Period analysis of variable stars by robust smoothing

Hee-Seok Oh

University of Alberta, Edmonton, Canada

Doug Nychka, Tim Brown and Paul Charbonneau

National Center for Atmospheric Research, Boulder, USA

Summary. The objective of this paper is to estimate the period and the light curve (or periodic function) of a variable star. Previously, several methods have been proposed to estimate the period of a variable star, but they are inaccurate especially when a data set contains outliers. We use a smoothing spline regression to estimate the light curve given a period and then find the period which minimizes the generalized cross-validation (GCV). The GCV method works well, matching an intensive visual examination of a few hundred stars, but the GCV score is still sensitive to outliers. Handling outliers in an automatic way is important when this method is applied in a "data mining" context to a vary large star survey. Therefore, we suggest a robust method which minimizes a robust cross-validation criterion (RCV) induced by a robust smoothing spline regression. Once the period is determined, a nonparametric method is used to estimate the light curve. A real example and a simulation study suggest that RCV and GCV are superior to existing methods.

Keywords: Period; periodic function; generalized cross-validation; robust spline regression; smoothing spline regression

1. Introduction

Variable stars are stars whose brightness changes over time. The class of periodic variable stars are stars whose maxima and minima brightness recur at constant time intervals. The variability of brightness allows for the classification of stars into different groups, according to information on physical properties such as magnitude, the range of period and light curve shape - a plot of the brightness variation of the star in time. It also provides important clues to the structure of the galaxies and stellar evolution (Brown and Gilliland, 1994; Gautschy and Saio, 1995, 1996; Hilditch, 2001). The primary statistical problem associated with classifying a variable star is to estimate its period and its light curve.

Consider a time series $\{y_i, t_i\}$,

$$y_i = f(t_i/p) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where y_i is the ith brightness measurement, t_i is the ith sampling time, ε_i is the ith measurement error, and f is a periodic function or light curve on [0,1] ($f(t/p)$ has period p). The observations from a variable star are unequally spaced, because the data are collected only at certain times of night, sometimes with long interruptions. For this reason we expect that $\{t_i\}$ are unequally spaced. The basic statistical problem is to estimate both f and p.

Several methods have been developed to estimate the period of a variable star. The periodogram and least squares are the two traditional methods for estimating the period using a simple cosine model (Deeming, 1975; Lomb, 1976; Scargle, 1982). Lafler and Kinman

(1965) found the period to minimize a measure of dispersion defined by the function LK(p):

$$\mathrm{LK}(p) = \sum_{i=1}^{n} \left\{ y^*_{i+1}(p) - y^*_i(p) \right\}^2 ,$$

where the y^*_i are the response values sorted by phase ($t_i \bmod p$). Dwortesky (1983) suggested a string-length method that depends on differences in phase as well as in response. The method minimizes the string length given by

$$\mathrm{STR}(p) = \sum_{i=1}^{n} \left\{ \left[y^*_{i+1}(p) - y^*_i(p) \right]^2 + \left[\phi^*_{i+1}(p) - \phi^*_i(p) \right]^2 \right\}^{1/2} ,$$

where the $\phi^*_i(p)$ are the ordered phase values. Stellingwerf (1978) proposed another method based on a measure of dispersion, called phase dispersion minimization (PDM). In this method, the period is chosen to minimize the residual sum of squares (RSS) of the one way analysis of variance, after the phase interval is divided into a number of bins and the mean response is calculated for each bin. This particular method has gained a wide use by astronomers. Recently, Reimann (1994) suggested a nonparametric method to fit the brightness as a function of phase at a given period, using the SuperSmoother, a variable-span local linear smoother developed by Friedman (1984). SuperSmoother performs three running-line smooths of the data (phase, brightness) with long, medium and short span length. Cross-validation is then used to determine the span length that gives the best fit at each phase value. This method finds the period that minimizes the sum of absolute residuals obtained by SuperSmoother fitting, which is given by

$$\mathrm{AR}(p) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i(p)|,$$

where $\hat{y}_i(p)$ are the fitted values from SuperSmoother assuming a period p. Reimann (1994) showed through a simulation study that the cosine method with least squares and the method with SuperSmoother perform better than others. However, an obvious disadvantage with the cosine method is that it does not work well when the true light curve is not sinusoidal. Finally it is clear that any of current methods are ineffective when the data have outliers. This lack of robustness is a practical concern for mining large data bases accumulated for light curve analysis. As one contribution of this paper, empirical results through real and numerical examples show that a properly constructed robust estimator remains high efficiency even when no outliers are present.

1.1. Data

The data sets of variable stars used in this paper are a derived product from the project STellar Astrophysics and Research on Exoplanets (STARE). The primary objective of the STARE project is to use precise photometry to search for extrasolar giant planets transiting their parent stars (Charbonneau et al., 2000). An important byproduct of STARE is a survey of variable stars. For each of thousands of stars in a field, photometry data from the STARE instrument can be used to produce a light curve. Most stars are essentially constant in brightness, but about 10% of the stars are variable. Figure 1-(a) shows the brightness versus time (nights) for a star classified as an eclipsing binary. The data in this

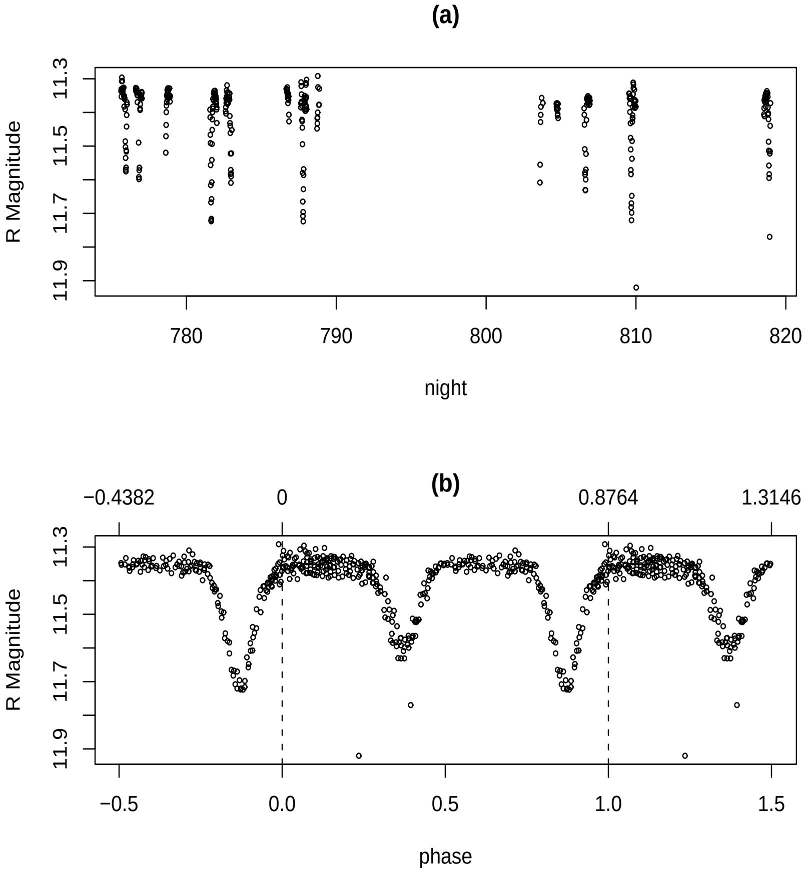

Fig. 1. (a) Brightness of a variable star (an eclipsing binary star) measured in stellar magnitude R, where $R = -2.5\log(F) + C$, F is the flux density from the star and C is a constant and (b) plot of brightness versus phase with $p = 0.8764$ days as period.

figure consist of 351 separate measurements of the star's brightness, taken on 13 nights contained in a 44-night interval. The precision of each measurement is about 0.01 stellar magnitude. When two stars orbit each other in the plane of the observer, the combined brightness decreases when one member of the pair eclipses the other. However, as seen in Figure 1-(a), when the observations are unequally spaced with very long interruptions, the periodicity of the variable star is not obvious. Because the brightness depends on the phase ($=$ time mod p), if the brightness is periodic in time with period, p, then a plot of brightness versus phase will reveal the periodicity. Figure 1-(b) presents a plot with a potential period in the phase domain. The light curve in Figure 1-(b) is produced by folding data over the period of variability. The plot with $p = 0.8764$ (day) clearly reveals a distinct light curve of the star. A light curve generated from the correct period will be useful for classifying the star. For instance, note that from Figure 1-(b), the light curve appears to be flat between eclipses. This feature is associated with the detached (Algol) type eclipsing binaries.

1.2. Outline

In the absence of outliers, we suggest the use of the generalized cross-validation (GCV) score to estimate the period of a variable star. In Section 2, a nonparametric method based on smoothing spline regression is proposed to determine the period of a variable star which minimizes the GCV score. However, with the recognition that a smoothing spline and thus the related GCV score are affected by outliers, we suggest a robust modification. In the robust cross-validation (RCV) method, we estimate the period to minimize the RCV score induced by a robust smoothing spline regression. Once the period is determined by either the GCV or the RCV method, the light curve can be estimated by a nonparametric method such as smoothing splines or SuperSmoother. Conceptually we have found it useful to separate the smoother used to determine the period with that used to estimate the light curve once p is estimated. A theoretical background of the RCV method is briefly mentioned at the end. In Section 3, we compare the GCV and the RCV methods with the existing methods using real brightness data and using a simulation study. As a related topic, we discuss a method to estimate multiple periodicity in Section 4. Some concluding remarks are made in Section 5.

2. Methodology

2.1. Estimation of period: The GCV and the RCV method

Given a period p, $\hat{f}_\lambda(t/p)$, the periodic cubic spline, is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(t_i/p)\}^2 + \lambda\int_{[0,1]}\left\{f^{''}(x)\right\}^2 dx \tag{2}$$

subject to $\int\{f^{''}(x)\}^2 dx < \infty$ and $f(0) = f(1)$, $f^{'}(0) = f^{'}(1)$ (Wahba, 1990).

The GCV score for estimating the period of a variable star is

$$\mathrm{GCV}(p,\lambda) = \frac{\sum_{i=1}^{n}\left\{y_i - \hat{f}_\lambda(t_i/p)\right\}^2}{n\left\{1 - n^{-1}\mathrm{trace}\left[A(p,\lambda)\right]\right\}^2}, \tag{3}$$

where $A(p,\lambda)$ is the smoothing matrix associated with the spline estimate (Hastie and Tibshirani, 1990). It is useful to let $\mathrm{GCV}(p)$ denote the minimum of $\mathrm{GCV}(p,\lambda)$ over $\lambda \in$

$[0, \infty)$ as

$$\mathrm{GCV}(p) = \min_{\lambda} \mathrm{GCV}(p, \lambda). \tag{4}$$

We minimize the GCV score in two steps: for each period p, $\mathrm{GCV}(p)$ is computed and then the period p^* is determined by minimizing $\mathrm{GCV}(p)$ for all p. Applying this method to the data shown in Figure 1-(a), the estimated p is obtained as 0.8762 days. This is not far from 0.8764 days obtained by a visual search method in Figure 1-(b). However, the GCV score (not shown) has some local minima around the global minimum. As mentioned earlier, the smoothing spline regression is a linear estimate of the data and can be severely affected by outliers. The local minima of the GCV score is apparently influenced by two outliers (determined visually) near nights 810 and 820 (Figure 1-(a)). If we compute $\mathrm{GCV}(p)$ ignoring these two outliers, then the GCV score (not shown) does not have any local minima.

Now consider a new method that adopts robust spline regression instead of the usual smoothing spline. The robust smoothing spline can be defined, by replacing the sum of squared errors in (2) by a different function of the errors, as follows: let $\hat{f}_\lambda(t/p)$ be the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left\{y_i - f(t_i/p)\right\} + \lambda \int_{[0,1]} \left\{f^{''}(x)\right\}^2 dx. \tag{5}$$

Here the function $\rho(x)$ is typically convex and increases slower than order x^2 as x becomes large. Huber's favorite is

$$\rho(x) = \begin{cases} x^2 & \text{if } |x| \leq C \\ C(2|x| - C) & \text{otherwise,} \end{cases}$$

where C is a cutoff point usually determined from the data. For C, we follow Huber (1981) and choose $\hat{C} = 1.345 * \mathrm{MAD}$ which ensures 95% efficiency with respect to the normal model in a location problem. Based on this characterization, we consider an idealized robust cross-validation for the smoothing parameter of robust smoothing spline regression as

$$\mathrm{RCV}^*(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \rho\left\{y_i - \hat{f}_{\lambda,-i}(t_i)\right\}, \tag{6}$$

where $\hat{f}_{\lambda,-i}(t_i)$ is the robust smoothing spline when the ith data point, (t_i, y_i) is omitted. The implementation of $\mathrm{RCV}^*(\lambda)$ is not feasible, because the robust spline is a nonlinear estimate and so exhaustive leave-one-out cross-validation is usually not possible. An approximation of $\mathrm{RCV}^*(\lambda)$ is needed and we propose a very effective scheme based on the concept of pseudo data. The (unobservable) pseudo data, $\boldsymbol{z}$ are defined as

$$\boldsymbol{z} = \psi(\boldsymbol{y} - \boldsymbol{f})/E\psi^{'} + \boldsymbol{f}, \tag{7}$$

where $\psi = \rho^{'}$. Note that the pseudo data can only be constructed with knowledge of the true function. However, based on this construction, Cox (1983) gave an interesting result: a robust smoothing spline fit is asymptotically equivalent to a least squares smoothing spline fit based on pseudo data. By using this fact, we suggest the following approximation

$$\mathrm{RCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \rho\left\{y_i - \tilde{f}_{\lambda,-i}(t_i)\right\}, \tag{8}$$

where $\tilde{f}_{\lambda,-i}(t_i)$ is the least squares smoothing spline with empirical pseudo data when the ith data point, (t_i, y_i) is omitted. The empirical pseudo data are defined as

$$\hat{\boldsymbol{z}} = \psi(\boldsymbol{y} - \hat{\boldsymbol{f}})/E\psi^{'} + \hat{\boldsymbol{f}}, \tag{9}$$

where $\hat{\boldsymbol{f}}$ is the robust spline applied to the full data. With our notation, the approximation of robust CV score for the variable star problem can be expressed as

$$\mathrm{RCV}(p, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \rho \left\{ y_i - \tilde{f}_{\lambda,-i}(t_i/p) \right\}, \tag{10}$$

where λ is the smoothing parameter for a period p. Define $\mathrm{RCV}(p)$ as the minimum of $\mathrm{RCV}(p, \lambda)$ over $\lambda \in [0, \infty)$ for fixed p

$$\mathrm{RCV}(p) = \min_{\lambda} \mathrm{RCV}(p, \lambda). \tag{11}$$

From the fact that smoothing spline regression can be severely affected by outliers, $\mathrm{RCV}(p)$ might be much less sensitive than $\mathrm{GCV}(p)$ of (4) with a least squares smoothing spline when data are perturbed by outliers. $\mathrm{RCV}(p)$ score (not shown) for the data in Figure 1-(a) has a global minimum at 0.8764. Unlike ordinary GCV, the minimum is unique and smooth.

2.2. *Light curve estimation*

Once the period is determined by either the GCV or the RCV method, a nonparametric curve fitting method can be used to estimate the light curve in the phase domain. Figure 2 shows the estimation of the light curve of the star in Figure 1 after the period is determined at 0.8764 by the RCV method. The top panel shows the estimates of the light curve using SuperSmoother and cosine method, the middle panel illustrates the fits using a smoothing spline and robust smoothing spline regression, and the bottom panel shows the fits based on two other robust smoothing methods described in Section 3 for comparison. All fitting methods have been used with optimal values (smoothing parameter and order) for appropriate criteria. As expected, the fit from a robust smoothing spline (the dashed line in the middle panel) provides a robust estimate relative to the two outliers. The main differences between estimated light curve using SuperSmoother and estimated light curves obtained by other smoothing methods are (1) the shape of the light curve such as the flatness between two eclipses and (2) the difference of amplitude between the primary minimum and the secondary minimum. The light curve (the solid line in the top panel) by SuperSmoother has almost the same amplitudes between two minima and is rounded between eclipses, while the light curves fitted by other smoothing methods have a different amplitude between two minima and are flat between eclipses.

The goal of estimating a light curve is not only to fit the light curve but also to obtain useful information to classify variable stars. If we classify a star as an eclipsing binary based on its light curve, further classification into a contact binary (W Ursa Majoris) type (the light curve by SuperSmoother) or a detached (Algol) type will depend on the relative amplitudes of the minima. Because the SuperSmoother typically underfits the true function, it is not well suited to detect all the features of the light curve shape that are necessary to classify the stars. Instead, as seen in Figure 2, the smoothing spline regression captures the local structures of the true function well. Especially when the data have outliers, robust

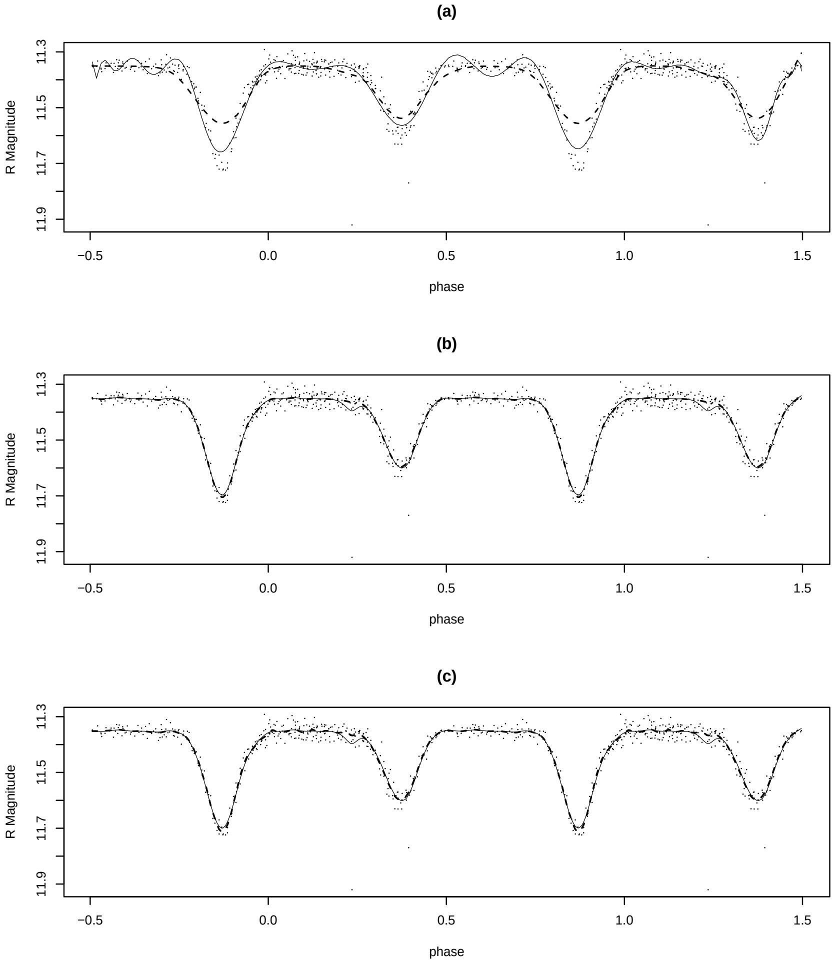

Fig. 2. The estimates of the light curve by several methods. (a) The fits by SuperSmoother (the solid line) and cosine method (the dashed line); (b) the solid line is smoothing spline fit and the dashed line robust smoothing spline fit; (c) the solid line is robust smoothing spline (`rss`) fit and the dashed line is robust `loess` fit.

smoothing spline regression appears to be superior for this application. Note that the cosine method (the dashed line in the top panel) can be used for estimating the light curve, but this method does not work well when the true light curve is not sinusoidal. These subjective observations are confirmed by the simulation study in Section 3.

As a topic related to estimating light curves, we suggest an approximate confidence interval for $f(t_i/p)$ with robust smoothing splines when the period, p is fixed. In order to accomplish this we first detail an explicit equivalence between robust splines and a least squares spline based on empirical pseudo data. The robust smoothing spline fit described in Section 2.1 can be obtained by coupling a least squares smoothing spline with empirical pseudo data in (9). With empirical pseudo data $\hat{\mathbf{z}}$, consider the least squares smoothing spline problem for a fixed period p which minimizes

$$\sum_{i=1}^{n}\{\hat{z}_i - f(t_i/p)\}^2 + \lambda \boldsymbol{f}^T R \boldsymbol{f}, \tag{12}$$

where R is a specific covariance matrix. The solution of (12) solves the normal equation $-2(\hat{\boldsymbol{z}} - \boldsymbol{f}) + 2\lambda R\boldsymbol{f} = \mathbf{0}$ and is equivalent to the normal equation of a robust smoothing spline $-\psi(\boldsymbol{y} - \boldsymbol{f}) + 2\lambda R\boldsymbol{f} = \mathbf{0}$ when $\boldsymbol{f} \equiv \hat{\boldsymbol{f}}$ and $E[\psi'(\varepsilon)] = 2$. Hence, the fit $\hat{\boldsymbol{f}}$ is a robust smoothing. Therefore, applying empirical pseudo data $\hat{\boldsymbol{z}}$ to a least squares smoothing spline produces a robust smoothing. To construct a confidence interval, we apply the pseudo data concept to the confidence intervals proposed by Wahba (1983). The connection between a smoothing spline and a posterior mean suggests a $100(1-\alpha)\%$ confidence interval for $f(t_i/p)$ with a fixed period p as follows

$$\hat{f}(t_i/p) \pm Z_{\alpha/2}\sqrt{\hat{\sigma}^2[A(\hat{\lambda})]_{ii}}, \tag{13}$$

where

$$\hat{\sigma}^2 = \frac{\|[I - A(\hat{\lambda})]\hat{\boldsymbol{z}}\|^2}{\mathrm{tr}[A(\hat{\lambda})]}.$$

Figure 5 shows a 95% confidence interval of the light curve of star 306 constructed by (13).

One reviewer raised some justifiable questions whether this confidence procedure can be trusted in view of the underlying distributions being non-Gaussian. First we note that the confidence interval procedure is being applied to the estimator derived from the empirical pseudo data not the data in the original outlier scale. The empirical and theoretical pseudo data will always have finite moments due to the boundedness of the transformation. Thus, formulas based on first and second moments are not unreasonable. It is an open area of research for us to prove the validity of these intervals, however, we can provide some evidence based on technical results and collateral theory as to why one might trust this procedure. But we should emphasize that the following discussion is far from a rigorous outline or even a sketch of a proof. The main point in our argument is to assume that the estimator based on empirical pseudo data and RCV (8) approximates (i.e. is asymptotically equivalent) to the estimator using theoretical pseudo data (7) and the optimal smoothing parameter. We note that Cox's results give some evidence for an asymptotic equivalence between (7) and (9) and the work of Hall and Jones (1990) suggests that cross-validation can provide a consistent estimate of the optimal bandwidth in the context of robust smoothing. Given the theoretical pseudo-data estimator evaluated at the optimal smoothing parameter one would expect the confidence intervals for Wahba to be reliable. Here we appeal to Wahba's

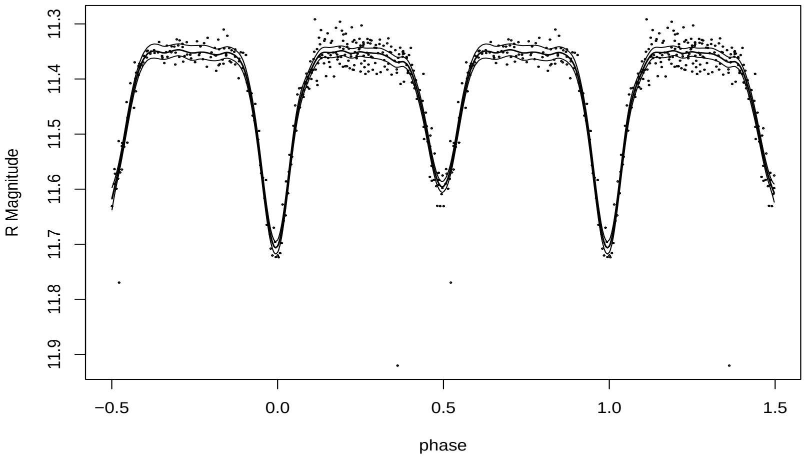

Fig. 3. 95% pointwise confidence interval.

work in this area and the fact that the spline is a special case of a geostatistics or Kriging estimator. Indeed, Wahba's intervals are based on the prediction standard errors under the assumption of a particular generalized covariance. Such Kriging standard errors do not depend on normality only finite moments. In summary, while we do not have a rigorous justification of these intervals we feel that there is enough suggestions in the available theory to make them useful measures of uncertainty for the estimated function. As in any procedure that depends on underlying assumptions, care should be taken when drawing inferences. But we feel that these companion confidence intervals are much better than simply reporting an estimate without any quantification of its uncertainty.

2.3. Theoretical motivation for RCV(λ)

When robust smoothing spline regression is performed for estimating a light curve, the smoothing parameter λ has to be selected automatically. As a selection method for λ, we believe that RCV(λ) may be useful. Note that RCV$^*(\lambda)$ mentioned here is the idealized, leave-one-out version defined as, for a fixed p,

$$\mathrm{RCV}^*(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \rho\left\{y_i - \hat{f}_{\lambda,-i}(t_i/p)\right\}.$$

This is different from our approximation defined in (8) and (10).

Given a period p, we conjecture that the minimizer of RCV$^*(\lambda)$ also minimizes in asymptotic mean squared error between the estimate of the robust smoothing spline regression and the truth. Denote the mean squared error as MSE $(\hat{f}, f) = \frac{1}{n}\sum_{i=1}^{n} E(f_i - \hat{f}_i)^2$. A robust extension to the result of Craven and Wahba (1979) gives the following: Given a fixed p, if λ^* is the minimizer of $\mathrm{E}[\mathrm{RCV}^*(\lambda)]$ over $[\lambda_n, \Lambda_n]$, then

$$\frac{\mathrm{MSE}\left(\hat{f}_{\lambda^*}, f\right)}{\min_\lambda \mathrm{MSE}\left(\hat{f}_\lambda, f\right)} \to 1, \qquad \text{as} \quad n \to \infty. \tag{14}$$

We now include a sketch of the proof for the property (14). More theoretical results of RCV^* and rigorous proofs of (14) are in progress and will appear elsewhere. Let $\tilde{f}_\lambda$ be a least-square smoothing spline fit with pseudo data in (7). By a Taylor expansion of RCV^*, $\mathrm{E}[\mathrm{RCV}^*(\lambda)]$ is asymptotically equivalent to $\mathrm{E}[\mathrm{CV}(\lambda)]$ based on pseudo data

$$\mathrm{E}[\mathrm{RCV}^*(\lambda)] \approx \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left(f_i - \tilde{f}_{\lambda,-i}\right)^2 + \text{constant}.$$

Thus, by using the result of Craven and Wahba (1979), it can be shown that $\mathrm{E}\left[\mathrm{RCV}^*(\lambda)\right] \approx \mathrm{MSE}(\tilde{f}_\lambda, f) + \text{constant}$. Finally, with Cox's result(1983): $\hat{f}_\lambda$ inherits the same asymptotics as $\tilde{f}_\lambda$, and we conclude that

$$\mathrm{E}\left[\mathrm{RCV}^*(\lambda)\right] \approx \mathrm{MSE}\left(\hat{f}_\lambda, f\right) + \text{constant}.$$

In comparison to these results, Hall and Jones (1990) discussed kernel M-estimates of the regression function using Huber's ρ-function. They showed that least squares cross-validation results in optimal bandwidth selection (and determining C) with respect to mean squared error. However, we have found it is difficult to extend their results to spline-type estimates based on pseudo data.

3. A comparison of methods

Here we report results of our analysis of several variable stars and one numerical experiment. These experiments are designed for comparing the practical performances of the proposed approaches with some existing methods. To assess the performance of the proposed method RCV when the data are perturbed by outliers, we use two robust smoothing approaches. One is a robust loess method based on the assumption of symmetric errors instead of Gaussian errors. Therefore, the robust loess estimate is not adversely affected if the errors have a long-tailed distribution (Chambers and Hastie, 1993). The other is a robust fit of a smoothing spline using the L_1 norm. The algorithm is an iterative reweighted smooth spline algorithm which performs a least squares smoothing spline at each step with the weights w equal to the inverse of the absolute value of the residuals for the last iteration step. Note that this robust smoothing splines is different from the robust smoothing spline fit based on the empirical pseudo data in Section 2.1. Both robust smoothing methods are applied to find the period that minimizes the sum of square residuals obtained from fitting.

For two experiments, the following eight methods are compared:

1. `rcv:` the robust cross-validation proposed in Section 2.1 as the target,
2. `gcv:` the generalized cross-validation described in Section 2.1,
3. `lk:` the measure of dispersion procedure of Lafler and Kinman (1965),
4. `pdm:` the phase dispersion minimization procedure,
5. `Fourier:` the cosine method with least-squares,

Table 1. The estimated period (in days) and its (bootstrap) standard deviation σ_B, given in parentheses, for several variable stars according to different methods

	Star						
Method	306	969	1164	1744	4699	4865	5954
`rcv`	0.8764	2.2124	1.4630	1.3328	3.2316	2.3250	0.9808
	(0.0001)	(0.0030)	(0.0004)	(0.0015)	(0.0295)	(0.0016)	(0.0003)
`gcv`	0.8762	2.2134	1.4634	1.3328	3.2858	2.3256	0.9806
	(0.0003)	(0.0028)	(0.0010)	(0.0020)	(0.0168)	(0.0024)	(0.0004)
`lk`	0.8764	2.2148	1.4632	1.3320	3.2220	2.3240	0.9804
	(0.0001)	(0.0022)	(0.0006)	(0.0013)	(0.0332)	(0.0014)	(0.0003)
`pdm`	0.8764	2.2182	1.4592	1.3316	3.2966	2.3196	0.9798
	(0.0001)	(0.0020)	(0.0003)	(0.0019)	(0.0322)	(0.0018)	(0.0003)
`Fourier`	0.8762	2.2144	1.4626	1.3322	3.2894	2.3210	0.9804
	(0.0002)	(0.0034)	(0.0007)	(0.0011)	(0.0190)	(0.0021)	(0.0003)
`sm`	0.8762	2.2116	1.4626	1.3316	3.2896	2.3278	0.9802
	(0.0003)	(0.0023)	(0.0014)	(0.0009)	(0.0322)	(0.0030)	(0.0003)
`rloess`	0.8760	2.2134	1.4632	1.3328	3.2888	2.3278	0.9804
	(0.0004)	(0.0021)	(0.0006)	(0.0009)	(0.019)	(0.0028)	(0.0003)
`rss`	0.8766	2.2142	1.4632	1.3334	3.2878	2.3262	0.9808
	(0.0003)	(0.0018)	(0.0003)	(0.0043)	(0.0313)	(0.0011)	(0.0003)

6. `sm:` the smoothing procedure by SuperSmoother,
7. `rloess:` the robust smoothing procedure by robust loess, and
8. `rss:` the robust smoothing procedure by robust smoothing spline.

Note that all smoothing methods have performed with some forms of smoothing parameter selection and the order for `Fourier` has been selected by Akaike's information criterion (AIC).

3.1. Analysis of data on variable stars

In this section, for several real variable stars, we compare the eight methods cited the above. For the real eclipsing binary star (306) data shown in Figure 1, based on the global minima from an exhaustive search, the period is estimated as 0.8762 and 0.8764 under the GCV and the RCV method respectively. Existing methods result in almost the same period as the RCV method (Table 1). However, as shown in Table 1, we find that the estimated periods vary with different procedures for some variable stars. Figure 4 displays light curves produced by folding data over the estimated period determined through `rcv`.

To estimate the error of the period estimate, we simply use a bootstrap method (Efron, 1979). The procedure is performed as follows: B training sets $\boldsymbol{y}^{*b}$, $b = 1, 2, \ldots, B(= 100)$ are drawn with replacement from the original dataset $\boldsymbol{y}$. Each sample has the same size as the original dataset. The period p^{*b} is estimated from each bootstrap training set, and then we compute its standard deviation to assess the accuracy of the period estimate

$$\sigma_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(p^{*b} - \bar{p^*})^2},$$

where $\bar{p^*} = \sum_{b=1}^{B} p^{*b}/B$. The errors of the period estimate, σ_B for several real variable

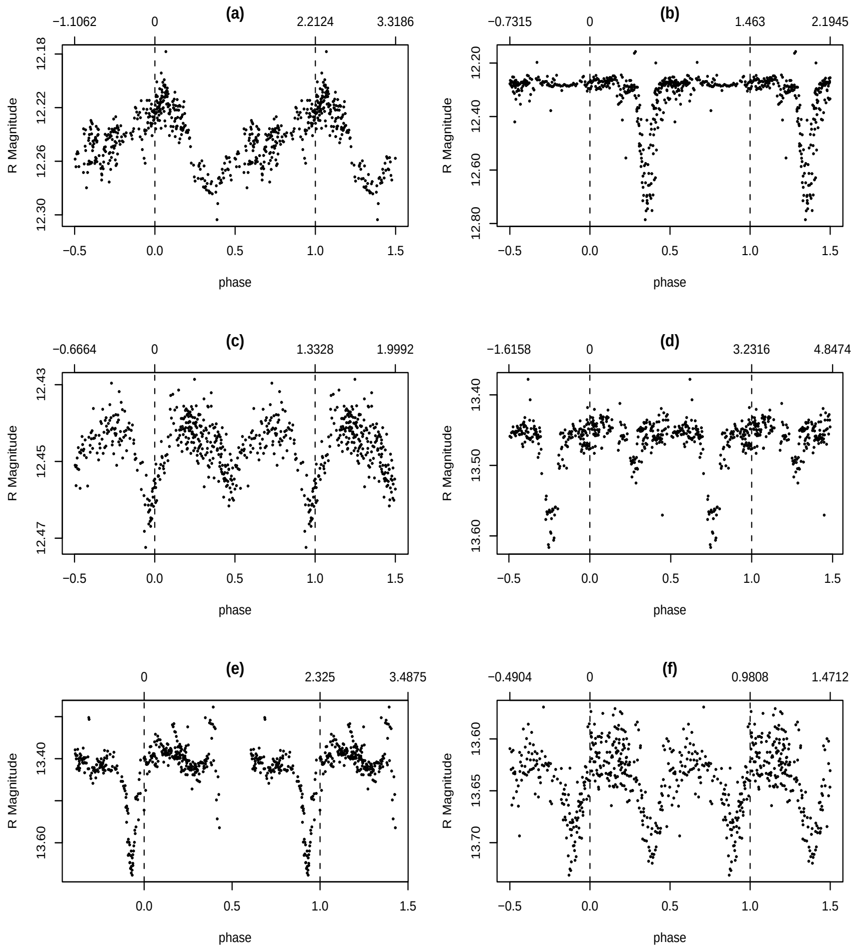

Fig. 4. Plots of brightness versus phase with period estimate by RCV. (a) star 969 (b) star 1164 (c) star 1744 (d) star 4699 (e) star 4865 (f) star 5954.

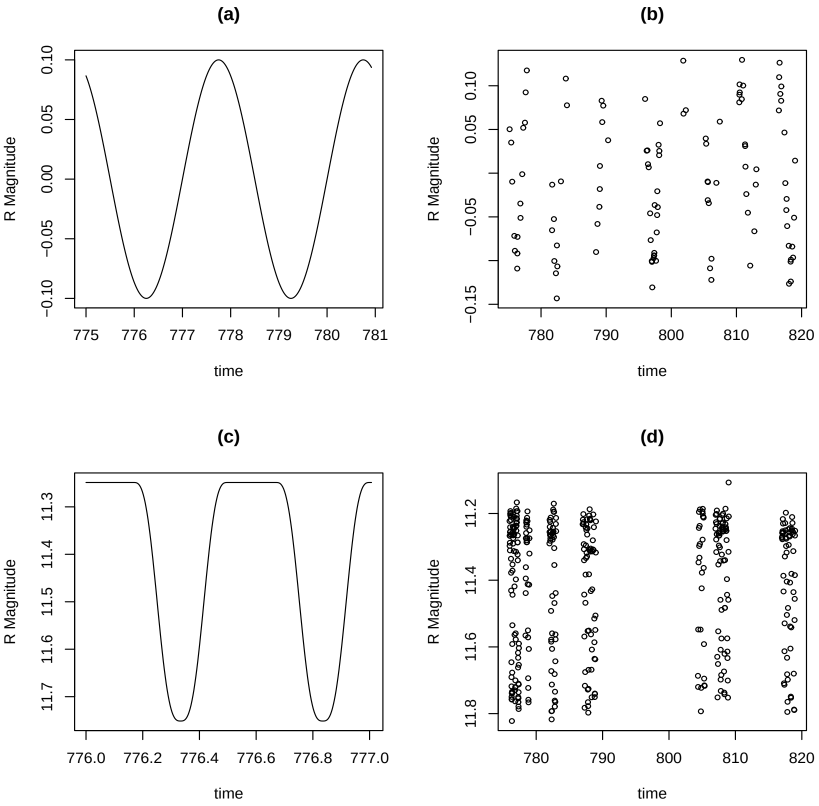

Fig. 5. Simulated sinusoidal signals with noise. (a) True light curve with the period 3 days. (b) Irregularly sampled brightness. (c) True light curve with the period 1 day. (d) Irregularly sampled brightness.

stars are given in parentheses in Table 1.

3.2. *Simulation study*

In this experiment the proposed RCV method is compared with existing methods through a simulation study based on two test functions and three noise levels. In summary, this is a 3 factor experiment with 2 test functions, 3 levels of noise and 8 methods of estimating the period. We use two test functions plotted in Figure 5. The first function is a sinusoidal signal, $f_i = 0.1 \sin(2\pi t_i/3)$ that is a curve resembling the shape for an eclipsing binary (EW). The period, 3 days, is typical for stars in the data set and we also match the irregular time sampling based on brightness data from observed data. For the noise models, we consider (1) Gaussian errors, $\mathrm{N}(0, 0.04)$, (2) Student t noise with three degrees of freedom scaled by 0.05, and (3) a mixture of 95% N(0, 0.04) and 5% N(0, 0.3) at random location. Note that noise model (1) represents the case that outliers are not present, and models (2)-(3) are considered for the case outliers are present. The noise level, 0.04, is consistent with the estimated noise level for several real EW stars.

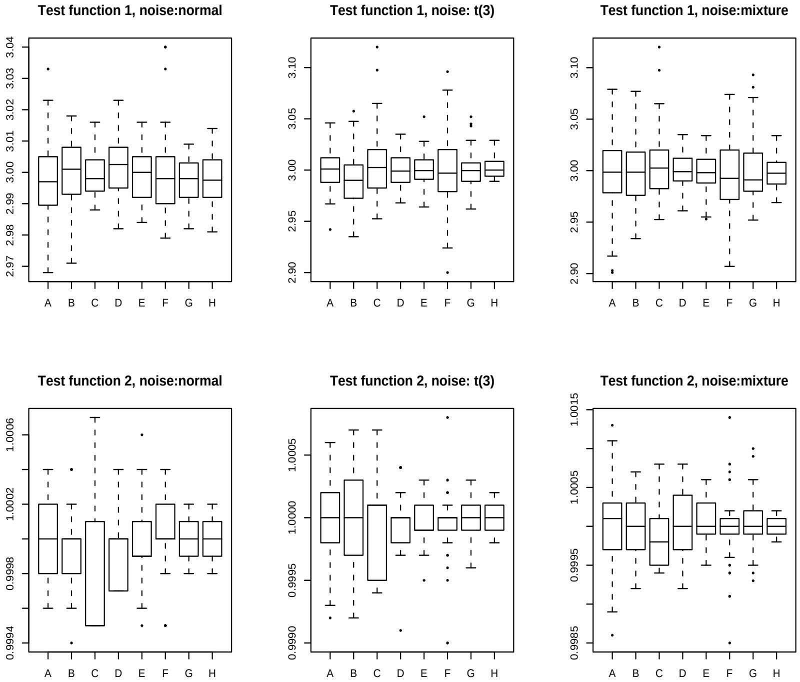

Fig. 6. Boxplots of period estimates with respect to two test functions and three noise scenarios. In all cases, A denotes the method suggested by Lafler and Kinman; B, PDM method; C, Fourier; D, SuperSmoother; E, Robust Loess; F, Robust Smooth Spline; G, GCV; H, RCV.

For the second test function, the signal is a periodic function similar to the actual Algol type (EA) star as shown in Figure 1-(b). The test function has a geometric interpretation. Assume two spheres with the same radius, r that are orbiting each other. We calculate overlapping and non-overlapping areas of the two spheres and equate non-overlapping areas to brightness. The test function obtained in this manner is:

$$f(t) = \begin{cases} \alpha[\pi r] + \beta, & a \le t < b \\ \alpha[\pi r - r^2(\theta - \sin\theta)]_{0 \le t < 2\pi} + \beta, & b \le t < c \\ \alpha[\pi r - r^2(\theta - \sin\theta)]_{2\pi \le t < 0} + \beta, & c \le t < d \\ \alpha[\pi r] + \beta, & d \le t < e \\ \alpha[\pi r - r^2(\theta - \sin\theta)]_{0 \le t < 2\pi} + \beta, & e \le t < f \\ \alpha[\pi r - r^2(\theta - \sin\theta)]_{2\pi \le t < 0} + \beta, & f \le t < g \end{cases}$$

$\alpha = 0.07$, $\beta = -11.5$ and radius $r = 1$ results in the "true" light curve in Figure 5-(c). We select a period $g - a = 1$ day. Irregular sampling provides the brightness shown in Figure 5-(d). Finally, three noise scenarios are considered: (1) Gaussian error $\mathrm{N}(0, 0.06)$, (2) Student t noise with three degrees of freedom scaled by 0.08, and (3) a mixture of 95% N(0, 0.06) and 5% N(0, 0.8) at random location. The noise level, 0.06, is a reasonable choice for an EA star.

The boxplots in Figure 6 summarize the results of period estimates based on 100 replications. From the simulation results, we have the following empirical observations: (1) `rcv` and `gcv` give nearly identical results for the Gaussian; (2) for the first test function (sinusoidal) and Gaussian noise, `Fourier` provides the best result with `gcv` and `rcv`; (3) both robust robust procedures, `rcv` and `rloess` outperform non-robust procedures when the error is $t(3)$ and a mixture of two Gaussian at random locations; (4) `rcv` always outperforms two other robust methods, `rloess` and `rss` for $t(3)$ and mixture noise scenario.

4. Multiple periodicity

Here we discuss the estimation of multiple periodicity of a variable star. Consider a time series $\{y_i, t_i\}$ with fixed L periodicity,

$$y_i = \sum_{l=1}^{L} f_l(t_i/p_l) + \varepsilon_i,$$

where f_l is lth period function with period p_l. The statistical problem with the above additive model is to estimate f_l and p_l, $l = 1, 2, \ldots, L$. Since the RCV method in Section 2.1 is induced by a smoothing technique (a robust smoothing spline), we can use a backfitting algorithm which is well known in fitting nonparametric additive regression (Chambers and Hastie, 1993). The backfitting algorithm can be briefly described as follows. For simplicity, let $L = 2$, that is $y_i = [f_1(t_i/p_1) + f_2(t_i/p_2)] + \varepsilon_i$. Consider the system of two equations:

$$\begin{aligned} \boldsymbol{f}_1 &= S_1(\boldsymbol{y} - \cdot - \boldsymbol{f}_2) \qquad \text{and} \\ \boldsymbol{f}_2 &= S_2(\boldsymbol{y} - \boldsymbol{f}_1 - \cdot), \end{aligned} \tag{15}$$

where the dots in the equation are placeholders showing the term that is missing in each row. Here a vector $\boldsymbol{f}_l$ denotes the function f_l evaluated at the sampling time t_i and the

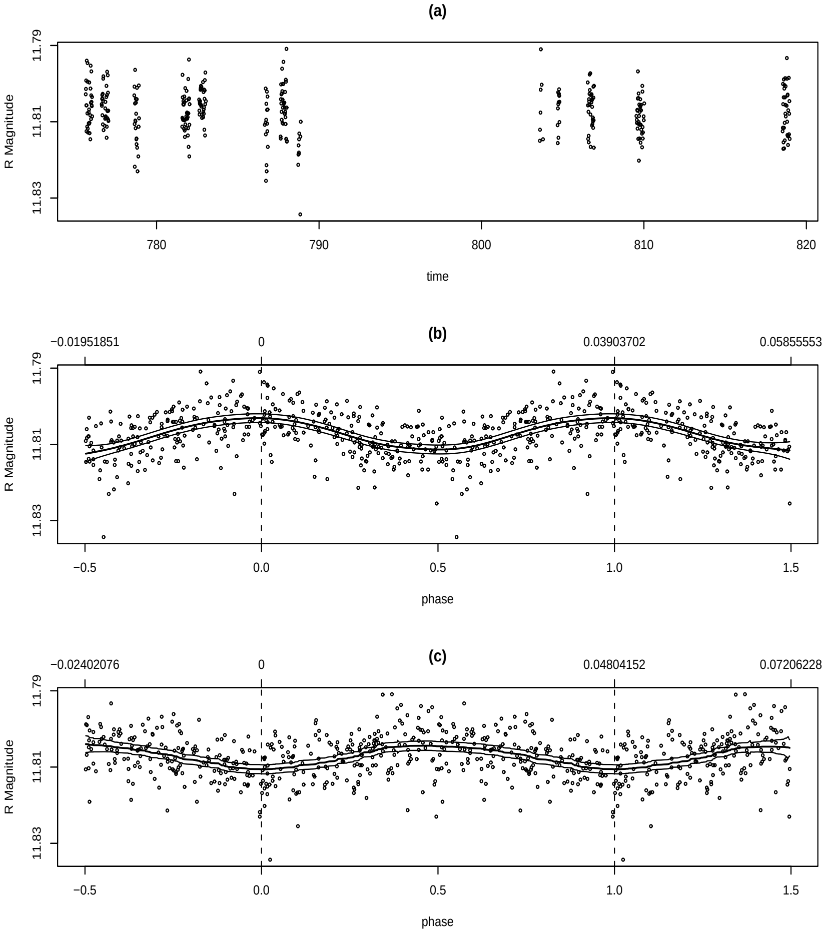

Fig. 7. (a) Brightness of a variable star with multiple periodicity. (b) Plot of brightness versus phase with the first period p=0.03903 (day) and the estimate of light curve with 95% pointwise confidence interval. (c) Plots of brightness versus phase with the second period p=0.04804 (day) and the estimate of light curve with 95% pointwise confidence interval.

period p_l. S_l represents the smoother operator matrix for smoothing against t_i at the period p_l. In this case, we use robust smoothing splines in (5) as all the smoothers. Thus, this system solves the following problem

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left\{y_i - [f_1(t_i/p_1) + f_2(t_i/p_2)]\right\} + \sum_{l=1}^{2}\lambda_l \int_{[0,1]} \left\{f_l^{''}(t)\right\}^2 dt. \tag{16}$$

To solve this system, the Gauss-Seidel iterative method loops through the equations substituting the most updated versions of functions in the right hand side with each iteration. For estimating multiple periods, we use the algorithm based on backfitting as follows. Suppose that we have an initial period estimate $\hat{p}_1$. And then

(1) Obtain residuals by $y_i - \hat{f}_1(t_i/\hat{p}_1)$.
(2) Estimate $\hat{p}_2$ by using the RCV method described in Section 2.1.
(3) Take residuals by $y_i - \hat{f}_2(t_i/\hat{p}_2)$.
(4) Estimate $\hat{p}_1$ by using RCV method.
(5) Repeat (1)-(4) until the estimates $\hat{p}_1$ and $\hat{p}_2$ converge.

Finally, we apply the algorithm to a real Delta Scuti type star (543) in the STARE database. The algorithm converges to two periods $\hat{p}_1 = 0.03903$ days and $\hat{p}_2 = 0.04804$ days after several iterations. Figure 7 shows the brightness of star 543 in the time domain and plots of brightness versus phase with period estimates determined by backfitting algorithm.

5. Conclusion

We have proposed two methods to estimate the period of a variable star from the irregularly observed brightness. The GCV method is to minimize GCV score generated by a smoothing spline, while the RCV method is based on robust smoothing spline regression as a robust version to the outliers. Based on actual light curve data and a simulation study, we have shown that the proposed method estimates the period more accurately than existing methods. In case the signal is perturbed by a few outliers, the RCV method followed by the robust smoothing spline regression is a useful tool for estimating the period. In general, the robust smoothing spline proposed in this paper provides a resistant method to outliers so that the advantage of this approach might be important for the next phase of our group's scientific project. Currently, we are applying RCV method to determine periods for a survey of approximately 6000 stars.

As a future direction for statistical research, we also note the pseudo data idea can be fruitfully applied to thin-plate smoothing spline and wavelet shrinkage to obtain robust estimators. It might also be interesting to apply RCV method to different database such as HIPPARCOS and MACHO.

Acknowledgement

We would like to thank the Editor and two reviewers of this paper for their constructive comments. This work was supported in part by the grant DMS-9312686 to the Geophysical Statistics Project at NCAR from U.S. National Science Foundation and the Natural Sciences and Engineering Council of Canada.

References

Brown, T. M. and Gilliland, R. L. (1994). Asteroseismology. *Annual Reviews of Astronomy and Astrophysics*, **32**, 37–82.

Chambers, J. M. and Hastie, T. J (1993) (eds). *Statistical Models in S.* Chapman and Hall, New York.

Charbonneau, D., Brown, T. M., Latham, D. W. and Mayor, M. (2000). Detection of planetary transits across a sun-like star. *The Astrophysical Journal*, **529**, L45–L48.

Cox, D. D. (1983). Asymptotics for M-type smoothing splines. *The Annals of Statistics*, **11**, 530–551.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the Method of generalized cross-validation. *Numerische Mathematik*, **31**, 377–403.

Deeming, T. J. (1975). Fourier analysis with unequally-spaced data. *Astrophysical and Space Science*, **36**, 137–158.

Dwortesky, M. M. (1983). A period-finding method for sparse randomly spaced observations or "How long is a piece of string ?". *Monthly Notices of the Royal Astronomical Society*, **203**, 917–924.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.

Friedmann, J. H. (1984). A variable span smoother. Technical report No. 5, Laboratory for Computational Statistics, Department of Statistics, Stanford University.

Gautschy, A. and Saio, H. (1995). Stellar pulsations across the HR diagram, Part 1. *Annual Reviews of Astronomy and Astrophysics*, **33**, 75–113.

Gautschy, A. and Saio, H. (1996). Stellar pulsations across the HR diagram, Part 2. *Annual Reviews of Astronomy and Astrophysics*, **34**, 551–606.

Hall, P. and Jones, M. C. (1990). Adaptive M-estimation in nonparametric regression. *The Annals of Statistics*, **18**, 1712–1728.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hilditch, R. W. (2001). *An Introduction to Close Binary Stars.* Cambridge University Press, Cambridge.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Lafler, J. and Kinman, T. D. (1965). An RR Lyrae survey with the Lick 20-inch astrograph II. The calculation of RR Lyrae periods by electronic computer. *Astrophysical Journal Supplement Series*, **11**, 216–222.

Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysical and Space Science*, **39**, 447–462.

Reimann, J. D. (1994). Frequency Estimation Using Unequally-Spaced Astronomical Data. Ph.D. dissertation, Department of Statistics, University of California at Berkeley.

Scargle, J. D. (1982). Studies in astronomical time series analysis II. Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, **263**, 835–853.

Stellingwerf, R. F. (1978). Period determination using phase dispersion minimization. *The Astrophysical Journal*, **224**, 953–960.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc.*, **B, 45**, 133–150.

Wahba, G. (1990). *Spline Models for Observational Data.* Society for Indurstial and Applied Mathematics, Pennsylvania.