A FAST CHI-SQUARED TECHNIQUE FOR PERIOD SEARCH OF IRREGULARLY SAMPLED DATA

DAVID M. PALMER
Los Alamos National Laboratory, B244, Los Alamos, NM 87545, USA; palmer@lanl.gov

ABSTRACT

A new, computationally and statistically efficient algorithm, the Fast χ^2 algorithm ($F\chi^2$), can find a periodic signal with harmonic content in irregularly sampled data with nonuniform errors. The algorithm calculates the minimized χ^2 as a function of frequency at the desired number of harmonics, using fast Fourier transforms to provide $O(N \log N)$ performance. The code for a reference implementation is provided.

Key words: methods: data analysis – methods: numerical – methods: statistical – stars: oscillations

1. INTRODUCTION

A common problem, in astronomy and in other fields, is to detect the presence and characteristics of an unknown periodic signal in irregularly spaced data. Throughout this paper I will use examples and language drawn from the study of periodic variable stars, but the techniques can be applied to many other situations.

Generally, detecting periodicity is achieved by comparing the measurements phased at each of a set of trial frequencies to a model periodic phase function, $\Phi(\phi)$, and selecting the frequency that yields a high value for a quality function. 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 frequencies searched, and the methods used for computational efficiency.

The Lomb algorithm (Lomb 1976), for example, uses a sinusoid plus constant for its model function. The quality is the amplitude of the unweighted least-squares fit at the trial frequency. A simple implementation takes $O(nN)$ computations, where n is the number of measurements and N is the number of frequencies searched. Press & Rybicki (1989) present an implementation that uses the fast Fourier transform (FFT) to give $O(N \log N)$ performance. Spurious detections can be produced at frequencies where the sample times have significant periodicity, power from harmonics above the fundamental sine wave is lost, and the algorithm is statistically inefficient in the sense that it ignores point-to-point variation in the measurement error.

Irwin et al. (2006) use a Lomb algorithm to find a candidate period, then test the measurements' χ^2 reduction between a constant value and a model periodic light curve obtained by folding the measurements at that period and applying smoothing.

Phase dispersion minimization (Stellingwerf 1978) divides a cycle into (possibly overlapping) phase bins. The quality is calculated from the χ^2 agreement among those data points that fall into each bin at the trial frequency. This also has $O(nN)$ performance. The size and number of the bins, and the time of $\phi = 0$, are choices that can affect the detection probability of a particular signal.

The Fast χ^2 (hereafter $F\chi^2$) technique presented here uses a Fourier series truncated at harmonic H:

$$\Phi_H(\{A_{0...2H}, f\}, t)$$
$$= A_0 + \sum_{h=1...H} A_{2h-1} \sin(h2\pi f t) + A_{2h} \cos(h2\pi f t) \quad (1)$$

for the model periodic function. The fit quality is the χ^2 of all data, jointly minimized over the Fourier coefficients, $A_{0...2H}$, and the frequency, f. Baluev (2009) investigates the statistics of these fits. The set of frequencies searched can have arbitrary density and range. The computational complexity of this implementation is $O(HN \log HN)$. The optimal choice of H depends on the (generally unknown) true shape of the periodic signal, but evaluations with multiple H values can share computational stages, resulting in determinations at $H' = 1, \ldots, H$ at only a slight premium. The fit follows standard χ^2 statistics and makes efficient use of measurement errors.

2. DESCRIPTION

χ^2 minimization is the standard method to fit a data set with unequal, Gaussian, measurement errors. (Other maximum likelihood methods can handle more general error distributions, but are beyond the scope of this paper.)

A hypothesis may be expressed as a model $M(\mathbf{P}, t)$, where $\mathbf{P}$ is the set of unknown model parameters and t is an independent variable. This is compared with measurements x_i, with standard errors σ_i, made at t_i. The comparison is made by finding the $\mathbf{P}$ that minimizes

$$\chi^2(\mathbf{P}) = \sum_{i=1...n} \frac{(M(\mathbf{P}, t_i) - x_i)^2}{\sigma_i^2}. \quad (2)$$

In this paper t is a scalar value which, for the variable star case, represents time. However, the extension of this technique to spatial or other dimensions, to higher dimension such as the (u, v) baselines of interferometry, and to nonscalar measurements, $\mathbf{x_i}$, is straightforward.

If this minimum χ^2 is significantly below that for a null hypothesis, then this is evidence for the model, and for the value of $\mathbf{P}$ at the minimum. The extent by which the model decreases the χ^2 compared with the null hypotheses, $\Delta\chi^2 = \chi_0^2 - \chi^2(\mathbf{P})$, is known as the minimization index and, if the null hypothesis is true and other conditions apply, has a χ^2 probability distribution with the same number of degrees of freedom as the model.

Many models of interest are linear in some parameters, and nonlinear in others. These models can be factored into the linear parameters, $\mathbf{A}$, and a set of functions $\mathbf{M}(\mathbf{Z}, t)$ of the nonlinear parameters, $\mathbf{Z}$, so that

$$M(\{\mathbf{A}, \mathbf{Z}\}, t) = \sum_i \mathbf{A}_i \mathbf{M}_i(\mathbf{Z}, t). \quad (3)$$

For any given $\mathbf{Z}$, the χ^2 minimized with respect to $\mathbf{A}$ can be rapidly found in closed form by linear regression. Finding the

global minimum of χ^2 thus reduces to searching the nonlinear parameter space $\mathbf{Z}$.

The truncated Fourier series function in Equation (1) is linear in the coefficients $A_{0...2H}$, and nonlinear in $\mathbf{Z} = f$, with component functions

$$\mathbf{M}_0(f, t) = 1, \tag{4}$$

$$\mathbf{M}_{2h-1}(f, t) = \sin h 2\pi f t, \tag{5}$$

$$\mathbf{M}_{2h}(f, t) = \cos h 2\pi f t. \tag{6}$$

Therefore, the minimum χ^2 can be quickly calculated at any chosen frequency. Complete, dense coverage of the frequency range of interest is required to find the overall minimum—the orthogonality of the Fourier basis tends to make the minima very narrow.

Proceeding in the usual manner for linear regression (see e.g., Press et al. 2002, Section 15.4 for a review), we produce the α matrix and β vector used in the "normal equations." The values of α and β depend on the frequency f.

The α matrix comes from the cross terms of the model components, as a weighted sum over the data points:

$${}^f\alpha_{jk} = \sum_{i=1...n} \frac{\mathbf{M}_j(f, t_i)\mathbf{M}_k(f, t_i)}{\sigma_i^2}. \tag{7}$$

The β vector is the weighted sum of the product of the data and the model components:

$${}^f\beta_j = \sum_{i=1...n} \frac{\mathbf{M}_j(f, t_i)x_i}{\sigma_i^2}. \tag{8}$$

A frequency-independent scalar, X, which is the χ^2 corresponding to a *constant* $= 0$ model:

$$X = \sum_i x_i^2/\sigma_i^2, \tag{9}$$

is also used in the calculation of χ^2.

The minimization of Equation (2) over the linear $\mathbf{A}$ parameters can be shown to give a minimum χ^2 subject to f of

$${}^f\chi^2_{\min} = X - \sum_j \sum_k {}^f\beta_j ({}^f\alpha^{-1})_{jk} {}^f\beta_k. \tag{10}$$

The Fourier coefficients corresponding to this minimum are

$${}^f\mathbf{A}_j = \sum_k ({}^f\alpha^{-1})_{jk} {}^f\beta_k. \tag{11}$$

We can use the trigonometric product relations

$$\sin(a)\sin(b) = \frac{1}{2}(\cos(a-b) - \cos(a+b)), \tag{12}$$

$$\sin(a)\cos(b) = \frac{1}{2}(\sin(a+b) + \sin(a-b)), \tag{13}$$

$$\cos(a)\sin(b) = \frac{1}{2}(\sin(a+b) - \sin(a-b)), \tag{14}$$

$$\cos(a)\cos(b) = \frac{1}{2}(\cos(a-b) + \cos(a+b)) \tag{15}$$

to reduce the cross terms of sines and cosines in α to

$${}^f\alpha_{2h-1,2h'-1} = \frac{1}{2}(C'((h-h')f) - C'((h+h')f)), \tag{16}$$

$${}^f\alpha_{2h-1,2h'} = \frac{1}{2}(S'((h+h')f) + S'((h-h')f)), \tag{17}$$

$${}^f\alpha_{2h,2h'-1} = \frac{1}{2}(S'((h+h')f) - S'((h-h')f)), \tag{18}$$

$${}^f\alpha_{2h,2h'} = \frac{1}{2}(C'((h-h')f) + C'((h+h')f)), \tag{19}$$

while the terms for β are

$${}^f\beta_{2h-1} = S(hf), \tag{20}$$

$${}^f\beta_{2h} = C(hf), \tag{21}$$

where

$$\begin{aligned}
C(f) &= \sum \cos(2\pi f t_i)x_i/\sigma_i^2 &= \operatorname{re} P(f) \\
S(f) &= \sum \sin(2\pi f t_i)x_i/\sigma_i^2 &= \operatorname{im} P(f) \\
C'(f) &= \sum \cos(2\pi f t_i)/\sigma_i^2 &= \operatorname{re} Q(f) \\
S'(f) &= \sum \sin(2\pi f t_i)/\sigma_i^2 &= \operatorname{im} Q(f)
\end{aligned}. \tag{22}$$

The functions C, C', S, and S' are the cosine and sine parts of the Fourier transforms of the weighted data and the weights:

$$\begin{aligned}
p(t) &= \sum_i \delta(t - t_i)\frac{x_i}{\sigma_i^2} & P(f) &= \int p(t)e^{i2\pi f t}dt \\
q(t) &= \sum_i \delta(t - t_i)\frac{1}{\sigma_i^2} & Q(f) &= \int q(t)e^{i2\pi f t}dt
\end{aligned}. \tag{23}$$

The function $p(t)$ is the weighted data, and $q(t)$ is the weight, as a function of time. These are both zero at times when there is no data, and Dirac delta functions at the time of each measurement. P and Q can be efficiently computed by the application of the FFT to p and q.

This FFT calculation of P and Q, their use in construction of α and β for a set of frequencies, and the χ^2 minimization by Equation (10) for each of those frequencies, are the components of the Fχ^2 algorithm.

3. IMPLEMENTATION

The steps in implementing an Fχ^2 search are: (1) choosing the search space, (2) generating the Fourier transforms, (3) calculating the normal equations at each frequency, (4) finding the minimum, and (5) interpreting the result.

(1) *Choosing the search space.* The frequency range of interest, the number of harmonics, and the density of coverage may be chosen based on physical expectations, measurement characteristics, processing limitations, or other considerations.

The maximum frequency searched may be where the exposure times of the observations are a large fraction of the period of the highest harmonic, or some other upper bound placed by experience or the physics of the source. At low frequencies, if there are only a few cycles or less of the fundamental over the span of all observations, a large χ^2 decrease may be evidence of variability but not necessarily of periodicity.

It is not necessary that the fitting function accurately represent all aspects of the physical process. Sharp features in the actual phase function may require high harmonics to reproduce as a Fourier sum, but can be detected with adequate sensitivity using

a small number of harmonics. Details of the phase function (e.g., the behavior on ingress and egress of an eclipsing binary) may be determined by other techniques once a candidate frequency has been found.

The density of the search—how closely the trial frequencies are spaced—affects the sensitivity as well. For a simple Fourier analysis of the fundamental, a spacing of one cycle over the span of observations can cause a reduction in amplitude to $1/\sqrt{(2)}$ if the true frequency is intermediate between two trial frequencies. For an analysis to harmonic H, the spacing of trial fundamental frequencies must be correspondingly tighter. The maximum sensitivity loss depends on the harmonic content of the signal, which is generally not a priori known, but a spacing much looser than $1/H$ cycles will typically lose the advantage of going to harmonic H, while a much tighter spacing will consume resources that might be better employed with a search to $H + 1$, depending on the characteristics of the source.

If ΔT is the time span of all observations and δt is the fastest timescale of interest, then a reasonable maximum frequency and spacing for the fundamental would be $f_{\max} \gtrsim 1/(2H\delta t)$ and $\delta f \lesssim 1/(H\Delta T)$.

(2) *Generating the Fourier transforms.* The calculation of $^f\chi^2_{\min}$ requires evaluation of the Fourier functions P, at $\{0, f, \ldots, Hf\}$, and Q, at $\{0, f, \ldots, 2Hf\}$.

The real-to-complex FFT, as typically implemented, takes as input $2\mathcal{N}$ real data points from uniformly spaced discrete times $t = \{0, \delta t, 2\delta t, \ldots (2\mathcal{N} - 1)\delta t\}$. If the data were not sampled at those exact times, they are "gridded" (e.g., by interpolation, or by nearest neighbor sampling) to estimate what the data values at those times would have been. The output of the FFT is the $\mathcal{N}+1$ complex Fourier components corresponding to frequencies $f = \{0, \delta f, 2\delta f, \ldots, \mathcal{N}\delta f\}$, where $\delta f = 1/(2\mathcal{N}\delta t)$.

To provide the frequency range and density required for the Fχ^2 method, the weighted-data and the weights are placed in sparse, zero-padded arrays, with δt bin width, that cover at least $2H$ times the observed time period. This produces discrete functions (using $\widehat{\text{hat}}$ symbols to indicate discrete quantities)

$$\hat{p}[\hat{t}] = \sum_i \hat{\delta}[\hat{t}, \hat{t}_i]\frac{x_i - \bar{x}}{\sigma_i^2}, \tag{24}$$

$$\hat{q}[\hat{t}] = \sum_i \hat{\delta}[\hat{t}, \hat{t}_i]\frac{1}{\sigma_i^2}. \tag{25}$$

In this notation, hatted times are integer indices offset by a starting epoch, $T_0 \leqslant$ all t_i, scaled by δt, and rounded down to the lower integer: $\hat{t} = \text{floor}((t - T_0)/\delta t)$; and $\hat{\delta}$ is the Kronecker delta:

$$\hat{\delta}[m, n] = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases}. \tag{26}$$

For numerical reasons, the measurements are adjusted by the mean value,

$$\bar{x} = \frac{\sum_j \frac{x_j}{\sigma_j^2}}{\sum_j \frac{1}{\sigma_j^2}}, \tag{27}$$

and the value of A_0 found by the algorithm should have $\bar{x}$ added back to find the true mean value of the source.

When implemented, these discrete functions can be represented as arrays with indices $[0, \ldots, 2\mathcal{N} - 1]$. To search an acceptable density of frequencies, you should choose $(\mathcal{N}/H)\delta t \gtrsim \Delta T$. Most practical implementations of the FFT place requirements on $\mathcal{N}$ for the sake of efficiency, such as being a power of 2 or having only small prime factors.

The real arrays $\hat{p}[\hat{t}]$ and $\hat{q}[\hat{t}]$ are then passed to an FFT routine to get the complex arrays $\hat{P}[\hat{f}]$ and $\hat{Q}[\hat{f}]$. The discrete frequency indices $\hat{f} = 0, \ldots, N - 1$ correspond to frequencies $f = \hat{f}\delta f$. Many FFT implementations use the imaginary part of the $\hat{f} = 0$ element to store the cosine component at the Nyquist frequency $f_{\text{Nyquist}} = \mathcal{N}\delta f$.

(3) *Calculating the normal equations at each frequency.* Equations (16)–(23) describe how to construct α and β. The α matrix at a given $\hat{f}$ is based on the terms of $\hat{Q}$ at indices $\{0, \hat{f}, 2\hat{f}, \ldots, 2H\hat{f}\}$. The β vector is based on the terms of $\hat{P}$ at $\{0, \hat{f}, 2\hat{f}, \ldots, H\hat{f}\}$.

Streamlining the notation so that $\hat{Q}[n\hat{f}] = {}_{\mathcal{R}}Q_n + i_{\mathcal{I}}Q_n$ and $\hat{P}[n\hat{f}] = {}_{\mathcal{R}}P_n + i_{\mathcal{I}}P_n$, the $H = 2$ case can be written as

$$^{\hat{f}}\alpha = \frac{1}{2}$$

$$\begin{pmatrix} 2_{\mathcal{R}}Q_0 & 2_{\mathcal{I}}Q_1 & 2_{\mathcal{R}}Q_1 & 2_{\mathcal{I}}Q_2 & 2_{\mathcal{R}}Q_2 \\ 2_{\mathcal{I}}Q_1 & {}_{\mathcal{R}}Q_0 - {}_{\mathcal{R}}Q_2 & {}_{\mathcal{I}}Q_2 & {}_{\mathcal{R}}Q_1 - {}_{\mathcal{R}}Q_3 & {}_{\mathcal{I}}Q_3 + {}_{\mathcal{I}}Q_1 \\ 2_{\mathcal{R}}Q_1 & {}_{\mathcal{I}}Q_2 & {}_{\mathcal{R}}Q_0 + {}_{\mathcal{R}}Q_2 & {}_{\mathcal{I}}Q_3 - {}_{\mathcal{I}}Q_1 & {}_{\mathcal{R}}Q_1 + {}_{\mathcal{R}}Q_3 \\ 2_{\mathcal{I}}Q_2 & {}_{\mathcal{R}}Q_1 - {}_{\mathcal{R}}Q_3 & {}_{\mathcal{I}}Q_3 - {}_{\mathcal{I}}Q_1 & {}_{\mathcal{R}}Q_0 - {}_{\mathcal{R}}Q_4 & {}_{\mathcal{I}}Q_4 \\ 2_{\mathcal{R}}Q_2 & {}_{\mathcal{I}}Q_3 + {}_{\mathcal{I}}Q_1 & {}_{\mathcal{R}}Q_1 + {}_{\mathcal{R}}Q_3 & {}_{\mathcal{I}}Q_4 & {}_{\mathcal{R}}Q_0 + {}_{\mathcal{R}}Q_4 \end{pmatrix}, \tag{28}$$

$$^{\hat{f}}\beta = \begin{pmatrix} {}_{\mathcal{R}}P_0 \\ {}_{\mathcal{I}}P_1 \\ {}_{\mathcal{R}}P_1 \\ {}_{\mathcal{I}}P_2 \\ {}_{\mathcal{R}}P_2 \end{pmatrix}. \tag{29}$$

The minimum $^f\chi^2$ at each frequency is less than or equal to that for the constant value null hypothesis:

$$\chi_0^2 = \sum_i \frac{(x_i - \bar{x})^2}{\sigma_j^2}, \tag{30}$$

$$^f\chi^2 = \chi_0^2 - \sum_{j,k} {}^f\beta_j({}^f\alpha^{-1})_{jk}{}^f\beta_k$$

$$= \chi_0^2 - {}^f\Delta\chi^2. \tag{31}$$

The values of the Fourier coefficients, $^f\mathbf{A}_j = \sum_k({}^f\alpha^{-1})_{jk}{}^f\beta_k$, are not required for finding the minimum. However, they will typically be calculated as an intermediate result and may be used for further analysis.

(4) *Interpreting the result.* The value of $\hat{f}$ with the largest value of $^f\Delta\chi^2 = \sum_{j,k} {}^f\beta_j({}^f\alpha^{-1})_{jk}{}^f\beta_k$ (and thus the lowest χ^2) provides the best fit among the searched frequencies.

The $^f\Delta\chi^2$ value tells how much better the model at that frequency fits the data than the constant-value null hypothesis does. $^f\Delta\chi^2$ must be compared to what is expected by chance, given the number of additional free parameters in the model ($2H$) and the number of trials (representing independent frequencies searched).

The number of independent frequencies searched can be made arbitrarily large by decreasing the gridding interval, and so any given χ^2 improvement can in theory be diluted away to insignificance. However, the number of trials is only unbounded toward higher frequencies. If you search starting at low frequencies, the number of trials "so far" can be treated as being proportional to f. In that case, the value to be minimized is

$$p(f) = \hat{f}\mathcal{P}_{2H, \geqslant \Delta\chi^2}({}^f\Delta\chi^2), \tag{32}$$

where $\mathcal{P}_{2H,\geqslant\Delta\chi^2}$ is the cumulative χ^2 distribution with $2H$ degrees of freedom. $p(f)$ is proportional to the probability, given the null hypothesis, of finding such a large decrease in χ^2 by chance at a frequency f or below.

The use of f to adjust $\mathcal{P}$ is straightforward from a frequentist perspective. From a Bayesian perspective it corresponds to a prior assumption that the distribution of $\log(f)$ is uniform (which is equivalent to $\log(period)$ being uniform) over the search interval. A different adjustment could be made if a different Bayesian prior were desired.

Because the true minimum might fall between two adjacent frequency bins, and because the gridding of the data causes some sensitivity loss, frequencies in the vicinity of the minimum, and in the vicinities of other frequencies that have local minima that are almost as good, should be searched more finely. These searches should use the ungridded data to directly calculate a locally optimum $^f\Delta\chi^2$ and adjustments.

Multiple candidate frequencies can be extracted and examined to see if they can reject the null hypothesis using other statistics in combination with $^f\Delta\chi^2$. For example, in a search for stars with transiting planets a particular shape of light curve is expected. Finding an otherwise marginal $^f\Delta\chi^2$ in combination with this light-curve shape would be a convincing detection. Fχ^2 can quickly find all candidate frequencies with marginal or better $^f\Delta\chi^2$.

The reduced χ^2, the ratio of the minimized χ^2 to the degrees of freedom, is a useful measure of how well the data fits the model. However, even the correct frequency can produce a poor reduced χ^2 under several circumstances. The source may have a light curve that has sharp features, or is otherwise poorly described by an H-harmonic Fourier series. The source may have "noisy" variations in addition to periodic behavior. There may be multiple frequencies involved, as in Blazhko effect RR Lyraes, overtone Cepheids, or eclipsing binaries where one of the components is itself variable. In all these cases, the detection of a period can provide a starting point for further analysis of the source's behavior.

Because the source may be noisier (compared with the model) than the measurement error, an adjusted $\Delta\chi^2$ may be defined as

$$\Delta\chi^2_{\text{adj}} = \frac{\chi^2_0 - \chi^2(\mathbf{P})}{\Delta\chi^2_{\text{best}}/N_{\text{dof}}}.$$

For a noisy source, this will allow significance calculations to be based on the noise of the source, rather than the measurement error, preventing false positives due to fitting the source noise. For truly constant sources, and for sources that are accurately described by the harmonic fit, the measurement error will dominate. This is an improvement over the traditional method of using the standard deviation of the data as a surrogate for measurement error because it continues to incorporate the known instrumental characteristics and because it does not overstate the source noise of a well-behaved variable source. For any given fit, the period with the best $\Delta\chi^2$ will also have the best $\Delta\chi^2_{\text{adj}}$.

The best-fit frequency is not necessarily the "right answer." There are several effects that can produce χ^2 minima at frequencies that are not the frequency of the physical system being studied. Some examples of this are shown in Section 4. For example, a mutual-eclipsing binary of two similar stars may be better fit at $2f_{\text{orbit}}$ than at f_{orbit} for a given H.

If the set of sampling times has a strong periodic component, then this can produce "aliasing" against periodic or long-term variations. Irregular sampling improves the behavior of χ^2-based period searches. For alias peaks to be strong, there must
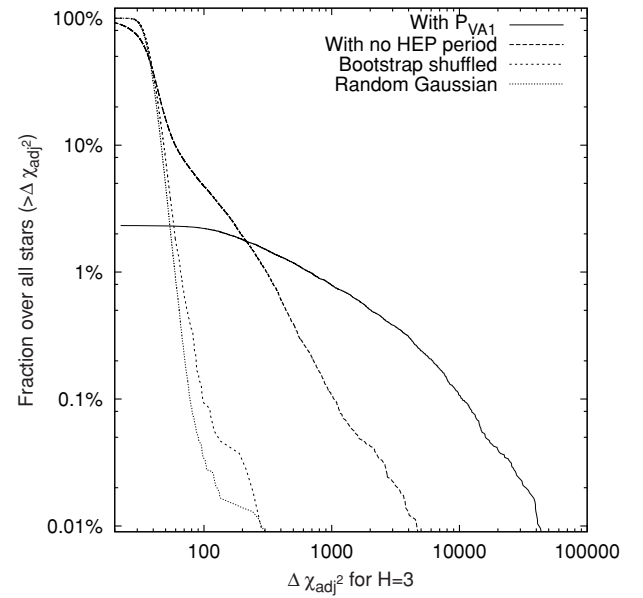
Figure 1. Cumulative distribution of $\Delta\chi^2_{\text{adj}}$ for HEP stars with known periods from VA1 (solid line) and those without VA1 periods (dashed line) as a fraction of all HEP stars. To test the false-positive rate, the analysis was repeated on two randomized samples of the data. In the *Bootstrap shuffle* (short-dashed line), the data for each star was shuffled so that each measurement value and its error estimate, $x_i \pm \sigma_i$, was randomly assigned to a measurement time t_i. For the *random Gaussian* (dotted line), the value at each t_i was replaced by random variables with constant mean and standard deviation $C \pm \sigma$.

be some frequency, f_{sample} for which a large fraction of the sampling times are clustered in phase to within a small fraction of the true source variation timescale. For many astronomical data sets taken from a single location only at night, this is the case with $f_{\text{sample}} = 1/\text{day}$ or $1/\text{siderealday}$. Depending on the observation strategy (e.g., observing each field as it transits the meridian vs. observing several times a night) the clustering in phase can be tighter or looser, and thus produce greater or lesser aliasing at short periods.

If there is long-term variation in the source, then there may be alias peaks near f_{sample} and its harmonics, even if the long-term variation is aperiodic. If the source combines a high-frequency periodicity with a higher amplitude long-term aperiodic variation, then the alias peaks can provide the lowest χ^2. To detect the short-period variation, the long-term variation may be removed with, e.g., a polynomial fit (although, as discussed in Section 4, this will not necessarily result in an improvement). An initial run of the Fχ^2 algorithm with $\delta t = 1/f_{\text{sample}}$ can be used to quickly test whether the long-term variations are themselves periodic.

One advantage of χ^2 methods over Lomb is that they do not produce alias peaks near f_{sample} if the source is constant. Near f_{sample}, the limited phase coverage of the samples provides only weak constraints on the amplitude of the Fourier components. Statistical fluctuations can produce large amplitudes (and thus high Lomb values), but the correspondingly large standard error on the amplitude ensures that such fits do not yield a large improvement in χ^2.

4. REFERENCE IMPLEMENTATION AND EXAMPLES

An implementation of the algorithm, written in C, is available at the author's Web site[1]. For maximum portability and clarity, it uses the GNU Scientific Library (GSL; Galassi et al. 2003)

---

[1] http://public.lanl.gov/palmer/fastchi.html
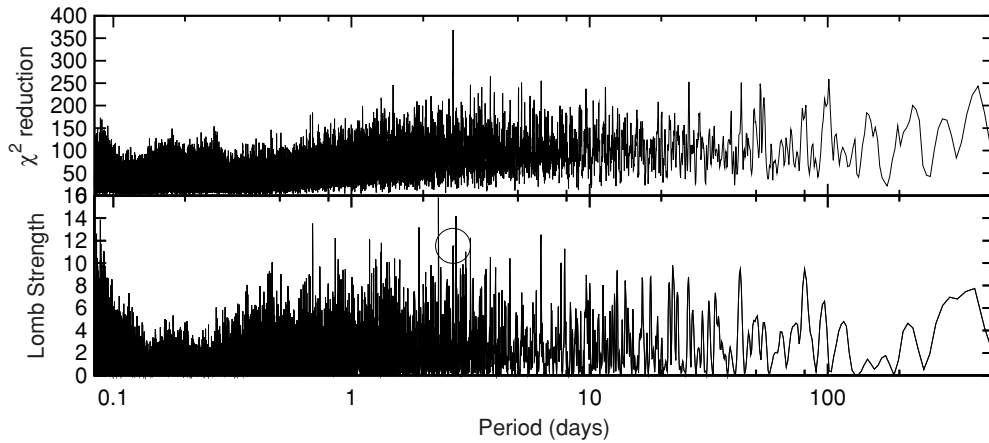
Figure 2. Strength of the signal as a function of period for HIP 69358 for $F\chi^2$ (top) and Lomb (bottom) algorithms. The Lomb peak corresponding to the $F\chi^2$ peak is circled.

and includes a driver interface that reads the input data as ASCII files.

The processing speed might be improved by using a different FFT package, such as FFTW (Frigo & Johnson 2005), by stripping the GSL and CBLAS layers from the core BLAS routines, and by using binary I/O. If many sources are measured at the same set of sampling times and have proportional measurement errors (e.g., an imaging instrument returning to the same field of view many times) then the same $^f\alpha^{-1}$ matrix may be reused on the $^f\beta$ vectors of all sources. Additional speedups are also possible.

However, before implementing such optimizations, users should balance the potential efficiency gains against the cost of the modification and determine whether the reference implementation is already fast enough for their purposes.

As a test case, the reference implementation was applied to the *Hipparcos* Epoch Photometry (HEP) Catalogue (van Leeuwen 1997).

Although the primary mission of *Hipparcos* was to provide high accuracy positional astrometry, it also produced a large easily available high-quality photometry data set. The processed data includes "Variability Annex 1" (hereafter VA1), listing variables with periods derived from the *Hipparcos* data, or periods from the previous literature consistent with the *Hipparcos* data. The HEP was later reprocessed by several groups, such as Koen & Eyer (2002), to discover additional periodic variables.

The complete data set of $>10^5$ stars, averaging $>10^2$ measurements each, spanning $>10^3$ days, takes <10 hr to search to frequencies up to $(2\ \text{hr})^{-1}$ and $H = 3$ on a standard desktop computer (a 2006 Apple Mac Pro with 2×2 Intel cores at 2.66 GHz). This is a factor of 2–3 times slower than the "Fast Computation of the Lomb Periodogram" code in *Numerical Recipes* Section 13.9, with comparable parameters for the search space, using the same computer and compiler. Profiling the code during execution finds that the bulk of the CPU time is split roughly evenly between the FFT and the linear regression stages.

Of the 118,218 stars in the HEP, 115,375 had sufficient high-quality data to be processed by the reference implementation (most of the others had interfering objects in the field of view). Of these, 2275 had periods listed in VA1, P_{VA1}, and had at least 50 measurements with no quality flags set spanning more than $3 \times P_{VA1}$.

In the subset of 2275 VA1 stars, 2066 (88.5%) had calculated periods that either agreed with (50.0%) or were harmonically re-
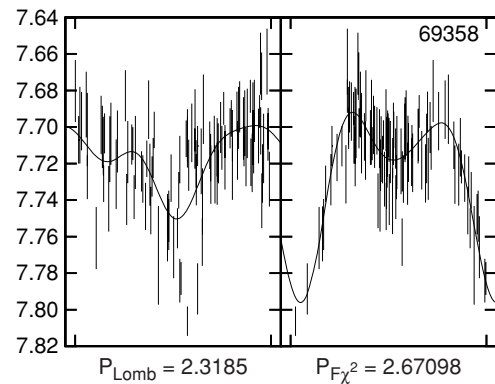

Figure 3. HEP data for HIP 69358 folded at the periods found by Lomb (left) and $F\chi^2$ (right) algorithms.

lated to P_{VA1}. In descending order of incidence, these harmonic relations were $2 \times P_{VA1}$ (15.7%); $1/2 \times P_{VA1}$ (12.8%); $3 \times P_{VA1}$ (9.5%); and $3/2 \times P_{VA1}$ (0.6%). The other 262 stars (11.5%) had calculated periods that were unrelated to the HEP value. The reference implementation has the ability to detrend the data with a polynomial fit (to search for periodic variation on top of a slow irregular variation). However, applying this to the HEP data, removing variation out to t^3, decreased the agreement with the VA1 periods, increasing the number of harmonically unrelated periods from 11.5% to 16.9%, and decreasing the number at P_{VA1} from 50.0% to 42.4%. Koen & Eyer recommend that if analysis with and without such detrending find different periods, then the periods should be considered unreliable.

Figure 1 shows the cumulative distribution of $\Delta\chi^2_{\text{adj}}$ for the HEP data and for randomized data. This demonstrates a well-behaved false-positive rate when observing constant sources. Excluding the VA1 stars (representing 2% of the population) a further $\sim$10% of the HEP stars are clearly distinguishable from randomized data by $F\chi^2$ algorithm. This indicates variability in these stars, although they do not have to be strictly periodic as long as there is significant power at some period.

Figure 2 shows a comparison of the results of the $F\chi^2$ and Lomb algorithms for an example star, HIP 69358, which is listed as an unsolved variable in the *Hipparcos* catalog. A single strong peak at $P_{F\chi^2} = 2.67098$ days is consistent with the period of 2.67096 days found by Otero & Wils (2005). However, the Lomb algorithm finds 12 peaks with higher strength than the
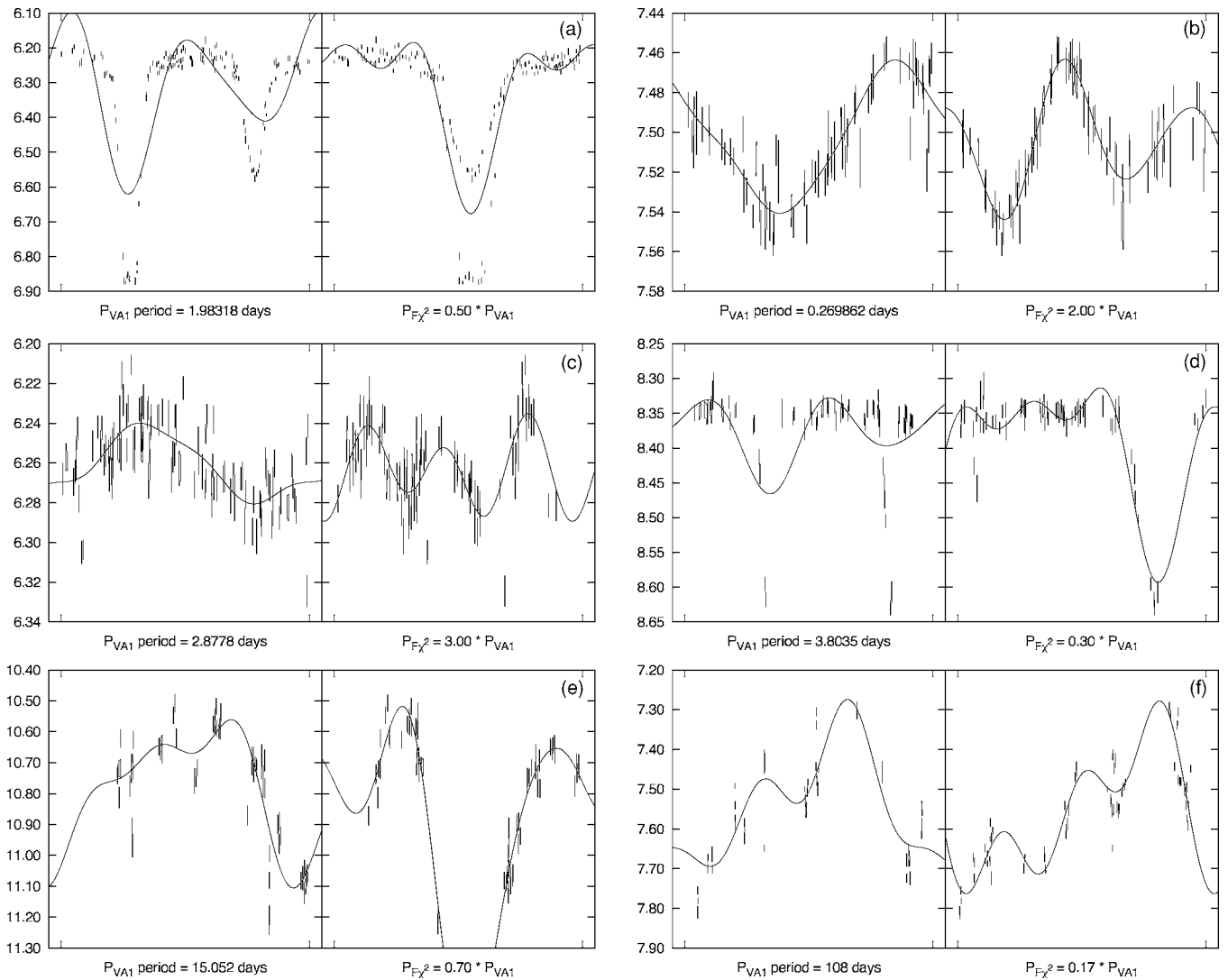
Figure 4. Light curves of various stars folded at $P_{\rm VA1}$ and $P_{\rm F\chi^2}$ to demonstrate some of the ways that $\rm F\chi^2$ can find a period other than the fundamental of the physical system. Star classification for these stars comes from the Simbad Database, operated at CDS, Strasbourg, France (http://simbad.u-strasbg.fr) (a) *HIP 109303 = AR Lac* is an RS CVn eclipsing binary with $P_{\rm VA1} = 1.98318$ days which $\rm F\chi^2$ best fits at half the period. Due to the narrow minima, the $H = 3$ Fourier expansion at the fundamental $\{1f, 2f, 3f\}$, provides a worse fit than the fit at half the period $\{2f, 4f, 6f\}$. However, the different depths of the two minima are clearly distinguishable, showing unambiguously that $P_{\rm VA1}$ is the true orbital period. (b) *HIP 10701 = AD Ari* has $P_{\rm VA1} = 0.269862$ days. However, $\rm F\chi^2$ finds a period of twice that, which appears to give a better fit to the data. This is a δ Sct type variable, which are characterized by multimodal oscillations, so power at multiple frequencies is to be expected. (c) *HIP 59678 = DL Cru* is an α Cyg variable. These stars have complex oscillations, so different cycles at the primary fundamental may have different apparent amplitudes. The $\rm F\chi^2$ algorithm to $H = 3$, for this data set, found a lowest χ^2 at triple the fundamental period. This splits the cycles during the observation into three sets, each of which can have a quasi-independent average amplitude adjusted by the Fourier components, allowing a slight additional χ^2 reduction through noise fitting. (d) *HIP 115647 = DP Gru* is an Algol-type eclipsing variable with narrow minima. A fit to the best $H = 3$ Fourier expansion is not a good model for the shape of this light curve, and so the $\rm F\chi^2$ would not be very effective in finding this period unless H is increased. However, the HEP data samples only three minima, spread over 145 orbits, and so may not be sufficiently constraining to uniquely determine a period, regardless of the algorithm used. (e) *HIP 98546 = V1711 Sgr* is a W Vir type variable. $P_{\rm VA1} = 15.052$ days, but $\rm F\chi^2$ gives 10.566 days. The $\rm F\chi^2$ fit is better in the formal sense of having a lower χ^2 with the line passing closer to the data points, but it does not look like a W Vir light curve. To add to the confusion, the General Catalog of Variable Stars (http://www.sai.msu.su/gcvs) gives $P = 28.556$ days, but this period is not apparent in the *Hipparcos* data. (f) *HIP 12557 = W Tri*, a semiregular pulsating star, has $P_{\rm VA1} = 108$ days, which is the third-best value found by $\rm F\chi^2$ (after 17.99 days and 22.68 days).

one at that period. As seen from the folded light curves at the strongest $P_{\rm Lomb}$ and $P_{\rm F\chi^2}$ (shown in Figure 3), $\rm F\chi^2$ found the characteristic light curve of an eclipsing binary while Lomb found a noise peak.

Cases where $\rm F\chi^2$ does not find the fundamental period are useful for examining the limitations of this algorithm. Some examples of this are presented in Figure 4 and discussed in its caption.

5. CONCLUSION

The $\rm F\chi^2$ technique is a statistically efficient, statistically valid method of searching for periodicity in data that may have irregular sampling and nonuniform standard errors.

It is sensitive to power in the harmonics above the fundamental frequency, to any arbitrary order.

It is computationally efficient, and can be composed largely of standard FFT and linear algebra routines that are commonly available in highly optimized libraries.

A reference implementation is available and can be easily applied to your data set.

Facilities: HIPPARCOS

REFERENCES

Baluev, R. V. 2009, MNRAS, in press
Frigo, M., & Johnson, S. G. 2005, Proc. IEEE, 93, 216
Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M., & Rossi, F. 2003, GNU Scientific Library Reference Manual (2nd ed.; Bristol: Network Theory Ltd.)
Irwin, J., et al. 2006, MNRAS, 370, 954
Koen, K., & Eyer, L. 2002, MNRAS, 331, 45
Lomb, N. R. 1976, Ap&SS, 39, 447
Otero, S. A., & Wils, P. 2005, IAU Inform. Bull. Var. Stars, 5644, 1
Press, W. H., & Rybicki, G. B. 1989, ApJ, 338, 277
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 2002, Numerical Recipes in C++: The Art of Scientific Computing (Cambridge: Cambridge Univ. Press)
Stellingwerf, R. F. 1978, ApJ, 224, 953
van Leeuwen, F. 1997, in ESA SP-402: *Hipparcos*—Venice 97, ed. B. Battrick, M. A. C. Perryman, & P. L. Bernacca (Noordwijk: ESA), 19