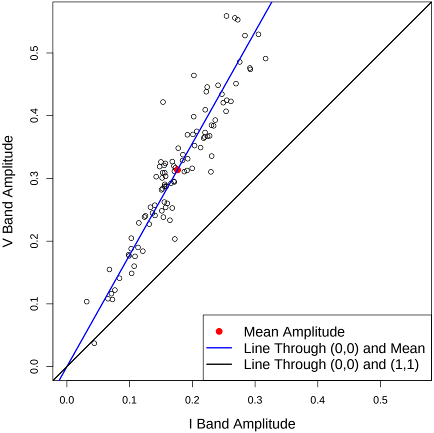
Could choose $J(\mathbf{a})$ in same manner as $J(\boldsymbol{\rho})$:

- Let $\bar{a} = \frac{1}{B}\sum_{b=1}^{B} a_b$.
- $J(\mathbf{a}) = (\gamma_2/2)\sum_{b=1}^{B}(a_b - \bar{a})^2$

But amplitudes tend to be different across bands.

Need better method.

I and V Amplitudes for 100 RR Lyrae AB



Choosing $J(\mathbf{a})$

- Let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^B$ be amplitude estimates for a set of well observed light curves.
- Define:

$$\tilde{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{a}_i \Big/ \left|\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{a}_i\right|\right|_2$$

- The penalty is:

$$\begin{aligned} J(\mathbf{a}) &= \frac{\gamma_2}{2}\left|\left|\mathbf{a} - (\mathbf{a}^T\tilde{\mathbf{a}})\tilde{\mathbf{a}}\right|\right|_2^2 \\ &= \frac{\gamma_2}{2}\mathbf{a}^T\left[\mathbf{I} - \tilde{\mathbf{a}}\tilde{\mathbf{a}}^T\right]\mathbf{a} \end{aligned}$$

$J(\mathbf{a})$ pushes the amplitudes towards the scaled mean vector $\tilde{\mathbf{a}}$.

Minimization of PNLL

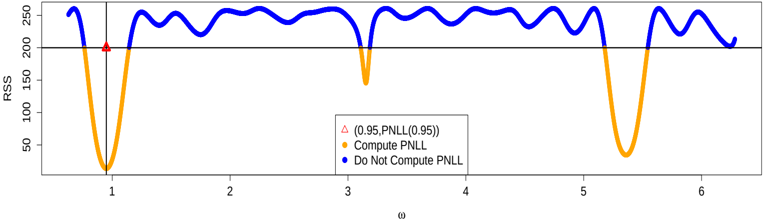
$$PNLL(\omega, \boldsymbol{\beta}, \mathbf{a}, \boldsymbol{\rho}) = \sum_{b=1}^{B} \sum_{i=1}^{n_b} \left(\frac{m_{bi} - a_b \sin(\omega t_{bi} + \rho_b) - \beta_b}{\sigma_{bi}} \right)^2 + \gamma_1 \boldsymbol{\rho}^T \left[\mathbf{I} - \frac{1}{B} \mathbf{1}\mathbf{1}^T \right] \boldsymbol{\rho} + \gamma_2 \mathbf{a}^T \left[\mathbf{I} - \tilde{\mathbf{a}}\tilde{\mathbf{a}}^T \right] \mathbf{a}$$

Outline of Minimization Procedure:

- Grid search on ω.
- For fixed ω, block coordinate descent on $\boldsymbol{\beta}, \mathbf{a}, \boldsymbol{\rho}$.
- For fixed ω and $\boldsymbol{\rho}$, closed form solutions for $\boldsymbol{\beta}$ and $\mathbf{a}$ updates.
- $\boldsymbol{\rho}$ update uses Majorization–minimization (MM).

Compute PNLL on Subset of Frequencies

- Do not compute $PNLL(\omega)$ at all ω.
  - $PNLL(\omega) > RSS(\omega) \, \forall \, \omega$
  - $PNLL(\omega_1) < RSS(\omega_2) \implies \omega_2 \neq \underset{\omega \in \Omega}{\text{argmin}} \, PNLL(\omega)$



- R package `multiband` on `CRAN`.

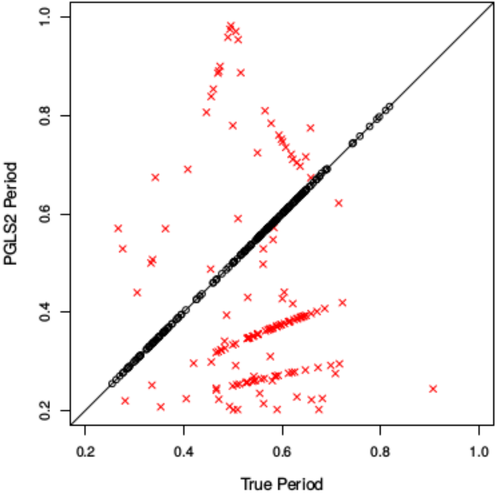
Notes

- computation time goes up
- uncertainty harder to quantify
- model is more tuned to specific variable star classes
- PNLL developed in Long [5]
- see Vanderplas [12] for related work

Testing Algorithm on Quasi–Real Data

- $\approx$ 400 well sampled RR Lyrae light curves in 5 bands (SDSS Strip 82)
- downsample to 10 observations / band.
- estimate period by minimizing PNLL

Results



Period estimates with 10 observations / band versus truth.

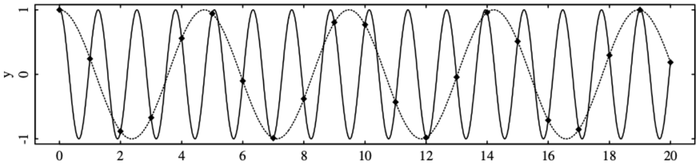
Pattern in Errors: Pseudoaliasing

Aliasing: Suppose we observe star at times $t_i = i$ and

$$\omega' = |\omega + 2\pi k|$$

for some $k \in \mathbb{Z}$. Then

$$\begin{aligned}\sin(\omega t_i + \phi) &= \sin((\omega + 2\pi k)t_i - 2\pi k t_i + \phi) \\ &= \sin((\text{sign}(\omega + 2\pi k)\omega')t_i - 2\pi k i + \phi) \\ &= \sin(\omega' t_i + (\pi - 2\phi)\mathbb{1}_{\{\omega + 2\pi k < 0\}} + \phi)\end{aligned}$$

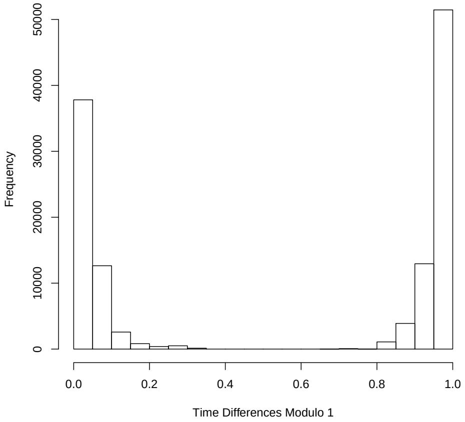


The frequencies ω and ω' are aliased and cannot be separated.

Pseudoaliasing: If observation times are approximately equally spaced, approximate aliasing.

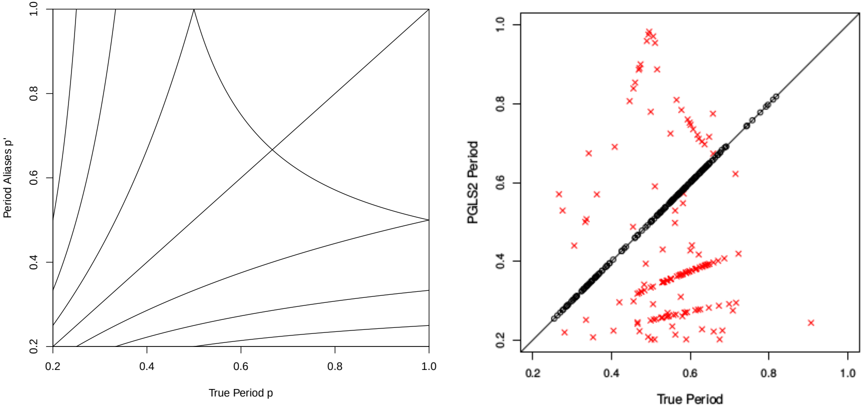
Histogram of Delta Times Modulo 1 Day

Compute $(t_{b,i+1} - t_{b,i})$ mod 1 for RR Lyrae in SDSS–Stripe 82.



Usually, we observe star at around same time each night.

Pseudoaliasing for 1 / day Sampling



Eventually asymptotics will take hold and dispersion around truth will be normally distributed. But the asymptotic approximation can provide a very poor approximation at small to medium *n*.

Outline

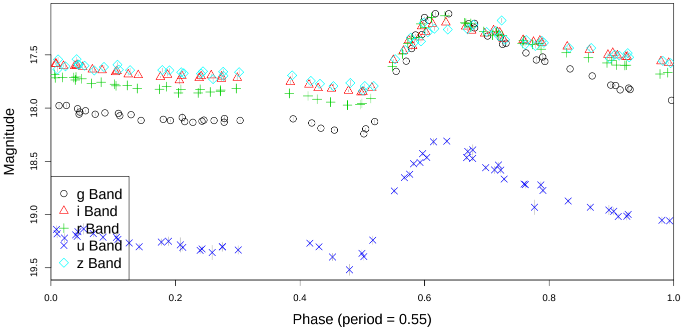
Period Uncertainty Quantification

Multiband Period Estimation

Data Based Models, Finding Structure in Milky Way Halo

Folded Light Curve of Periodic Variable

Folded light curve: Brightness versus time modulo period.

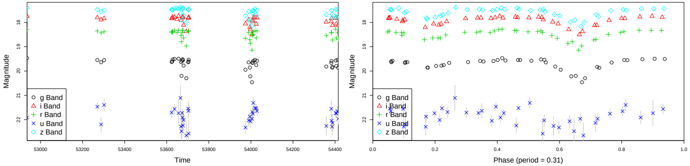


SDSS–III Stripe 82 [2]

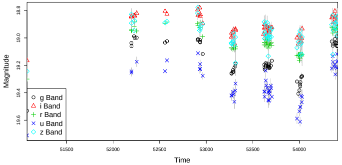
- collected $\approx 60,000$ variables in Stripe 82 Region
- variables belong to different classes
- some periodic, some not

Three Examples

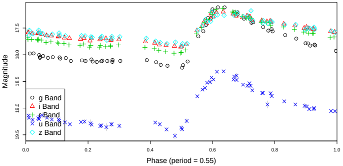
Eclipsing Binary (Unfolded and Folded)



Quasar (unfolded)



RR Lyrae (folded)

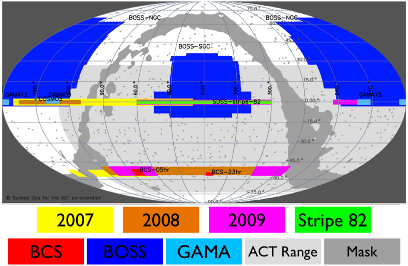


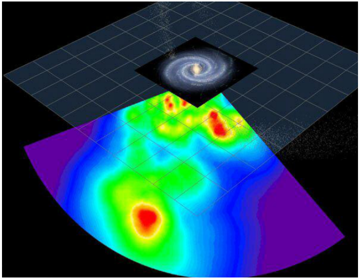
Example of Variable Star Science

- spiral galaxy formation simulation: `https://www.youtube.com/watch?v=s-25dEcY-WU`
- RR Lyrae variable star distance is function of mean brightness in V band:

$$d = 10^{(V - M_V + 5)/5}$$

- discover structure in the Milky Way halo by finding, estimating distances to RR Lyrae





Sesar [10] used SDSS Stripe 82 Data to make Halo Map

- compare observed structure to simulations

Mapping Galactic Halo with DES

Dark Energy Survey (DES)

- 10 photometric measurements in (g,i,r,z,Y) over five years
- depths to 24 mag in i
- 5000 square degrees
- few million variable stars

DES is deeper and wider than SDSS Stripe 82, but much more sparsely sampled.

Sinusoid Based Models

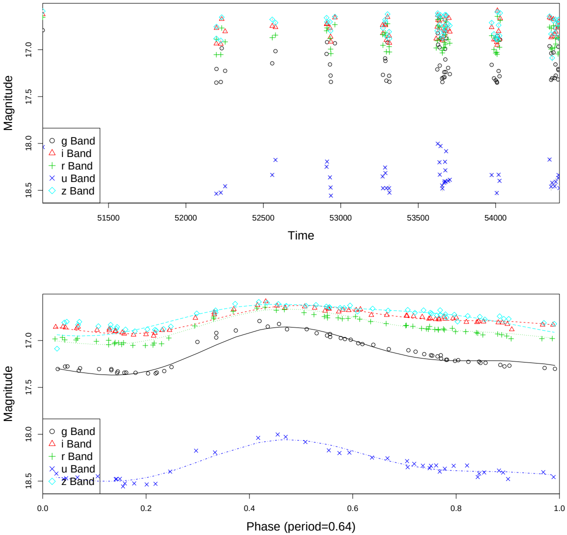
Method: Model light curve variation in each filter as sinusoid with K harmonics. [6, 14, 4, 8, 9]

$$m_{jb} = \beta_b + \sum_{k=1}^{K} a_{bk} \sin(\omega t_{jb} + \phi_{bk}) + \epsilon_{jb}$$

where $\epsilon_{jb} \sim N(0, \sigma_{jb}^2)$.

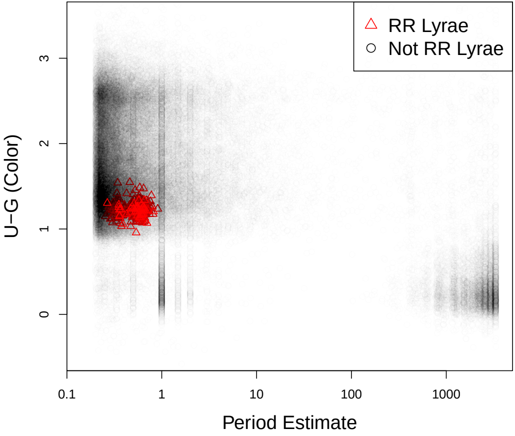
- $(2K+1)B + 1$ parameters. Pure sine $3B + 1$ parameters.
- Maximum likelihood has closed form solution at fixed ω
  - Maximization strategy: Grid search on ω.

Example of Maximum Likelihood Fit with $K = 2$

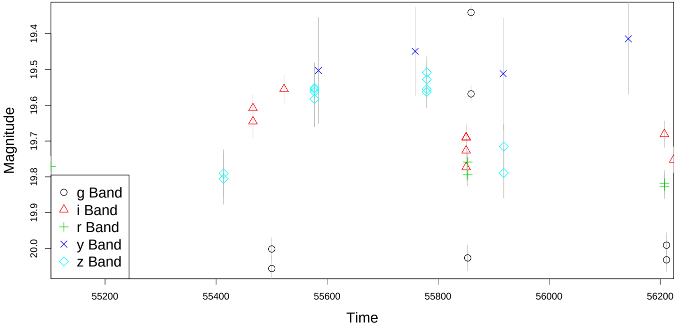


Total of 26 parameters, model estimates period well.

Output from Model Useful for Classification



Problem: Sparsely Sampled Light Curves



Pan–STARRS light curve. Period estimation difficult for this quality data.

Building Physics Into Models

Idea: Use physics / existing data to build model specifically targeted at particular variable star class. Sometimes called templates.

Specifically:

- amplitude ratio between bands approx constant
- one phase parameter for all bands
- shape varies by band, but within band same for different lcs
- mean
  - distance reduces brightness
  - certain bands are brighter than others
  - dust decreases brightness
    - more line of sight dust to some stars than others
    - proportion for dimming (extinction) fixed ratio across bands

Parsimonious RR Lyrae Model

global parameters fit once for all RR Lyrae

magnitude offset band b | extinction law band b | **RR Lyrae shape**

$$m_b(t) = \mu + M_b + E[B-V]R_b + a\gamma_b(\omega t + \phi)$$

distance modulus | amount of dust | amplitude | frequency | phase

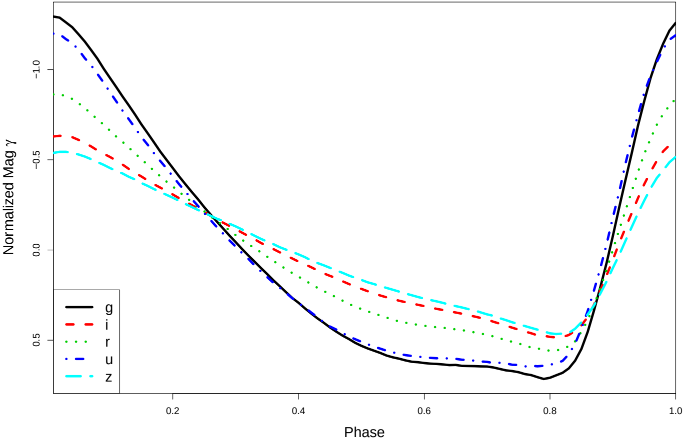
individual parameters fit for each RR Lyrae

- data $D = \{\{t_{jb}, m_{jb}, \sigma_{jb}\}_{j=1}^{n_b}\}_{b=1}^{B}$
- normal measurement error model:

$$m_{jb} = m_b(t_{jb}) + \epsilon_{jb}$$

where $\epsilon_{jb} \sim N(0, \sigma_{jb}^2)$.

γ_b Estimated from SDSS Stripe 82 RR Lyrae



Parameter Estimation

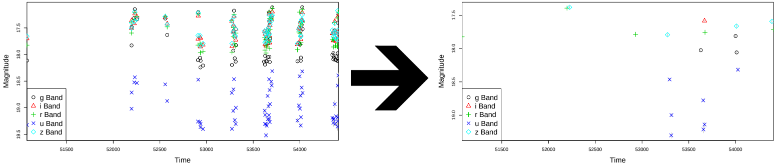
$$
\begin{aligned}
& RSS(\omega, \mu, E[B-V], a, \phi) \\
\equiv & \sum_{b=1}^{B} \sum_{j=1}^{n_b} \left(\frac{m_{jb} - \mu - M_b - E[B-V]R_b - a\gamma_b(\omega t_{jb} + \phi)}{\sigma_{jb}} \right)^2
\end{aligned}
$$

Estimate parameters with maximum likelihood (χ^2 minimization):

- Likelihood is highly multimodal in ω, grid search.
- Model is linear in μ,$E[B-V]$, and a, closed form updates.
- Warm start Newton–Raphson updates for ϕ.

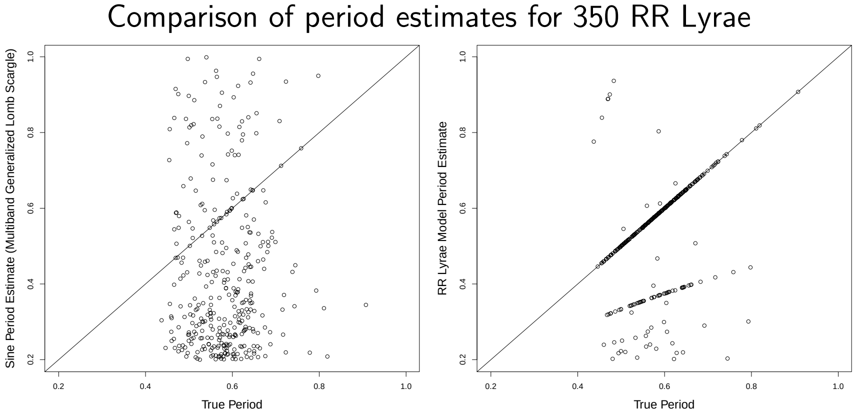
Simulation

Downsample Stripe 82 variables to 20 observations across all bands:

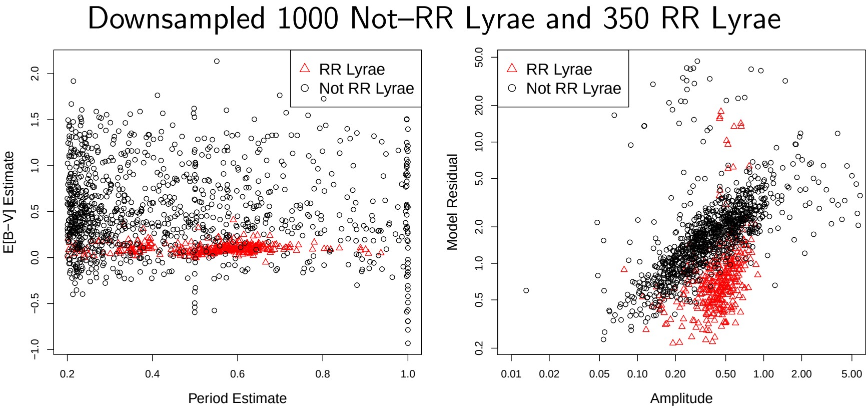


- Can we estimate periods correctly for RR Lyrae?
- Can we separate RR Lyrae from non–RR Lyrae?
- Can we estimate distances accurately?
- Can we reproduce halo maps of Sesar 2010?

Simulation Results for Period Estimation



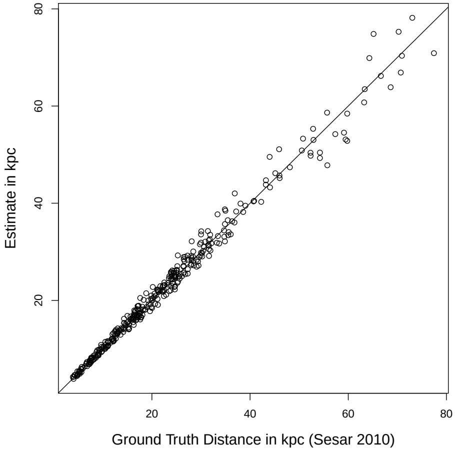
Simulation Results for Classification



- visually good separation with small number of features
- potential to use model output as feature input to classifier

Simulation Results for Distance Estimation

- Use model to estimate distance modulus (μ) for RR Lyrae.
- Convert μ to distance (d) in parsecs: $d = 10^{\mu/5+1}$



Ongoing Work

- model refinements:
  - M_b dependence on period, metallicity.
  - bands other than g, i, u, r, z
  - parameterize γ_b to account for shape differences across RRL
- propagate uncertainty to 3D halo density estimate
  - treat as Inference Problem: Bayesian posterior or frequentist confidence bands. Hierarchical model?
  - treat as Prediction Problem: Evaluate distance / density estimates using Stripe 82 data or follow–up observations
- combining modeling and feature extraction
  - output from RRL model useful for classification
  - features / summary statistics computed also useful
- Milky Way Halo maps using DES data

Bibliography I

[1] Nina Hernitschek, Edward F Schlafly, Branimir Sesar, Hans-Walter Rix, David W Hogg, Željko Ivezić, Eva K Grebel, Eric F Bell, Nicolas F Martin, WS Burgett, et al.
Finding, characterizing, and classifying variable sources in multi-epoch sky surveys: Qsos and rr lyrae in ps1 3π data.
The Astrophysical Journal, 817(1):73, 2016.

[2] Željko Ivezić, J Allyn Smith, Gajus Miknaitis, Huan Lin, Douglas Tucker, Robert H Lupton, James E Gunn, Gillian R Knapp, Michael A Strauss, Branimir Sesar, et al.
Sloan digital sky survey standard star catalog for stripe 82: The dawn of industrial 1% optical photometry.
The Astronomical Journal, 134(3):973, 2007.

[3] Géza Kovács and G Kupi.
Computation of the fourier parameters of rr lyrae stars by template fitting.
Astronomy & Astrophysics, 462(3):1007–1016, 2007.

[4] Nick R Lomb.
Least-squares frequency analysis of unequally spaced data.
Astrophysics and space science, 39(2):447–462, 1976.

[5] James P Long, Eric C Chi, and Richard G Baraniuk.
Estimating a common period for a set of irregularly sampled functions with applications to periodic variable star data.
arXiv preprint arXiv:1412.6520, 2014.

[6] N Mondrik, JP Long, and JL Marshall.
A multiband generalization of the multiharmonic analysis of variance period estimation algorithm and the effect of inter-band observing cadence on period recovery rate.
The Astrophysical Journal Letters, 811(2):L34, 2015.

[7] Ondřej Pejcha and Christopher S Kochanek.
A global physical model for cepheids.
The Astrophysical Journal, 748(2):107, 2012.

Bibliography II

[8] Jeffrey D Scargle.
Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data.
The Astrophysical Journal, 263:835–853, 1982.

[9] A Schwarzenberg-Czerny.
Fast and statistically optimal period search in uneven sampled observations.
The Astrophysical Journal Letters, 460(2):L107, 1996.

[10] Branimir Sesar, Željko Ivezić, Skyler H Grammer, Dylan P Morgan, Andrew C Becker, Mario Jurić, Nathan De Lee, James Annis, Timothy C Beers, Xiaohui Fan, et al.
Light curve templates and galactic distribution of rr lyrae stars from sloan digital sky survey stripe 82.
The Astrophysical Journal, 708(1):717, 2010.

[11] RF Stellingwerf.
Period determination using phase dispersion minimization.
The Astrophysical Journal, 224:953–960, 1978.

[12] Jacob T VanderPlas and Željko Ivezic.
Periodograms for multiband astronomical time series.
The Astrophysical Journal, 812(1):18, 2015.

[13] Peter Yoachim, Les P McCommas, Julianne J Dalcanton, and Benjamin F Williams.
A panoply of cepheid light curve templates.
The Astronomical Journal, 137(6):4697, 2009.

[14] M Zechmeister and M Kürster.
The generalised lomb-scargle periodogram-a new formalism for the floating-mean and keplerian periodograms.
Astronomy & Astrophysics, 496(2):577–584, 2009.