

Introduction to Time Series Analysis

Extracted from my Ph.D. Dissertation (1994)

W. T. Bridgman

cygnusx1@mac.com

1. Introduction to Time Series Analysis

1.1. Classifications

1.1.1. Characteristics of Distributions

A set of random numbers x_i can be characterized by their distribution. Two measures of this distribution are its mean and standard deviation, also described as the first and second moments of the distribution. For a continuous distribution with a probability density $p(x)$, the mean, μ_x , and standard deviation, σ_x^2 , are given by

$$\mu_x = E[x] = \int_{-\infty}^{+\infty} x p(x) dx \quad (1-1)$$

$$\sigma_x^2 = E[(x - \mu_x)^2] = \int_{-\infty}^{+\infty} (x - \mu_x)^2 p(x) dx \quad (1-2)$$

where $E[]$ is called the expectation value operator. However, in experimental situations, we don't always have the luxury of a known probability function. We are given a set of discrete data points and must work from there. In these cases, we define the sample mean, denoted \bar{x} or $\hat{\mu}_x$, and the sample variance, denoted $\hat{\sigma}_x^2$, by

$$\bar{x} = \hat{\mu}_x = \frac{1}{N} \sum_{j=1}^N x_j \quad (1-3)$$

and

$$var(\{x_j\}) = \hat{\sigma}_x^2 = \frac{1}{N-1} \sum_{j=1}^N (x_j - \bar{x})^2 \quad (1-4)$$

where x_j is a set of N discrete random numbers. Note that the sample variance is divided by $N-1$ rather than N . This reduces the bias error in the estimate but in practice N should be so large that this difference doesn't matter. The quantity σ_x is called the standard deviation.

Another function often used to characterize distributions is the third moment, or *skewness*, defined as

$$Skew(\{x_j\}) = \frac{1}{N} \sum_{j=1}^N \left[\frac{x_j - \bar{x}}{\sigma_x} \right]^3$$

where σ_x is the standard deviation of the distribution. Dividing by the standard deviation enables the skewness to be defined as dimensionless. Skewness measures the asymmetry of a distribution –

positive (negative) skewness indicates the asymmetry extends towards positive (negative) x(Press et al. 1992, p. 612).

The fourth moment, or *kurtosis*, of a distribution measures its peakedness or flatness relative to the normal distribution. A distribution with a positive kurtosis is sharply peaked, or *leptokurtic*, while a negative kurtosis indicates a flattened or *platykurtic* distribution. In between, near that of a normal distribution, is *mesokurtic*. The definition of kurtosis is chosen as

$$Kurt(\{x_j\}) = \left\{ \frac{1}{N} \sum_{j=1}^N \left[\frac{x_j - \bar{x}}{\sigma_x} \right]^4 \right\} - 3$$

where the -3 term makes the value zero for the normal distribution(Press et al. 1992, p. 612).

1.1.2. Basic Definitions

A time series is defined as an ordered sequence of observations. *Ordered* is a key term here, for the sequence of the observations also carries information. There are a number of terms for characterizing time series.

A *continuous* time series is one that is defined for all real values of some time parameter. A good example of a continuous series would be some function of time, say, $f(t)$, where for any real value of t you obtain an observational value. Continuous time series are more often the results of modeling rather than data.

Time series are most commonly used in the *discrete* form where you have a sequence of observations taken at fixed points in time. In general, the points in time when the measurements are taken may be evenly spaced or unevenly spaced.

A time series is called *stationary* if the mean and variance of the distribution are constant over some time interval. If the mean or variance change with time, the series is called *non-stationary*.

Another important distinction that must be made in statistical studies is the difference between a parameter and its *estimator*. For example, equation 1-1 defines the mean of a time series. However, since this definition is rather difficult to apply to an actual time series, we define the estimator for the mean, in this case, equation 1-3. Since we use different expressions for a parameter and its estimator, it is important to know how good an estimator is. For this we define the *bias*, $B(\alpha)$, of an estimator $\hat{\alpha}$ of the parameter α which is the difference between the true value of the estimator and the expected value of the estimate,

$$B(\alpha) = \alpha - E[\hat{\alpha}]. \tag{1-5}$$

Ideally, we want an estimator to be *unbiased*, or $B(\alpha) = 0$.

The mean, variance, skewness, kurtosis, and other statistical quantities have been defined as statistical ensemble averages. In practice, we are given some time series, not a statistical ensemble,

from which we must estimate these quantities. Ideally, we would like to use the time averages we compute from the data for the ensemble averages. For this to work, the random process must be *ergodic*. The process is ergodic if all its statistics can be computed from the single time series and those statistics are equal to the ensemble average. For a time series to be ergodic, it must be stationary up to its fourth moments(Jr. 1987, p. 121).

1.2. Continuous and Discrete Fast Fourier Transforms

The continuous Fourier transform converts a function $g(t)$ in the time domain to a function $G(f)$ in the frequency domain. The transform is usually given in pairs,

$$G(f) = \int_{-\infty}^{\infty} g(t) e^{-i 2 \pi f t} d t$$

and

$$g(t) = \int_{-\infty}^{\infty} G(f) e^{i 2 \pi f t} d f$$

where f is the frequency. The normalization outside the integral varies from author to author. The function $G(f)$ contains information on the frequency and phase of a continuous sequence of eikonal ($e^{i(2\pi ft+\theta)}$) waves which must be added to generate the input signal $g(t)$.

The above forms of the Fourier transform are suitable for analytic computation, but suppose we wish to search a time series of data for possible periodic components. We immediately encounter two problems: (1) we only have a finite amount to data so the infinite integration limits are useless; (2) the data are binned at discrete times so the function is no longer continuous in our conventional definition. These limitations force us to convert from integration to summation and define the Discrete Fourier Transform pair(Brigham 1974, pp. 96-97)

$$G\left(\frac{n}{N T}\right) = \sum_{k=0}^{N-1} g(k T) e^{-i 2 \pi n k / N}, \quad n = 0, 1, \dots, N - 1$$

and

$$g(k T) = \frac{1}{N} \sum_{n=0}^{N-1} G\left(\frac{n}{N T}\right) e^{i 2 \pi n k / N}, \quad k = 0, 1, \dots, N - 1$$

where N is the number of points in the time series and T is the width of the time bins. The total length of the time series is therefore NT . The frequency corresponding to bin n is given by

$$f_n = \frac{n}{N T}$$

and the time corresponding to bin k is

$$t_k = k T.$$

At present, the fastest known algorithm for computing the discrete Fourier transform is called the Fast Fourier Transform (FFT). Most software numerics libraries implement this algorithm for

the computation of the Fourier transform. When using such routines, always be aware of the normalization chosen for the routine.

The discretization process places two new restrictions on the resulting transform. First is that the transform has a finite frequency resolution, the width of the frequency bins given by

$$f_o = \frac{1}{NT}.$$

Also, the transform is determined only up to a maximum frequency, called the Nyquist frequency, given by

$$f_{max} = f_c = \frac{1}{2T}.$$

The transition from the continuous to discrete Fourier transform also creates a number of side effects related to the two problems above. First, sampling the signal function using the value of the function at regular discrete times, introduces a phenomenon called aliasing in the frequency domain. In addition, the finite truncation of the time series produces rippling in the frequency domain. In general, truncation of the function in one domain introduces oscillations in the other (Brigham 1974, p. 140). This phenomenon is called leakage and is discussed in more detail below. Their effects are illustrated in Figure 1.

1.2.1. Leakage

If a periodic function is truncated at other than an integral multiple of the period, the truncation in the time domain generates side lobes in the frequency domain. Figure 1 illustrates this effect for the simple function

$$f(t) = e^{-|t|/\tau}$$

defined from $+\infty < t < -\infty$ and τ is a time-scaling parameter. The Fourier transform of this function is

$$F(f) = \frac{2\tau}{1 + 4\pi^2 f^2 \tau^2} \tag{1-6}$$

which is graphed as a long-dash line in the Fourier transform portion of the graph (note the amplitude is logarithmically scaled). If the function is truncated at some time $\pm NT$, the transform becomes

$$F(f) = \frac{2\tau}{1 + 4\pi^2 f^2 \tau^2} \left\{ 1 + e^{-N T/\tau} [2\pi f \tau \sin(2\pi f N T) - \cos(2\pi f N T)] \right\}$$

which is the short-dashed line in the figure.

These side lobes create an additional frequency component which we call leakage. This truncation is equivalent to convolving the frequency domain with the function $\sin(2\pi f NT)/(2\pi f NT)$. If the truncation interval is chosen equal to an integral multiple of the period, NT , the frequency

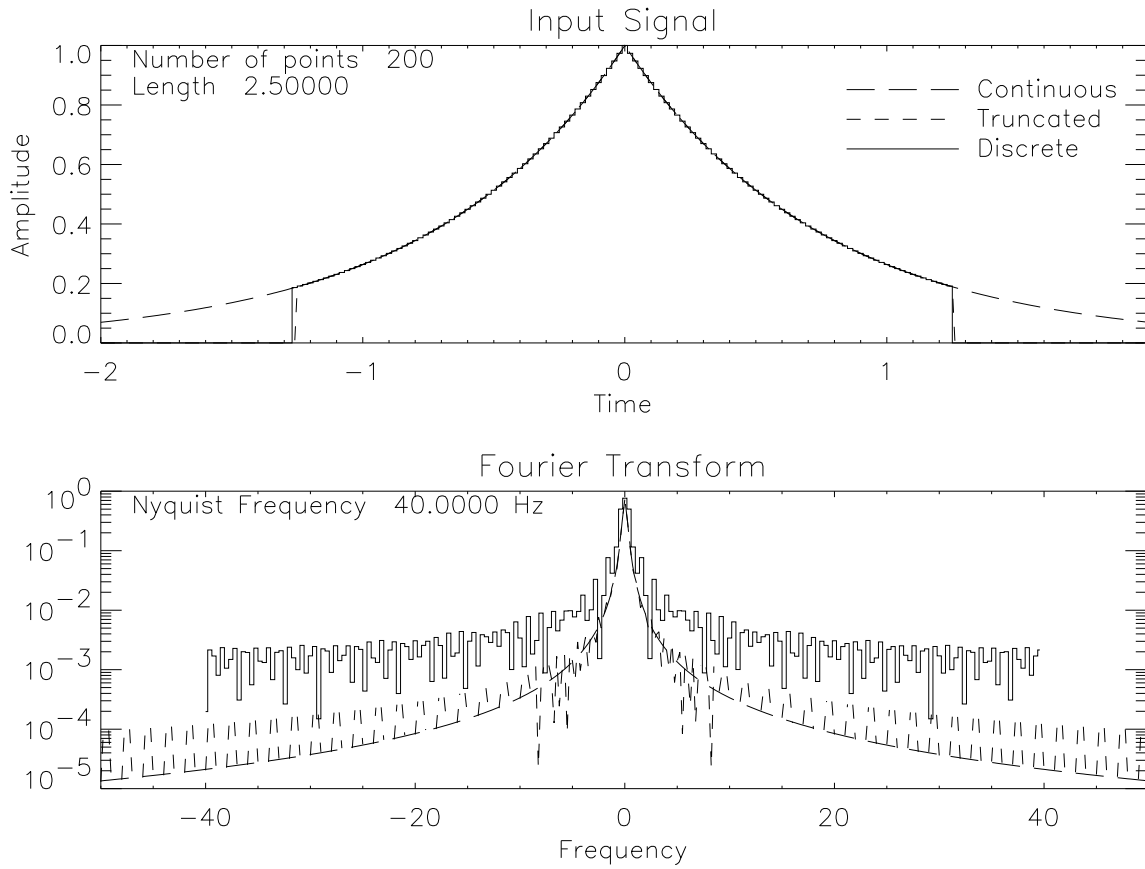


Fig. 1.— Conversion from a continuous Fourier transform to the fast Fourier transform. Conversion from a continuous Fourier transform to the fast Fourier transform. Note that for the fast Fourier transform, at some frequencies the amplitude was negative which is not manageable on a logarithmic scale. To get around this, the absolute value of the FFT was plotted.

domain sampling is coincident with the zeros of $\sin(2\pi fNT)/(2\pi fNT)$ and the side lobes vanish. In this case, the discrete Fourier transform yields the same results as the continuous Fourier transform.

To reduce this leakage, we need to use a time domain truncation function which has side lobes of smaller magnitude than $\sin(2\pi fNT)/(2\pi fNT)$. The smaller side lobes generate less leakage effects in the discrete transform. However, such leakage reduction does not come free. Any non-zero frequency components are broadened with respect to the original frequency. In general, the more leakage is reduced, the broader the spectral lines appear in the discrete transform (Brigham 1974, pp. 140-146).

The most popular function for this purpose is the Hanning filter which is applied in the time domain (before performing the Fourier Transform) and has the functional form

$$H(k) = \frac{1}{2} \left[1 - \cos \left(\frac{2\pi k}{N} \right) \right] \quad (1-7)$$

where k is the time bin number and N is the total number of bins in the series.

1.2.2. Aliasing

The Sampling Theorem states that if a signal is composed of no frequencies above the Nyquist frequency for a given sampling rate, the Discrete Fast Fourier Transform will completely determine the frequency characteristics of the signal. However, if the signal contains frequency components above this Nyquist frequency, the power from those frequencies will be “wrapped-around” or *aliased* into the Nyquist range. This is obvious in the discrete portion of the plot in Figure 1 where the “wings” of the absolute value of the Fourier transform are much higher than the continuous form. In general, for frequency f in the Nyquist range $0 \leq f < f_c$ of a Fast Fourier Transform, there will be contributions from higher frequencies given by

$$(2f_c \pm f), (4f_c \pm f), \dots, (2nf_c \pm f), \dots$$

(Bendat and Piersol 1986, p. 338).

1.2.3. Effects of a Low Signal-to-Noise Ratio

Consider a photon source whose emission rate is oscillating at a single well-defined frequency. If the intensity is so low that the detector records in each time bin either one or no photons from the source, the time series of the source (not including noise) appears similar to a sequence of square pulses. When an FFT is performed on this time series, harmonics appear at integral multiples of the original frequency. The noise level can affect how prominently these harmonics appear in the power spectrum.

1.3. Power Spectra

1.3.1. The One-sided Power Spectral Density

If we use Parseval’s Theorem

$$\int_{-\infty}^{+\infty} g(t) h^*(t) dt = \int_{-\infty}^{+\infty} G(f) H^*(f) df$$

and set $g(t) = h(t)$, then we obtain the Energy Theorem

$$E = \int_{-\infty}^{+\infty} |h(t)|^2 dt = \int_{-\infty}^{+\infty} |H(f)|^2 df$$

which simply states that energy is conserved in transformations between the time and frequency domains. We can therefore define a quantity

$$S(f) = |H(f)|^2$$

called the power spectral density (PSD) of the signal $h(t)$. The power spectral density represents the distribution of energy as a function of frequency (Jr. 1987, p. 33).

Since the power spectral density is symmetric in frequency space (i.e. $S(-f) = S(f)$), then it is convenient to define the *one-sided power spectral density*, $G(f)$, as

$$G(f) = S(-f) + S(f) = 2S(f).$$

This is the quantity measured when direct filtering procedures are used in time series analysis and is the quantity used in this work.

1.3.2. Errors in the Basic Estimates

For the purposes of detecting periodicities in a time series, what we wish to compute is the one-sided power spectral density, $G(f)$. However, with a real data set comes the question of how good is the discrete fast Fourier transform (DFFT) as an estimator of the power spectral density (PSD)? Unfortunately, the answer is, not very good. Through some very elaborate mathematics, it is possible to show that not only is the the PSD calculated from the DFFT a biased estimate of the PSD, but that the variance is as great as the PSD itself; namely,

$$\text{var} [G(f)] \geq E [G(f)]^2 \tag{1-8}$$

(Jr. 1987, pp.127–129). It is important to note that this result is valid for a noise spectrum which is white (the same over all frequencies) and Gaussian. This effect is illustrated in the first two panels of Figure 2. The first panel shows the analytic computed PSD for an ARMA model¹. After using

¹See Appendix B for a description of ARMA models.

the ARMA model to generate a time series of known underlying properties, the FFT is used to calculate the PSD from the time series. The second panel illustrates the results of this calculation. It is easy to see that the power can vary significantly between neighboring frequency bins. For clarity, no error bars were plotted on this panel. With a single time series, therefore, it could be easy to misidentify a statistical fluctuation as a spectral line.

So how is this problem resolved? The best solution, but somewhat counterintuitive, is to compute the weighted average of the PSDs of equivalent, uncorrelated data sets. The given time series can be divided into N equal-sized data sets and the PSD, $\hat{G}(f)$, computed for each. The results are then averaged

$$\hat{G}(f) = \frac{1}{N} \sum_{i=1}^N \hat{G}_i(f). \quad (1-9)$$

If the number of averages is large, say 30 or more, the PSD estimates, $\hat{G}(f)$, will have an approximately normal sampling distribution and the variance of the estimate will be reduced to

$$var [\hat{G}(f)] \cong \frac{1}{N} [\hat{G}(f)]^2 \quad (1-10)$$

(Bendat and Piersol 1986, pp. 353–355). The third panel of Figure 2 shows the PSD and the resulting error bars when the time series is divided into eight segments and sent through this procedure. The fourth panel demonstrates this for the series split into sixty-four equal segments. The solid line is the analytic model, the dashed line is the PSD averaged for the segmented series, and the dot-dashed line illustrates the result if the PSD from the second panel is bin-averaged. These results are in very good agreement with the error bars computed from the analytic model. I found this to be a surprising result. On the surface, it appears that the square of the sum of a sequence of FFTs (squared to produce the PSD) is the same as the sum of the squares of the sequence of FFTs. Apparently cross-terms between the FFT sequences almost perfectly cancel out (if they exactly canceled out, the two lines would be identical). This is also a very convenient result since in many cases the time series you want to analyze may be too long to run as one single power spectrum. Breaking the series up into smaller units for analysis and averaging the results becomes an easy solution for this problem.

1.4. Correlation Functions

The correlation function is a measure of the a signal’s evolution in time, how its behavior at some time t is related to its behavior at some later (or earlier) time $t + \eta$. There are a myriad of definitions that fall under the name of correlation functions (sometimes called covariance functions), each with its own advantages and disadvantages. The functional form I have used for the cross-correlation function between two data sets $\{x_j\}$ and $\{y_j\}$ is

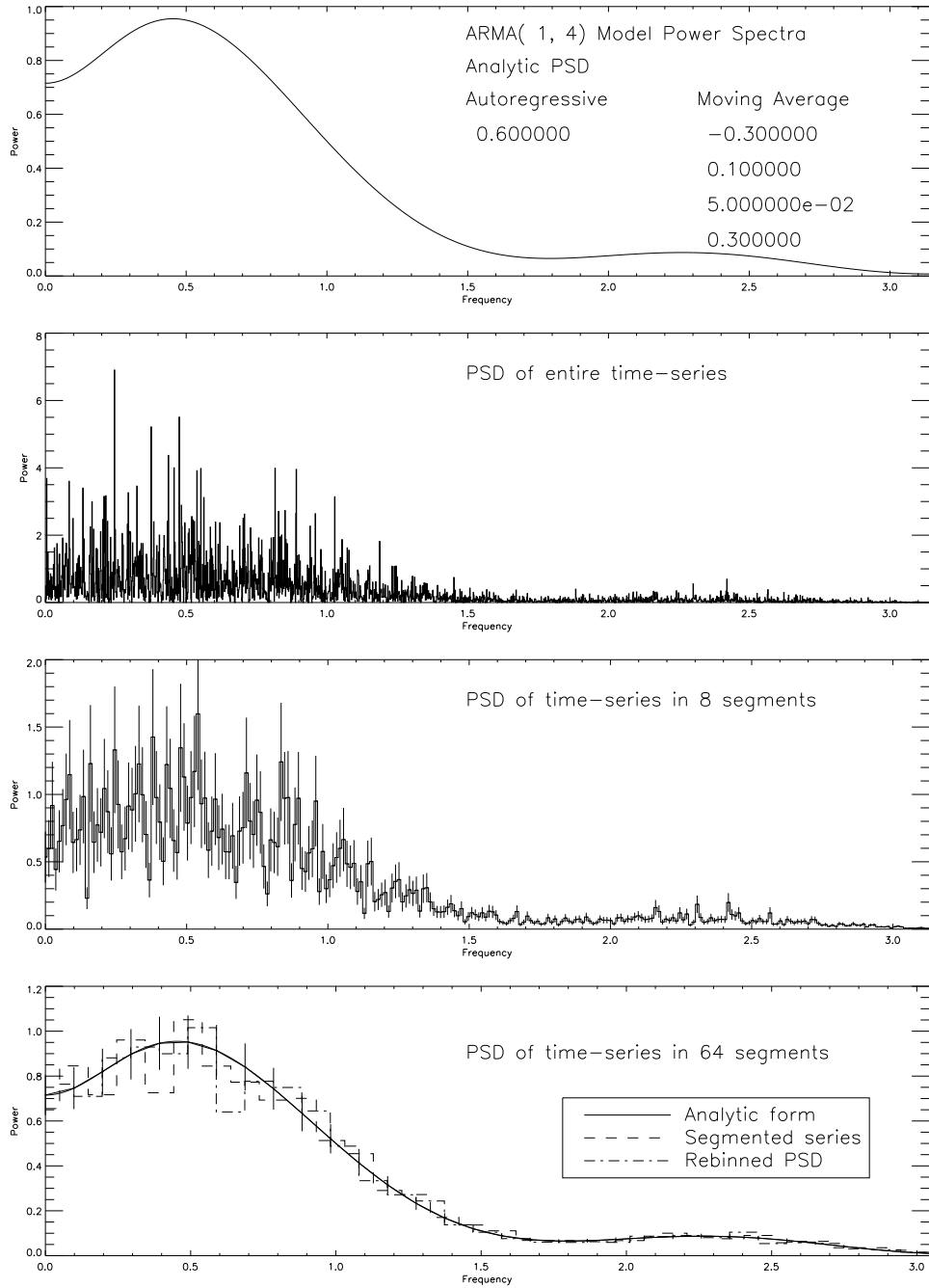


Fig. 2.— **Improving PSD error estimates by segmenting time-series records.** For clarity, no error bars are shown in the second panel. For the bottom panel, the error bars are computed from the analytic model.

$$r_u^{xy} = \frac{1}{N-u} \sum_{i=0}^{N-u-1} (x_i - \bar{x})(y_{i+u} - \bar{y}) \quad (1-11)$$

where the index u represents the time lag bin. This is an *unbiased* estimator for the cross-correlation function (Jr. 1987, pp. 146–149). To generate the auto-correlation function, the data set $\{y_j\}$ is replaced by $\{x_j\}$. It is important to understand what form has been chosen for the correlation function when deriving results based on the data analysis.

The auto-correlation function is directly related to the power spectral density. Taking the Fourier transform of the ACF generates the PSD as in

$$G(f) = 2T \sum_{m=0}^{\infty} r_m^{xx} \cos(2\pi f m T) \quad (1-12)$$

where r_m^{xx} is the auto-correlation function at lag bin m and use has been made of the fact that the ACF is real and symmetric (Jr. 1987, pp. 117–118). One might ask, since the PSD and ACF contain the same information, why do both? The best response to this very legitimate question is a question of sensitivity to what you may be looking for in the analysis. The ACF and PSD are sensitive to different types of behaviors in the time series. Consider for example a simple sinusoidal signal. Such a signal would appear as a sharp line in a PSD, whereas it would be spread out across fairly large lag times in an ACF and might be undetectable. Similarly, suppose the signal consists of some pulse which occurs at random times but is always followed, say a second later, by a second pulse of half the intensity of the original pulse. In the ACF, the second pulse would appear as a sharp line at a lag of one second, yet this line would be spread out over all frequencies in the PSD and might become much more difficult to detect. The best reason to look at both the ACF and PSD is that they are sensitive to *different* characteristics of the time series.

1.5. Lomb-Scargle Periodogram

One serious limitation of the Discrete Fourier Transform is that it requires the data points for analysis to be equally separated in time. For some data gathering operations, that can be a serious limitation. Astronomy in particular must conduct data gathering operations when weather and scheduling permit so equally spaced data points are hard to come by. A periodogram originally defined by Lomb has been modified by Scargle (1982) to handle this problem. It is occasionally referred to as the Lomb-Scargle periodogram (Press et al. 1992, pp.575–581). The basic form

of this periodogram is

$$P_N(\omega) \equiv \frac{1}{2\sigma^2} \left\{ \frac{\left[\sum_{j=1}^N (h_j - \bar{h}) \cos[\omega(t_j - \tau)] \right]^2}{\sum_{j=1}^N \cos^2[\omega(t_j - \tau)]} + \frac{\left[\sum_{j=1}^N (h_j - \bar{h}) \sin[\omega(t_j - \tau)] \right]^2}{\sum_{j=1}^N \sin^2[\omega(t_j - \tau)]} \right\} \quad (1-13)$$

where $\omega \equiv 2\pi f$ is the frequency, \bar{h} is the mean, σ^2 is the variance, and τ is defined by

$$\tan(2\omega\tau) = \frac{\sum_{j=1}^N \sin 2\omega t_j}{\sum_{j=1}^N \cos 2\omega t_j}. \quad (1-14)$$

This offset τ makes $P_N(\omega)$ independent of shifting the times, t_i , by a constant. In addition to reducing to the discrete Fourier transform when the data are equally spaced, Equation 1-13 also has the form one would obtain if the data set were linear least-squares fitted to the model

$$h(t) = A \cos(\omega t) + B \sin(\omega t) \quad (1-15)$$

for a given frequency ω . Another advantage of the Lomb-Scargle periodogram is that $P_N(\omega)$ has an exponential probability distribution with a mean of unity. If we calculate the periodogram at N independent frequencies, the probability that one frequency will have an amplitude larger than z is

$$p(> z) \equiv 1 - (1 - e^{-z})^M \quad (1-16)$$

and is the significance level of the peak. If this probability is small, the peak indicates a highly significant periodic signal. Since we are usually interested in the region of low significance, the probability can be expanded so

$$p(> z) \cong M e^{-z}. \quad (1-17)$$

2. ARMA (Autoregressive-Moving Average) Models

2.1. The Basics

When working with time series, it is desirable to have a method of modeling a time series driven by random processes. Ideally this would be a parameterized model and it would be possible to derive the power spectral density (PSD) and auto-correlation function (ACF) from the parameters of the model. A class of models which satisfy this criterion are the Autoregressive (AR), the Moving Average (MA), and the Autoregressive–Moving Average (ARMA) models. For each of these models, we can completely describe the PSD and ACF in terms of the parameters of the model.

Moving Average (MA) models, designated $MA(q)$, have the form

$$x_n = \sum_{k=1}^q b_k u_{n-k} + u_n \quad (2-1)$$

where u_n is a random process which serves as an input driving sequence. The parameter q designates the order of the process.

Autoregressive (AR) models, designated $AR(p)$, have the form

$$x_n = - \sum_{k=1}^p a_k x_{n-k} + u_n \quad (2-2)$$

where again a random process, u_n is used to drive the series. The parameter p designates the order of the process.

These forms are combined in the ARMA model, usually designated $ARMA(p, q)$, where p is the order of the autoregressive portion and q is the order of the moving average portion of the sequence. The ARMA model is generated by the sequence

$$x_n = - \sum_{k=1}^p a_k x_{n-k} + \sum_{k=0}^q b_k u_{n-k} \quad (2-3)$$

where we can set $b_0 = 1$ with no loss of generality as the series can always be scaled appropriately.

2.2. Power Spectra of ARMA Models

If the noise generator process is Gaussian, it can be shown that the PSD of the $ARMA(p, q)$ model is given by

$$G_{ARMA}(f) = T \sigma_w^2 \left| \frac{B(f)}{A(f)} \right|^2 \quad (2-4)$$

where T is the size of a time bin and

$$A(f) = 1 + \sum_{k=1}^p a_k e^{-2 \pi i f k T} \quad (2-5)$$

and

$$B(f) = 1 + \sum_{k=1}^q b_k e^{-2\pi i f k T} \tag{2-6}$$

are the characteristic polynomials of the autoregressive and moving average components, respectively, of the system. The quantity σ_w^2 is the variance of the white noise process (assumed to have zero mean)(Jr. 1987, pp. 174–177)².

2.3. Auto-Correlations of ARMA Models

In constructing ARMA models, the auto-correlation function provides the major key. The autoregressive and moving average components have very strong characteristics which appear in the ACF. Qualitative aspects of the model can be determined from a cursory examination of the auto-correlation function and these are summarized belowWei (1990).

Pure Autoregressive: AR(p). A purely autoregressive model ACF exhibits an exponential decay or a damped sine wave behavior at large lags.

Pure Moving Average: MA(q). For a pure moving average model of order q, the ACF has a sharp decay which terminates after lag q.

Autoregressive–Moving Average: ARMA(p,q). In the combined autoregressive–moving average model, the auto-correlation function exhibits a gentle decay after lag $q - p$.

The parameters of the ARMA model can be used to construct the auto-correlation function(Jr. 1987, pp. 181–182)

$$r_m^{xx} = \begin{cases} r_{-m}^{xx} & m < 0 \\ -\sum_{k=1}^p a_k r_{m-k}^{xx} + \sigma_w^2 \sum_{k=m}^q b_k h_{k-m}^* & 0 \leq m \leq q \\ -\sum_{k=1}^p a_k r_{m-k}^{xx} & m > q \end{cases} \tag{2-7}$$

where h_k^* is determined from the polynomial solution of $H(z) = B(z)/A(z)$ with

$$A(z) = 1 + \sum_{k=1}^p a_k z^{-k}$$

²White noise is characterized as having the same power at *all* frequencies (an idealization, since in reality it would require infinite power). The auto-correlation function of a white noise process is equal to the variance at zero lag and is zero at all lags other than zero. A white noise process can be generated with a uniform, Poisson, Gaussian, or some other probability distribution, so long as it has no correlations in *time*.

$$B(z) = 1 + \sum_{k=1}^q b_k z^{-k}$$

and

$$H(z) = 1 + \sum_{k=1}^{\infty} h_k z^{-k}.$$

However, the inversion operation — deriving the ARMA parameters from the ACF — is far from simple. A simple ARMA model with the corresponding power spectra and auto-correlation function is presented in Figure 3.

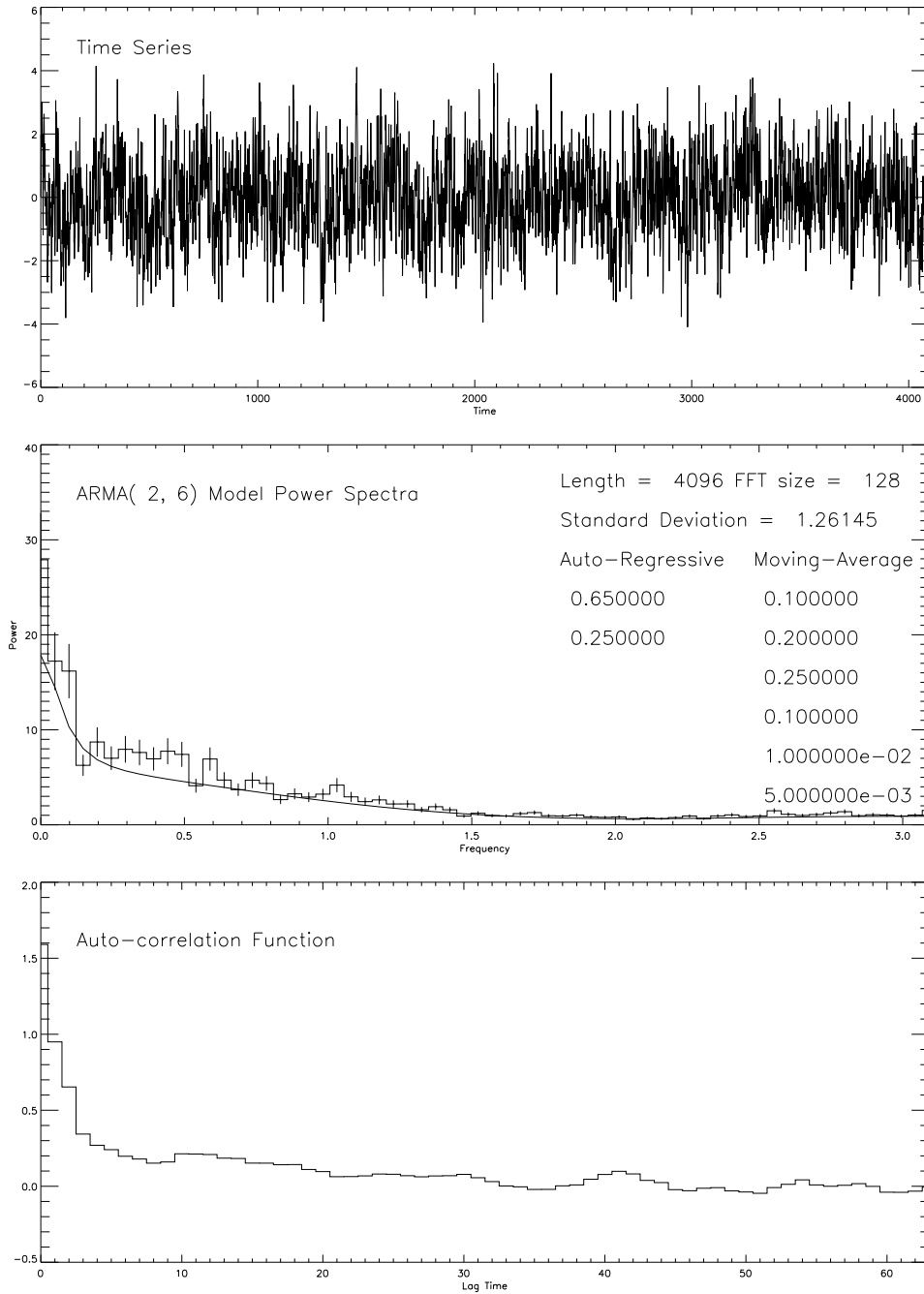


Fig. 3.— An ARMA model with its PSD and ACF. While not a precise model of Cygnus X-1's behavior, it does illustrate some of the general features of the PSDs and ACFs. The PSD shows most of the power at low frequencies (note the scale is linear, not logarithmic). The ACF also exhibits the sharp decline at short lags which changes to a more gentle decline at larger lags.

REFERENCES

- Julius S. Bendat and Allan G. Piersol. *Random Data: Analysis and Measurement Procedures, 2nd Edition*. Wiley-Interscience, 1986.
- Oran E. Brigham. *The Fast Fourier Transform And Its Applications*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1974.
- S. Lawrence Marple Jr. *Digital Spectral Analysis with Applications*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1987.
- W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C. The art of scientific computing*. Cambridge: University Press, —c1992, 2nd ed., 1992.
- J. D. Scargle. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *ApJ*, 263:835–853, December 1982. doi: 10.1086/160554.
- William Wei. *Time Series Analysis—Univariate and Multivariate Methods*. Addison-Wesley Publishing, Reading, MA, 1990.