11 Time series analysis

11.1 The astronomical context

Time-domain astronomy is a newly recognized field devoted to the study of variable phenomena in celestial objects. They arise from three basic causes. First, as is evident from observation of the Sun's surface, the rotation of celestial bodies produces periodic variations in their appearance. This effect can be dramatic in cases such as beamed emission from rapidly rotating neutron stars (pulsars).

Second, as is evident from observation of Solar System planets and moons, celestial bodies move about each other in periodic orbits. Orbital motions cause periodic variations in Doppler shifts and, when eclipses are seen, in brightness. One could say that the birth of modern time series analysis dates back to Tycho Brahe's accurate measurement of planetary positions and Johannes Kepler's nonlinear models of their behavior.

Third, though less evident from naked eye observations, intrinsic variations can occur in the luminous output of various bodies due to pulsations, explosions and ejections, and accretion of gas from the environment. The high-energy X-ray and gamma-ray sky is particularly replete with highly variable sources. Classes of variable objects include flares from magnetically active stars, pulsating stars in the instability strip, accretion variations from cataclysmic variable and X-ray binary systems, explosions seen as supernovae and gamma-ray bursts, accretion variations in active galactic nuclei (e.g. Seyfert galaxies and quasi-stellar objects, quasars and blazars), and the hopeful detection of gravitational wave signals. A significant fraction of all empirical astronomical studies concerns variable phenomena; see the review by Feigelson (1997) and the symposium *New Horizons in Time Domain Astronomy* (Griffin *et al.* 2012).

The nature of time series behavior in astronomical data exhibit a great range in properties. Orbital, rotational and pulsational behavior produces periodic variations in position, velocity and/or brightness. Binary-star orbits range from minutes to centuries, neutron star rotations range from milliseconds to minutes, and millions of harmonics have been detected in seismological studies of our Sun's oscillations. Accretion onto compact stars, stellar mass and supermassive black holes produces stochastic variability on time-scales of microseconds to decades. The statistical behavior of the emission is often not Gaussian white noise; $1/f$-type long-term memory processes and other patterns are common. Explosive phenomena produce a bewildering variety of temporal variations that defy simple classification. These include magnetic reconnection events on stellar surfaces, interstellar and interplanetary scintillation of radio sources, relativistic jets in blazars, microlensing

events when one star passes in front of another, supernova explosions at the end of a massive star's life, and thermonuclear explosions producing X-ray bursts, and gamma-ray bursts from the violent birth of black holes.

Some astronomical time series are multivariate because the phenomenon is measured (quasi)simultaneously in different wavelength bands. In situations like orbital eclipses or microlensing events, the variations should be simultaneous (though not necessarily with equal amplitude) in all bands. In situations like supernova light curves, blazar synchrotron jets variations, and gamma-ray burst afterglows, the variations may be correlated with some lag-time of astrophysical interest. Major time-domain survey projects such as Pan-STARRS and the Large Synoptic Survey Telescope are underway which will obtain essentially a "movie" of the sky at hundreds of epochs with billions of individual time series measured in several wavelength bands.

The statistical analysis of astronomical time series must confront difficulties associated with data acquisition, in addition to the intrinsic complexity of the emission processes.

1. The measurements of brightness or other variable properties very often are unevenly spaced. The schedule of ground-based visible light observations are regulated by weather, daily solar and monthly lunar cycles. Space-based observations are often regulated by satellite orbits. Many observations are limited by telescope allocation committees which often do not give sufficient access to acquire datasets optimized for later statistical analysis. While standard methods for evenly spaced time series can sometimes be adapted for occasional gaps in the data stream, they cannot be used for datasets that intrinsically consist of measurements at unevenly spaced times. Unfortunately, relatively little attention has been paid to such problems outside of the astronomical community; of the hundreds of volumes written on the analysis of temporal data, only one is specifically devoted to this issue (Parzen 1984). Common approaches are to project the data onto a regularly spaced grid either by binning or by interpolation. Various methods to reconstruct a smooth or evenly spaced time series from irregularly sampled data are discussed by Vio *et al.* (2000). Interpolation procedures may work adequately for smoothly varying behavior but are ineffective for objects exhibiting stochastic short-term autocorrelated variations or periodicities.
2. Individual observations are subject to heteroscedastic measurement errors. The variances of these errors are known from independent measurements of instrumental conditions (such as signal-to-noise ratios, blank sky measurements and instrumental calibrations). Thus, statistical analyses can, in principle, be weighted to account for differences in measurement errors. However, most time series procedures assume homoscedastic errors, and statistical treatments of heteroscedasticity with known variances have not been developed in a time series context.
3. The variations of astronomical sources, and background noise to be subtracted, often show correlation structure. Common deviations are $1/f$-type long-memory processes.

The time series methods described here may be useful for other types of astronomical datasets where one measures a signal along an independent ordered, time-like variable. This includes astronomical spectra where the intensity of light is measured as a function of its wavelength and astronomical images where the intensity is measured as a function of a

pixelated spatial location in an image. This connection is rarely made; an interesting example is the adaptation of Bayesian modeling procedures designed for change-point detection in a one-dimensional Poisson time series to source detection in a two-dimensional Poisson image (Scargle 2003).

11.2 Concepts of time series analysis

A broad goal of time series analysis is to represent behavior that may be exhibited over many points in a dataset in a model with a small number of parameters. No single procedure works for all situations. Datasets dominated by periodic behavior may be effectively modeled by Fourier transforms or related types of harmonic analysis. Datasets dominated by deterministic trends may be best modeled by regression. Datasets dominated by stochastic but autocorrelated behavior may be best modeled using autoregressive models. Composite models can be used for complicated time series with combinations of behaviors. In all of these cases, the temporal structure violates the assumption of independence (that is, the data are no longer i.i.d., Section 2.5.2) underlying many statistical methods discussed elsewhere in this volume.

Procedures for all these types of time series modeling are well-developed for evenly spaced data. A large academic library has hundreds of monographs and texts on these methodologies, some in mathematics and statistics, some addressing engineering signal processing, and others oriented towards econometrics. Astronomers tend to use a narrow suite of methods, particularly Fourier analysis, for a wide range of problems. They may also be unfamiliar with the many tools available for model selection and validation.

When the average value of a time series changes with time, one of several types of **trend** is present. A global linear trend has a formulation similar to linear regression treated in Section 7.2,

$$X_t = \alpha + \beta t + \epsilon_t. \tag{11.1}$$

The notation here is common in time series analysis and differs from the notation used elsewhere in the volume: X (or a vector of variables $\mathbf{X}$) is the response variable and time t is the independent variable. The notations X_t and ϵ_t are equivalent to $X(t)$ and $\epsilon(t)$.

More complex trends can be modeled with nonlinear functions or with smoothers and local regression procedures discussed in Chapter 6. Adapting language from signal processing, smoothing to reduce variance from short-term variations is a **low-pass filter**, while examining residuals after fitting a regression model to reduce long-term trends is a **high-pass filter**. Stochastic trends are often treated in the context of **autoregressive models**, such as ARMA models (Section 11.3.3). A **differencing filter**, such as

$$Y_t = X_t - X_{t-1}, \tag{11.2}$$

is commonly used to remove trends of various types (global, periodic or stochastic) to reveal short-term structure. Finally, these procedures can be used in combinations; for example,

a likelihood-based **state-space model** can be constructed to fit simultaneously long-term deterministic trends, periodicities, and short-term autocorrelated variations.

Autocorrelation is a basic measure correlated structure in a time series. With a calculation similar to the bivariate linear correlation coefficient presented in Section 3.6.2, the **autocorrelation function** establishes the degree to which values at different separations in time vary together,

$$ACF(k) = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2} \tag{11.3}$$

where $\bar{X}$ is the mean of the time series. The integer parameter $k > 0$ here is called the **lag time**, the numerator (properly normalized) is the sample autocovariance function, and the denominator is the sample variance. The plot of $ACF(k)$ for a range of k is often called the **correlogram**. Complicated behavior in the time series may appear as simple behavior in the ACF. Random uncorrelated noise, often called **white noise**, produces near-zero values in the ACF (except for the obvious value $ACF(0) = 1$). Time series with short-term autocorrelation have strong ACF coefficients at small k, while time series with **long-term memory** have ACF signals for a wide range of k. Periodic variations in the time series produce periodic variations in the ACF; these are most efficiently modeled using the Fourier transform. ACFs are not easily interpreted when both trend and stochastic variations are present, so they are often applied to **residual time series** after trend models have been fit and removed.

Stationarity is the property that the temporal behavior, whether stochastic or deterministic, is statistically unchanged by an arbitrary shift in time. An evenly spaced dataset is strictly stationary when the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_m})$ and $(X_{s+t_1}, X_{s+t_2}, \ldots, X_{s+t_m})$ is the same for any shift s of time. The concept of a mean value, where the sample mean converges to the population mean, is not relevant to many nonstationary processes. **Weakly stationary** phenomena have constant moments, such as the mean and autocovariance, but may change in other ways. Often the concept of stationarity is applied to the residuals after a model has been fitted; a successful model should produce stationary residuals with negligible autocovariance even if the original time series was nonstationary.

While astronomical systems may occasionally show gradual changes in variability, the clearest cases of **nonstationarity** are systems that change abruptly from one state to another. A remarkable example is the X-ray emission from the accreting Galatic X-ray binary star system GRS 1915+105. Here over a dozen distinct variability states – some dominated by noise, others by quasi-periodicities or explosions – are present (Fender & Belloni 2004). The instant when a time series changes variability characteristics is called a **change-point**.

When periodic phenomena are present, the signal becomes far more concentrated after transformation from the time domain to the frequency domain. This type of analysis has various names: **spectral analysis**, **harmonic analysis** (this term is sometimes restricted to cases where the period is known in advance) or **Fourier analysis** (although some spectral methods do not involve the Fourier transform). Frequency- and time-domain methods are often closely related; for example, the Fourier spectrum can be directly mapped onto the time-domain autocorrelation function, and a periodogram can sometimes be viewed as a partitioning of the variance related to analysis of variance (ANOVA).

Since all data points in a time series contribute to each value of a spectral density function, these methods are based on strong assumptions that the data points are equally weighted (homoscedastic errors) and the temporal behavior does not change during the observation, including changes in mean levels (stationarity). Fourier analysis has additional restrictive assumptions, such as evenly spaced data and sinusoidally shaped variations. Due to these limitations, the results of spectral analysis are often difficult to interpret; for example, the significance of a peak in a periodogram may not be readily computable if a trend is present, the data are not evenly spaced or autoregressive behavior is present.

Finally, we note that most of the methods discussed here assume the temporal behavior is mathematically linear with respect to previous values or ancillary variables. Many types of **nonlinear models** can be treated: time series with variable volatility and other types of heteroscedasticity, piecewise linear models, models with thresholds, autoregressive models with variable coefficients, dependences on high-order autocorrelation (e.g. nonzero bispectrum), or chaotic behavior (Fan & Yao 2003). **Chaos** is a particular form of nonlinear behavior where small changes in initial conditions can, under particular conditions, lead to large changes in outcomes. Chaotic time series have specific characteristics – strange attractors and other organized trajectories in phase space, nonzero Lyapunov exponents, period doubling – which are rarely seen in astronomical datasets (Scargle 1992). Only certain orbits in Solar System dynamics are convincingly chaotic. Nonlinear astronomical time series are more commonly dominated by stochastic behavior rather than deterministic chaos.

11.3 Time-domain analysis of evenly spaced data

11.3.1 Smoothing

In Chapter 6, we discussed a number of important data smoothing operations such as kernel density estimation which transform a series of discrete observations into a continuous function. These can readily be applied to time series data. We outline here a variety of other smoothing methods commonly used to reduce variance in evenly spaced time series.

Perhaps the simplest of these operations is the **central moving average (CMA)** with bandwidth of j time intervals. For even j,

$$\hat{X}_{i,CMA}(j) = \frac{1}{j+1} \sum_{k=-j/2}^{j/2} X_{i+k}. \tag{11.4}$$

Figure 11.3 below shows this and other smoothers applied to an astronomical dataset. This can be modified by averaging only past values rather than past-and-future values, using the robust median rather than the mean, and weighting each value by some predetermined factor such as the measurement error variance.

The **exponentially weighted moving average (EWMA)** is a common smoothing procedure for time series with short-term autocorrelaton

$$\hat{X}_{i,EWMA} = \alpha X_i + (1-\alpha) X_{i-1} \tag{11.5}$$

where $0 \leq \alpha \leq 1$ and $X_{1,EWMA} = X_1$. The estimator is called exponential because the current value is a weighted average of all previous values with weights decreasing exponentially with time. The value of α can be set by the user, or estimated by least squares to minimize $\sum(\hat{X}_{i,EWMA} - X_i)^2$. The EWMA is closely related to the random walk stochastic process (Section 11.3.3 below), and ARIMA models.

11.3.2 Autocorrelation and cross-correlation

The **autocorrelation function (ACF)** for evenly spaced data shown in Equation (11.3), a fundamental measure of the serial correlation in a time series, is a second-order statistic measuring the ratio of the variance with lag-time k to the sample covariance. Under the null hypothesis that the time series has no correlated structure and the population ACF is zero for all lags (except for ACF(0) which is always unity), the distribution of the sample is asymptotically normal with mean $-1/n$ and variance $1/n$; that is, the distribution of the null case ACF is $ACF(k) = N(-1/n, 1/n)$.

Plots of ACFs (**correlograms**) are frequently displayed with confidence intervals based on this normal approximation to test this null hypothesis that no autocorrelation is present. Comparing a sample ACF to its confidence band gives a simple test for randomness in the correlation structure of a time series. Some ACF patterns are readily interpretable though others are not. Strong values at low k rapidly decreasing at higher k imply a low-order autoregressive process is present (Section 11.3.3). Both monotonic trends and long-memory stochastic processes like $1/f$ noise (Section 11.9) produce slow decreases in ACF values from $ACF(0) = 1$ to large k. Data with a single characteristic time-scale for autocorrelation will produce a wave-like pattern of positive and negative values in the ACF even if phase coherence is not maintained. For strictly periodic sinusoidal variations, these variations follow a cosine curve.

Simple quantities like the sample mean are valid estimates of the population mean for stationary processes, but its uncertainty is not the standard value when autocorrelation is present,

$$\widehat{Var}(\bar{X}_n) = \frac{\sigma^2}{n}\left[1 + 2\sum_{k=1}^{n-1}(1 - k/n)ACF(k)\right]. \tag{11.6}$$

Qualitatively this can be understood as a decrease in the number of independent measurements due to the dependency of the process. This has a practical application: the comparison of means of two time series (either of different objects or the same object at different times) must take into account this increase in variance. Similar limitations apply to other statistical quantities estimated from autocorrelated time series. The mathematical theory of ergodicity provides the rules of convergence in such cases.

The **partial autocorrelation function (PACF)** at lag k gives the autocorrelation at value k removing the effects of correlations at shorter lags. The general PACF is somewhat complicated to express algebraically; see Wei (2006, Chapter 2) for a derivation. The value of the p-th coefficient $PACF(p)$ is found by successively fitting autoregressive models with order $1, 2, \ldots, p$ and setting the last coefficient of each model to the PACF parameter. For

example, a stationary AR(2) model with normal noise will have

$$PACF(2) = \frac{ACF(2) - ACF(1)^2}{1 - ACF(1)^2}. \tag{11.7}$$

The significance of PACF values can be assessed with respect to the normal approximation for the null hypothesis that no partial correlation is present at the chosen lag.

The PACF thus does not model the autocorrelation of the original time series, but gives insight into the time-scales responsible for the autocorrelated behavior. This can be very useful to the astronomer who often seeks to understand the drivers of a temporal phenomenon. In other fields, it is used as a diagnostic to help in model selection, as in choosing the order of an autoregressive model in the well-established **Box–Jenkins approach** to time series modeling. Astronomers interested in understanding the underlying behavior of an autocorrelated time series are encouraged to view both the ACF and PACF, as often one will have a simpler structure than the other. Figure 11.5 illustrates the ACF and PACF for an astronomical dataset.

Diagnostic **lag k scatter plots** sometimes reveal important structure in an autocorrelated time series (Percival & Walden 1993, Chapter 1). Here one plots all x_t values against x_{t+k} values where k is chosen by the scientist. A random scatter points toward uncorrelated noise, while a linear relationship can point towards stochastic autoregressive behavior. A circular distribution suggests a periodic signal, and clusters of points would suggest a nonstationary time series with distinct classes of variability behavior.

We finally mention a simple measure of autocorrelation in an evenly spaced time series, the **Durbin–Watson statistic** (Greene 2003, Chapter 19). Commonly applied to residuals to assist in econometric regression, the Durbin–Watson statistic is

$$d_{DW} = \frac{\sum_{i=2}^{n}(X_i - X_{i-1})^2}{\sum_{i=1}^{n} X_i^2} \tag{11.8}$$

which is approximately equal to $2(1 - R)$, where R equals the sample autocorrelation between successive residuals. The statistic ranges from 0 to 4; values around 2 indicate the absence of autocorrelation while values below (above) 2 indicate positive (negative) autocorrelation. For large samples, d_{DW} is asymptotically normal with mean 2 and variance $4/n$. For small samples, critical values of d_{DW} are tabulated but are applicable only to test the alternative hypothesis of an AR(1) process.

11.3.3 Stochastic autoregressive models

The simplest stochastic autocorrelated model is

$$X_i = X_{i-1} + \epsilon_i \tag{11.9}$$

where the noise term is usually taken to be a homoscedastic normal with zero mean and variance σ^2, $\epsilon_i \sim N(0, \sigma^2)$. This is similar to the classical **random walk** or **drunkard's walk**. In this case, recursive back-substitution shows that

$$X_i = X_0 + \epsilon_1 + \epsilon_2 + \cdots + \epsilon_i. \tag{11.10}$$

While the mean value remains zero, the ACF starts at unity and decreases slowly with lag-time k,

$$\begin{aligned} ACF_{RW}(i,k) &= \frac{Cov(X_i, X_{i+k})}{\sqrt{Var(X_i)Var(X_{i+k})}} && (11.11)\\ &= \frac{i\sigma^2}{\sqrt{i\sigma^2(i+k)\sigma^2}} && (11.12)\\ &= \frac{1}{\sqrt{1+k/i}}. && (11.13) \end{aligned}$$

As the differences between adjacent points in a random walk time series is white noise, the ACF of the first-order differences should have no significant values. This fact can be used to test the presence of structure beyond a random walk.

The obvious generalization of the random walk model and the EWMA smoothing procedure is to permit dependencies on more than one past value in the time series with different weights. This is the **autoregressive (AR) model**

$$X_i = \alpha_1 X_{i-1} + \alpha_2 X_{i-2} + \cdots + \alpha_p X_{i-p} + \epsilon_i. \tag{11.14}$$

The AR(1) model has simple analytical properties; the expected mean is zero and the autocorrelation function for lag k is now time-invariant as

$$ACF_{AR(1)}(k) = \alpha_1^k \tag{11.15}$$

for $|\alpha_1| < 1$. Here the ACF decays rapidly if α_1 is near zero, remains high if $\alpha \simeq 1$, and the $PACF(k)$ values average around zero above the lag $k = 1$. Figure 11.7 shows the spectral pattern produced by a high-order AR model fit to an astronomical dataset. Autoregressive models for Poisson distributed event counts have been developed that may be useful for certain astronomical problems.

In the **moving average (MA) model**, the current value of the time series depends on past values of the noise term rather than past values of the variable itself,

$$X_i = \epsilon_i + \beta_1 \epsilon_{i-1} + \cdots + \beta_q \epsilon_{i-q} \tag{11.16}$$

where $\epsilon_i = N(0, \sigma_i^2)$. A wide range of time series can be understood as a combination of AR and MA autoregressive processes. The **ARMA(*p*,*q*) model** is

$$X_i = \alpha_1 X_{i-1} + \cdots + \alpha_p X_{i-p} + \epsilon_t + \beta_1 \epsilon_{i-i} + \cdots + \beta_q X_{i-q}. \tag{11.17}$$

Simple cases have analytic properties; for example, assuming homoscedastic normal noise, the ARMA(1,1) model has variance and autocovariance

$$\begin{aligned} Var_{ARMA(1,1)}(x) &= \sigma^2 + \sigma^2 \frac{(\alpha+\beta)^2}{1-\alpha^2}\\ ACF_{ARMA(1,1)}(k) &= \frac{\alpha^{k-1}(\alpha+\beta)(1+\alpha/\beta)}{1+\alpha\beta+\beta^2}. && (11.18) \end{aligned}$$

The ACF and PACF can help identify what type of autoregressive behavior is present (Wei 2006, Chapter 6). For an ARMA(p,q) process, the ACF and PACF fall off exponentially

at large lag. For a monotonic trend, the ACF remains high for all lags. The ACF shows oscillating behavior when the time series behavior is periodic.

Once we have a model family of autocorrelation behavior, the model coefficients can be estimated by least-squares or maximum likelihood regression methods. Often MLEs are more accurate than least-squares estimators. Error analysis on ARMA parameters can be obtained by some sophisticated mathematical calculations or by bootstrap simulation. Model selection in time series analysis is often based on the **Akaike information criterion** (Percival & Walden 1993, Section 9.9). For a model with p parameters θ_p,

$$AIC(\theta_p) = -2\ln(L(\theta_p)) + 2p \tag{11.19}$$

where $L(\theta_p)$ is the likelihood for the model under consideration. For AR(p) models, the AIC reduces to the simple form

$$AIC(p) = N\ln(\hat{\sigma}_p^2) + 2p \tag{11.20}$$

where $\hat{\sigma}_p^2$ is the maximum likelihood estimator of the variance of the ϵ noise term.

The ARMA model is stationary and thus cannot treat time series with systematic trends, either (quasi)periodic or secular. The **autoregressive integrated moving average (ARIMA) model**, proposed by statisticians G. Box and G. Jenkins in 1970, combines ARMA and random walk stochastic processes to treat such problems. The ARIMA(p,d,q) model has p parameters giving the autoregressive dependence of the current value on past measured values, d parameters giving the number of differencing operations needed to account for drifts in mean values ($d = 1$ corresponds to the random walk), and q parameters giving the autoregressive dependence on past noise values.

Another form of nonstationarity is when the variance, rather than the local mean level, of the time series changes during the observation. Commonly called **volatility** in econometric contexts, mathematically it is heteroscedasticity, although with a different origin than the heteroscedasticity arising from measurement errors. Use of the **autoregressive conditional heteroscedastic (ARCH) model** is a major development to treat such problems. Here the variance is assumed to be a stochastic autoregressive process depending on previous values, as in the ARCH(1) model with

$$Var_{ARCH(1)}(\epsilon_i) = \alpha_0 + \alpha_1 Var(\epsilon_{i-1}). \tag{11.21}$$

The **generalized ARCH (GARCH) model** adds another type of nonlinear dependency on past noise levels. Note that these advanced models must be applied with care; for example, an ARIMA-type model might first be applied to treat long-term trends, followed by applying an ARCH-type model to the residuals for modeling volatility variations. Treatments of nonstationary autoregressive times series in astronomy (including change-points, CUSUM plots, ARCH and state-space models) are discussed by Koen & Lombard (1993). Research on incorporating heteroscedastic measurement errors into ARMA-type modeling has just begun (Tripodis & Buonaccorsi 2009).

It is important to realize that familiar statistical formulae may be very inaccurate for autoregressive processes. Consider, for example, the variance of a time series or the variance of its mean value (Section 3.4.4). For a stationary autocorrelated time series, these variances can be an order of magnitude larger than the values obtained from i.i.d. data.

Furthermore, the sample mean is not asymptotically normally distributed, and convergence to the asymptotic value is slow with a bias that the sample variance is systematically smaller than the true variance (Beran 1989, Chapter 1; Percival 1993).

11.3.4 Regression for deterministic models

The time series above are modeled by stochastic behavior which can be appropriate for astrophysical situations involving accretion, turbulence, rich stellar systems or other phenomena which are so physically complex that the temporal behavior cannot be deterministically predicted. But many astronomical time series – few-body Keplerian orbits, stellar rotation, supernova light curves – can be modeled as outcomes of functional relationships with time and/or covariate time series.

The statistical treatment of time series modeling can be formulated in a similar way to the regression problems discussed in Chapter 7,

$$y(t) = f(t) + g(\mathbf{x}(t)) + \epsilon(t) \tag{11.22}$$

where $y(t)$ is the observed time series, f represents some functional dependence representing trend, g represents some functional (often assumed linear) dependence on a vector of covariate time series, and ϵ represents a noise term. Unless f and g are both constant, the time series is nonstationary. The noise is often assumed to follow the normal distribution $N(0, \sigma^2)$; if σ varies with time, the time series is called heteroscedastic.

However, there is a critical difference between modeling temporal and other phenomena: time is not a random variable like mass or luminosity, and observations must be taken along a unidirectional time line. The residuals between a time series and its model are themselves a time series, and deviations between the model and data are often correlated in time. In such situations, error analysis emerging from least-squares or maximum likelihood estimation assuming i.i.d. variables will underestimate the true uncertainties of the model parameters. This occurs because, when residuals are autocorrelated, there are fewer effectively independent measurements than the number of data points.

After a model is fitted, it is critical to examine the residual time series for outliers and autocorrelation. The distribution of the **residual ACF** depends on the model in a nontrivial fashion. Useful statistics are generated from the sum of low orders of the residual ACF, such as the test statistic (Chatfield 2004, Section 4.7)

$$Q = n \sum_{k=1}^{K} ACF_{resid}(k)^2. \tag{11.23}$$

For an ARMA(p, q) model, Q is asymptotically distributed as χ^2 with $(K - p - q)$ degrees of freedom. The **Ljung–Box statistic** is similar to Q. Attention to correlated residuals can be extremely important in astronomy. van Leeuwen (2007) obtains a several-fold improvement in stellar parallactic distances from the Hipparcos astrometric satellite by modeling of the temporal behavior of the model residuals.

A strong advantage of formulating this regression problem from a likelihood approach is that the full capabilities of maximum likelihood estimation and Bayesian analysis become available (Chapters 3 and 7). In MLE, the likelihood can be maximized using the

EM algorithm and parameter uncertainties estimated using the Fisher information matrix. In Bayesian analysis, the parameter ranges and interdependencies can be evaluated using MCMC calculations, including marginalization over parameters of low scientific interest. The appropriate level of model parsimony and complexity of the model can be evaluated with model selection tools such as the Bayesian information criterion (BIC); time series analysts historically have used the closely related Akaike information criterion (AIC).

11.4 Time-domain analysis of unevenly spaced data

The vast majority of time series statistical methodology is designed for data measuring some variables at evenly spaced time intervals. Some limited methodology is available for datasets with intrinsically unevenly spaced observations (Parzen 1984). For example, theorems of convergence and asymptotic normality for moments, autocorrelations and regressions of a time series sampled with various observing sequences have been derived. But few general results emerge when the observing sequence depends on the data, or when either the time series or observing sequence is nonstationary (e.g. with trends or $1/f$-type noise). Econometricians have studied stochastic autoregressive models for irregular sampling, including cross-correlations between multiple time series that are evenly spaced but with different sampling rates (Engle & Russell 1998, Ghysels *et al.* 2007). Parametric modeling of unevenly spaced datasets in the time domain can be pursued within the flexible framework of likelihood-based state-space modeling discussed in Section 11.7.

We discuss here three statistical approaches for time-domain analysis of unevenly spaced data developed by astronomers over four decades. They are widely used, particularly to study the aperiodic variable emission from accreting massive black holes in quasars, BL Lac objects and other active galactic nuclei (e.g. Hufnagel & Bregman 1992).

11.4.1 Discrete correlation function

The discrete correlation function introduced by Edelson & Krolik (1988) is a procedure for computing the autocorrelation function that avoids interpolating the unevenly spaced dataset onto a regular grid. The method can treat both autocorrelation within one time series or cross-correlation between two unevenly spaced time series. Consider two datasets (x_i, t_{xi}) and (z_j, t_{zj}) with $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$ points respectively. For autocorrelation within a single dataset, let $z = x$. Construct two matrices, the **unbinned discrete correlation function (UDCF)** and its associated time lags,

$$\begin{aligned} UDCF_{ij} &= \frac{(x_i - \bar{x})(z_j - \bar{z})}{\sigma_x \sigma_z} \\ \Delta t_{ij} &= t_j - t_i \end{aligned} \tag{11.24}$$

where σ is the sample standard deviation of each dataset. The UDCF is then grouped into a univariate function of lag-time τ by collecting the $M(\tau)$ data pairs with lags falling within

the interval $\tau - \Delta\tau/2 \le \Delta t_{ij} < \tau + \Delta\tau/2$. The resulting **discrete correlation function** (DCF) and its variance are

$$DCF(\tau) = \frac{1}{M(\tau)} \sum_{k=1}^{M(\tau)} UDCF_{ij} \tag{11.25}$$

$$Var(\tau) = \frac{1}{(M(\tau)-1)^2} \sum_{k=1}^{M(\tau)} [UDCF_{ij} - DCF(\tau)]^2.$$

Edelson & Krolik provide advice, though without mathematical demonstration, for a number of situations. When homoscedastic measurement errors ϵ contribute significantly to the scatter, the denominator might be replaced by $\sqrt{(\sigma_x^2 - \epsilon_x^2)(\sigma_z^2 - \epsilon_z^2)}$. If a time series has many correlated observations within a small time interval, then the number of points contributed to $M(\tau)$ might be reduced to represent the number of uncorrelated UDCF values in the interval. The DCF can be calculated for negative τ as well as positive τ to reveal possible asymmetries in the correlated behavior. The DCF is undefined for any lag bin that has no associated data pairs.

A similar strategy of grouping lags between data pairs into evenly spaced bins is used in the **slot autocorrelation** estimator for the ACF developed for problems in experimental physics (Benedict *et al.* 2000). For a dataset x_i measured at times t_i, the ACF in a chosen range of lag-times $k\Delta\tau$ is

$$\widehat{ACF}_{slot}(k\Delta\tau) = \frac{\sum_{i=1}^{n} \sum_{j>i} x_i x_j b_k(t_j - t_i)}{\sum_{i=1}^{n} \sum_{j=1}^{n} b_k(t_j - t_i)} \quad \text{where} \tag{11.26}$$

$$b_k(t_j - t_i) = \begin{cases} 1 & \text{for } |(t_j - t_i)/\Delta\tau - k| < 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Here k are the chosen slots for which the ACF is computed, $\Delta\tau$ is the chosen width of the lag bins, and the Kronecker delta function b selects the data pairs with the appropriate lag-times to enter into the k-th slot. A similar procedure has also been developed for econometrics (Andersson 2007).

There has been some discussion of the accuracy and reliability of peaks in cross- or auto-correlation functions of unevenly spaced data. The centroid may give a more reliable estimate of a correlation lag than the peak, and significance levels can be estimated through resampling and other Monte Carlo procedures (Peterson *et al.* 1998).

In astronomical usage, the choice of bin width $\Delta\tau$ is generally not guided by mathematical considerations. The question is analogous to the choice of bin width in density estimation, as cross-validation to minimize the mean integrated square error (MISE) can be used (Section 6.4.2). Alexander (1997) recommends an equal number of data points contributing to each lag bin, rather than equal lag width. If at least ∼11 points are present in each bin, then the bin's distribution is approximately binomial and Fisher's z transform can be applied,

$$z(\tau) = \frac{1}{2} \ln \left(\frac{1 + DCF(\tau)}{1 - DCF(\tau)} \right). \tag{11.27}$$

The $z(\tau)$ values are then normally distributed with known mean and variance, after correction for a small known bias. Based on a few simulations, Alexander argues that $z(\tau)$ has greater sensitivity and smaller variance than the untransformed DCF.

11.4.2 Structure function

The **structure function** (sometimes called the **Kolmogorov structure function**) is a measure of autocorrelation originating in the field of stochastic processes and used in engineering signal processing, geospatial analysis and other applications. The structure function was introduced to astronomy by Simonetti *et al.* (1985) with application to unevenly spaced datasets.

The q-th order structure function is

$$D^q(\tau) = \langle |x(t) - x(t+\tau)|^q \rangle \tag{11.28}$$

where the angular brackets $\langle\ \rangle$ indicate an average over the time series and q is the order of the structure function (not necessarily integer). The $q = 2$ structure function is also called the **variogram**, an alternative to the autocorrelation function often used in spatial statistics (Section 12.4.1). The structure function is often more useful than the autocorrelation function when the time series is nonstationary when deterministic trends or stochastic ARIMA-type behavior are present.

When a dataset has a characteristic time-scale of variation τ_c, the structure function is small at shorter τ_c, rises rapidly to a high level around τ_c, and stays at this plateau for longer τ. The shape of the rise can roughly discriminate different noise or trend processes. If a structure function exhibits a power-law dependence on lag, $D^q(\tau) \propto \tau^\alpha$, the time series has a multi-fractal behavior. In this case, the Wiener–Khinchine theorem links the power-law dependence of the $q = 2$ structure function to the power-law dependence of the Fourier power spectrum. Study of the structure function, together with the closely related **singular measures**, can assist in differentiating between white noise, $1/f$ noise, fractional or standard Brownian motion, randomly located steps, smooth deterministic trends and intermittent processes such as turbulence (Davis *et al.* 1994).

The validity and accuracy of structure functions have recently been called into question. Nichols-Pagel *et al*. (2008) find that multi-taper spectral and wavelet estimators have smaller variances than structure function estimators when applied to simulations of turbulent physical systems. Emmanoulopoulos *et al.* (2010) argue from simulation studies that the structure function is unreliable for characterizing autocorrelated time series from unevenly spaced data. Structure function shapes can be very sensitive to the gap structure of the observations and can exhibit features that are not in the time series.

11.5 Spectral analysis of evenly spaced data

Whereas some temporal phenomena can be understood with parsimonious models in the time domain involving deterministic trends or stochastic autoregressive behavior, others

are dominated by periodic behavior that is most effectively modeled in the frequency domain. Harmonic analysis studies the variance of a time series as a function of frequency rather than time. Frequency-domain analysis is familiar to the astronomer because many astronomically variable phenomena involve orbits, rotations or pulsations that are exactly or nearly periodic. Time series analysis involving aperiodic phenomena, such as autocorrelated variations commonly seen in human affairs, are best treated in the time domain.

11.5.1 Fourier power spectrum

Classical spectral analysis is based on the Fourier transform and its associated power spectrum. **Fourier analysis** is designed for stationary time series with infinite duration and evenly spaced observations of sinusoidal periodic signals superposed on Gaussian (white) noise. Many astronomical time series deviate from these assumptions with limited-duration observations, gaps or unevenly spaced observations, nonsinusoidal periodic signals, phase incoherence, Poissonian or $1/f$ noise, or nonstationarity. Even when the assumptions of Fourier analysis are approximately met, many judgments must be made. We will briefly review some of the extensive methods in spectral analysis developed to treat these problems. The literature is vast: elementary presentations can be found in Warner (1998), Chatfield (2004) and Cowpertwait & Metcalfe (2009) while more advanced treatments include Priestley (1983) and Percival & Walden (1993).

A time series dominated by sinusoidal periodic variations can be modeled as

$$X_t = \sum_{k=1}^{n} A_i \cos(\omega_k t + \phi_k) + \epsilon_k \tag{11.29}$$

where the parameters A and ϕ are the amplitudes and phases ($0 \le \phi \le 2\pi$) at frequencies ω. This time series model is nonstationary except under special conditions. This function is mathematically defined for $n \to \infty$, but we consider the realistic situation where the data are discrete and finite.

While one could proceed in obtaining estimates of the parameters A_k using regression techniques, it is often more useful to consider the **power spectral density function**, or **power spectrum**, defined to be the Fourier transform of the autocovariance function (Section 11.3.2),

$$\begin{aligned} f(\omega) &= \frac{\sigma_X^2}{\pi} \sum_{k=-\infty}^{\infty} ACF e^{-i\omega k} \\ &= \frac{\sigma_X^2}{\pi} \left[1 + 2 \sum_{k=1}^{\infty} ACF(k) \cos(\omega k) \right] \end{aligned} \tag{11.30}$$

where σ_X^2 is the sample variance. Here $f(\omega)/\sigma_X^2$ can be interpreted as the fraction of the sample variance that is attributable to periodic variations in a narrow frequency interval around ω. The power spectrum can be written in a number of ways – involving trigonometric functions, exponentials and autocovariance functions – and with different normalizations. Note that the power spectrum and the autocorrelation function can be derived from one another and contain the same information.

The power spectrum can be characterized in closed form for some simple temporal models. For a white-noise process, the autocorrelation function is zero and the power spectrum $f(\omega) = \sigma_\epsilon^2/\pi$ is a constant for all frequencies. For a deterministic sinusoidal signal with variable phase, $X_t = \cos(\omega_0 + \phi)$ with ϕ a uniform random variable, the autocorrelation function is $ACF(k) = \cos(\omega_0 k)/(2\sigma_X^2)$ which is periodic and the power spectrum is infinite at ω_0 and zero at other frequencies. For an autoregressive AR(1) model, $x_t = \alpha X_{t-1} + \epsilon_t$, where $ACF(k) = \alpha^{|k|}$, the power spectrum is

$$f(\omega) = \frac{\sigma_\epsilon^2}{\pi[1 - 2\alpha\cos(\omega) + \alpha^2]}. \tag{11.31}$$

When $\alpha > 0$ the spectral density is strong at low frequencies and monotonically declines towards high frequencies. High-order AR or ARMA processes can give more complicated nonmonotonic spectral densities as illustrated in Figure 11.7.

For nontrivial scientific problems, the power spectrum cannot be derived analytically and must be estimated from the time series measurements x_t at evenly spaced increments Δt. This process is often called **spectral analysis**. A common model formulation is the finite Fourier series

$$X_t = a_0 + \sum_{p=1}^{n/2-1}\left[a_p\cos\left(\frac{2\pi pt}{n}\right) + b_p\sin\left(\frac{2\pi pt}{n}\right)\right] + a_{n/2}\cos(\pi t). \tag{11.32}$$

The least-squares parameter expected values are

$$\begin{aligned} a_0 &= \bar{X} = \frac{1}{n}\sum_{t-1}^{n} X_i \\ a_p &= 2\left[\sum X_t\cos(2\pi pt/n)\right]/n \\ b_p &= 2\left[\sum X_t\sin(2\pi pt/n)\right]/n \\ a_{n/2} &= \sum(-1)^t X_t/n \end{aligned} \tag{11.33}$$

for the p harmonics $p = 1, 2, \ldots, n/2 - 1$. Spectra are calculated at increments of $2\pi/n, 4\pi/n, \ldots, \pi$ between the low fundamental frequency $2\pi/(n\Delta t)$ and the high Nyquist frequency $\pi/\Delta t$. The spectrum is often plotted as a histogram of the quantity $I(\omega)$

$$I(\omega_p)_S = n\sqrt{a_p^2 + b_p^2}/4\pi \tag{11.34}$$

against frequency ω. This is the **periodogram** developed by A. Schuster in 1898 and very widely used since. An example of a Schuster periodogram for a variable astronomical source is shown in Figure 11.4.

Some astronomical objects, particularly pulsating stars and multi-planetary systems, exhibit multiple periods. In such cases, it is common to seek the dominant periodicity from a periodogram, remove that sinusoid from the time series, seek the next periodicity from the periodogram of the residuals, and iterate until no significant spectral peaks are present. This method is known in signal processing as **iterative pre-whitening**. In astronomy, it is a widely used procedure, often following the **CLEAN algorithm** originally designed for deconvolution in interferometric imaging (Roberts *et al.* 1987).

11.5.2 Improving the periodogram

Schuster's periodogram emerges directly from Fourier analysis, but has statistical limitations. It exhibits a high level of noise, even when the time series has high signal-to-noise, that does not decrease as the observation length increases. In statistical language, it is a biased estimator of the power spectrum $f(\omega)$ in the sense that its variance does not approach zero as n approaches infinity. As mentioned above, the analysis also makes various assumptions (such as infinitely long and evenly spaced data with stationary sinusoidal signals and Gaussian noise) which may not hold for a real astronomical dataset. For example, **spectral leakage** into nearby frequencies occurs if the time series is relatively short with respect to the periods under study and is not an integer multiple of the cycle length. These problems can be considerably alleviated with spectral analysis techniques including detrending, smoothing and tapering.

For nonstationary signals which vary in amplitude in some systematic way, we can **detrend** the dataset with regression or nonparametric smoothers to create a quasi-stationary time series for further harmonic analysis. This reduces strong spectral amplitudes at low frequencies and improves noise throughout the spectrum. High-pass filtering, low-pass filtering and iterative pre-whitening can also be useful to reduce undesired signals or noise in a spectrum.

Smoothing introduces bias but reduces variance in the spectral density estimator. Smoothing can be performed in the frequency domain or in the time domain applied to the original data or the autocorrelation function. A simple smoothing function proposed by Daniell is a moving average which convolves the time series with equal weights across a window of width m. It produces a spectral window with significant side-lobes. Bartlett's window convolves the data with a triangular rather than rectangular kernel and produces smaller side-lobes. Other common choices, such as the Tukey–Hanning, Parzen or Papoulis windows, smoothly decrease the weight with distance from the central frequency. As with all data smoothing operations (Section 6.4.1), a wider window (larger m) reduces variance but increases bias, particularly blurring signals at closely spaced frequencies. Strategies for selecting a window shape and width m are discussed by Percival & Walden (1993, Chapter 6); one seeks to balance the reduction of smoothing window leakage, avoid the introduction of spectral ripples, and maintain sufficient spectral resolution and dynamic range. Figure 11.4 illustrates the effects of two levels of smoothing on an astronomical periodogram.

Tapering, a convolution which reduces the spectral side-lobes and leakage arising from ringing off of the edges of the dataset, treats the limitation that the data do not have infinite duration. The effect of tapering is to reduce the sample size towards the edge of the dataset. Opposite to smoothing, tapering decreases bias due to spectral leakage, but introduces variance as the effective data size is reduced. Here, the time series x_i is replaced by the product $h_i x_i$ where h is a taper function. A large literature exists on the choice of the taper bandwidth and functional form. One common choice is a cosine taper where h_{cos} is applied to the first and last p% of the dataset,

$$h_{j,cos} = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi j}{|pn|+1}\right)\right]. \tag{11.35}$$

The 100% cosine taper which affects all data-points is called the Hanning taper. The average of different orthogonal tapers, or **multi-tapering**, is an important procedure permitting, for example, quantitative tradeoff between bias and variance. The discrete prolate spheroidal sequences or Slepian sequences are a favored taper sequence for this purpose.

11.6 Spectral analysis of unevenly spaced data

Most of the methodological work to treat spectral analysis when the observation times are intrinsically unevenly spaced has emerged from astronomy, as these situations do not often occur in common engineering or econometric applications.

Brillinger (1984) gives a broad, mathematical discussion of the spectral analysis of unevenly sampled time series showing that the power spectrum of the observed process will depend both on the underlying process of interest and on its convolution with the observational process. The estimation of the spectrum of the underlying process is a tractable inverse problem if both the observed and observational patterns are stationary random processes. However, the astronomical observational constraints often depend on various extraneous cycles (solar, lunar, satellite orbits), autocorrelated processes (cloud cover) and the vagaries of telescope allocation committees (Section 11.1). The astronomical observing times thus often cannot be considered a stationary random process.

11.6.1 Lomb–Scargle periodogram

Classical Fourier analysis requires time series measurements acquired in evenly spaced intervals, and such datasets are common in many fields of engineering and social science. The uneven spacings common in astronomical time series remove this mathematical underpinning of classical Fourier calculations. Deeming (1975) gives an important and influential discussion of these effects.

A major step was provided by the periodogram developed by N. Lomb (1976) an J. Scargle (1982) which generalizes the Schuster periodogram for unevenly spaced data. The **Lomb–Scargle periodogram (LSP)** can be formulated either as a modified Fourier analysis or as a least-squares regression of the dataset to sine waves with a range of frequencies. Its common expression is

$$P_{LS}(\nu) = \frac{1}{2\sigma^2}\left[\frac{[\sum_{i=1}^{n} X_i \cos(2\pi\nu t_i)]^2}{\sum_{i=1}^{n} \cos^2 2\pi\nu(t_i - \tau(\nu))} + \frac{[\sum_{i=1}^{n} X_i \sin(2\pi\nu t_i)]^2}{\sum_{i=1}^{n} \sin^2 2\pi\nu(t_i - \tau(\nu))}\right] \tag{11.36}$$

where σ^2 is the sample variance of the x_i and the parameter τ is defined by

$$\tan(4\pi\nu\tau) = \frac{\sum_{i=1}^{n} \sin(4\pi\nu t_i)}{\sum_{i=1}^{n} \cos(4\pi\nu t_i)}. \tag{11.37}$$

Press & Rybicki (1989) give an efficient computational algorithm for the LSP based on the fast Fourier transform.

Bretthorst (2003) places the LSP on a broader statistical foundation, showing that it is a solution to Bayes' theorem under reasonable conditions. He constructs a model of sinusoidal variations with time-variable amplitudes at different frequencies superposed by Gaussian noise. Uniform priors are assigned to the range of frequencies and amplitudes, and a Jefferys' prior to the variance. Solving Bayes' theorem analytically gives a generalized periodogram that reduces to the Schuster periodogram for uniformly sample data and to the LSP for arbitrary data spacings. The formalism can also treat nonsinusoidal shapes and multiple periodic signals. All of these are sufficient statistics, and marginal posterior probabilities for each parameter can be computed using Markov chain Monte Carlo procedures.

A number of recent studies have addressed methodological issues concerning the LSP. Reegen (2007) derives the LSP from a modified discrete Fourier transform with rotation in Fourier space, emphasizing methods for estimating reliable false alarm probabilities. Zechmeister & Kürster (2009) propose a generalized LSP based on weighted regression in the time domain with a floating mean, emphasizing application to nonsinusoidal behavior typical of exoplanets with eccentric orbits. Vio *et al.* (2010) rederive the LSP with a matrix algebraic formalism, allowing tests for periodic signals in the presence of nonstationary noise. Baluev (2008) and Sturrock & Scargle (2009) give further strategies for reliable assessment of peaks in the LSP, the latter based on Bayesian inference. Townsend (2010) presents an algorithm for efficient calculation of the LSP on graphics processing units, and another computational approach is proposed by Palmer (2009).

While the LSP is popular in the astronomical community for spectral analysis of unevenly spaced datasets, other approaches have been developed for applications in engineering and physics. Marquardt & Acuff (1984) present a **direct quadratic spectrum estimator (DQSE)**

$$Z_{DQSE}(\omega) = \frac{1}{T}\sum_{i=1}^{m}\sum_{j>i} D(t_j - t_i)F(t_i, t_j)\cos(2\pi\omega(t_j - t_i))x(t_i)x(t_j) \qquad (11.38)$$

where T is the observation duration, D is the Tukey–Hanning lag window, and F is a data spacing factor: $F = 1$ for equally spaced data and $F = 1/\lambda^2$ for random Poisson sampling with intensity λ. For arbitrary data spacings,

$$F(t_i, t_j) = \frac{1}{2}(t_{i+1} - t_{i-1})(t_{j+1} - t_{j-1}). \qquad (11.39)$$

A limit on F is introduced when the data spacing is very large.

The laboratory study of fluid turbulence using a laser Doppler anemometer gives velocity measurements at erratic times when tracer particles flow through the measurement device. This gives an unevenly sampled time series of a strongly autocorrelated, $1/f$-type (long-memory) process. This problem has led to the independent development of estimators of autocorrelation functions and power spectra (Benedict *et al.* 2000). The power spectral density is then derived from the **slot autocorrelation function** (Section 11.3.2) with smoothing or windowing to reduce the variance. The calculation can be computationally expensive.

Benedict *et al.* (2000) compare the performance of several spectral estimators for simulated datasets with $1/f$-type noise. They find that power spectra based on the slot correlation estimator with certain sophisticated smoothing options have excellent performance. The DQSE has low bias but shows high variance while the Lomb–Scargle estimator systematically overestimates the power at high frequencies. Rivoira & Fleury (2004) present an iterative procedure for estimating the number of lags and the window function that smooths the ACF in a fashion that balances bias and variance to minimize the mean integrated square error (MISE, Section 6.4.1). Their estimators give a lower MISE than the smoothed slot estimator and their computation is more efficient.

11.6.2 Non-Fourier periodograms

Astronomers have developed and extensively used a variety of methods for periodicity searches in unevenly spaced datasets and nonsinusoidal signals where Fourier methods have difficulties. The most common strategy involves **folding the data modulo a trial period**, computing a statistic on the folded time series (now a function of phase rather than time), and plotting the statistic for all independent frequencies. These methods have many advantages. They measure the strength of signals that are strictly periodic, but not necessarily sinusoidal in shape. This is particularly appropriate for variations due to orbital or rotational eclipses where the periodic signal often has a characteristic flat-bottom **U**-shape that can occupy an arbitrary phase fraction of each period. They are also relatively insensitive to the duration and uneven spacing of the dataset, and some methods readily permit heteroscedastic weighting from measurement errors. As with any harmonic analysis, however, adjudicating the significance of a peak in the periodogram can be difficult, particularly as the observation spacings interact with a periodic or autocorrelated signal to producing spurious aliased structure in the periodograms.

The simplest statistic is the **minimum string length** of the dataset folded modulo a trial period (Dworetsky 1983). This is the sum of the length of lines connecting values of the time series as the phase runs from zero to 2π. When an incorrect period is chosen, the signal is scattered among many phases and the line length is large, while for the correct period, the signal is collected into a small phase range and the line length is reduced. As distances rather than squared distances are considered, this is an L_1 statistic that is robust against non-Gaussianity and outliers (Chapter 5). This statistic is similar to two commonly used procedures in statistics: least absolute deviation (LAD) regression (Section 7.3.4) and the first element of the autocorrelation function of residuals from a regression function (Box & Pierce 1970). But in these cases, the line lengths are computed vertically for each individual point rather than diagonally between adjacent points.

Three related methods are based on least-squares (L_2) statistics of the time series data folded modulo trial periods, but then grouped into bins. They are very widely used in variable-star research. The **Lafler & Kinman (1965) statistic** for each trial period is

$$\theta_{LK}^2 = \frac{\sum_{j=1}^{m}(\bar{\phi}_j - \bar{\phi}_{j+1})^2}{\sum_{j=1}^{m}(\phi_j - \bar{\phi})^2} \tag{11.40}$$

where $\bar{\phi}_j$ the average value within the j-th phase bin, $\bar{\phi}$ is the global mean for the full sample, and m is the number of phase bins. In econometrics, this statistic (with slightly

different normalization) is known as the Durbin–Watson statistic commonly used for analysis of correlation in time series. J. von Neumann (1941) calculated its distribution for normally distributed variables.

The **Stellingwerf (1978) statistic** is the ratio of the sum of inter-bin variances to the sample variance,

$$\theta_S^2 = \frac{\sum_{j=1}^{m} \sum_{j=1}^{n_j} (\phi_{ij} - \bar{\phi}_j)^2}{(n-m) \sum_{i=1}^{n} (\phi_{ij} - \bar{\phi})^2} \tag{11.41}$$

where n_j is the number of data points in the j-th phase bin. The mean values can be weighted by heteroscedastic measurement errors. This is known as the **phase dispersion minimization** periodicity search method.

The **ANOVA (analysis of variance) statistic** presented by Schwarzenberg-Czerny (1989) is the ratio of the sum of inter-bin variances to the sum of intra-bin variances,

$$\theta_{AoV}^2 = \frac{(n-m) \sum_{j-1}^{m} n_j (\bar{\phi}_j - \bar{\phi}_{ij})^2}{(m-1) \sum_{j=1}^{m} \sum_{j=1}^{n_j} (\phi_{ij} - \bar{\phi}_j)^2}. \tag{11.42}$$

This ANOVA statistic (assuming normal homoscedastic errors and sufficiently large n_j values) follows an F distribution with $(m-1)$ and $(n-m)$ degrees of freedom,

$$\begin{aligned} E[\theta_{AoV}^2] &= \frac{n-m}{n-r-2} \\ Var[\theta_{AoV}^2] &= \frac{2(n-m)^2(n-3)}{(m-1)(n-m-2)^2(n-m-4)}. \end{aligned} \tag{11.43}$$

The ANOVA statistic may be preferable over the Stellingwerf statistic because the numerator and denominator are independent of each other. Note that the ANOVA statistic takes no cognizance of the shape of the variation in the binned phase time series, while the Lafler–Kinman (or Durbin–Watson) statistic is sensitive to smooth variations across the binned phase time series. They are not equivalent, and both can be used to give periodograms sensitive to different characteristics of the putative periodicity.

As with any histogram, there is no mathematical guidance regarding the choice of phase origin or bin boundaries; evenly spaced bins are commonly adopted, but this is not required (Section 6.3). The emerging periodogram may be sensitive to these choices. This limitation is addressed by the **Gregory-Loredo Bayesian periodogram** designed for nonsinusoidal periodic variations (Gregory & Loredo 1992). After folding the data modulo trial periods, an ensemble of grouping options is considered with different bin zero-points and widths. A likelihood is constructed including arbitrary levels for each bin level. The Bayesian odds ratio for periodic vs. constant emission is calculated in a high-dimensional parameter space, and the results are marginalized over uninteresting binning options to maximize the significance of the periodicity.

Schwarzerberg-Czerny (1999) has compared the performances of these period-finding procedures for a variety of simulated datasets. No single method outperforms others for all situations. It is important, if possible, to match the method to the complexity of the model. For periodic behavior with fine-scale structure, methods with narrow phase bins (or the unbinned string length statistic) are most sensitive while for broad structures that fill the

phased time series, Lomb–Scargle or ANOVA procedures are better. Gregory & Loredo's method that accounts for the model structure is an attractive approach.

The search for exoplanets is an important new application of periodicity searching and has triggered a flood of methodological studies. Here, a planet periodically passes in front of a star and produces a partial eclipse seen as a slight (typically 0.1–0.01%) diminution in the stellar brightness. The variation has a characteristic U-shape with a flat bottom. Data acquisition may be either evenly or unevenly spaced, and the data weightings may be either homo- or heteroscedastic. Proposed techniques include a least-squares **box-fitting algorithm** (Kovács *et al.* 2002) and a simplified version of the Gregory–Loredo approach using a χ^2-type statistic (Aigrain & Irwin 2004).

The problem is particularly challenging because the signal is very weak, and uninteresting sources of variability (instrumental effects, night-time conditions and stellar magnetic activity) may dominate. These effects can have $1/f$-type characteristics which can be very difficult to treat (Section 11.9). Approaches for treating these problems include principal components analysis (Tamuz *et al.* 2005), a gauge filter from image processing (Guis & Barge 2005), detrending and low-pass filtering (Renner *et al.* 2008), and application of covariance-corrected significance levels (Pont *et al.* 2006).

Statistician P. Hall (2008) has discussed some mathematical issues of periodicity searches using epoch-folded time series. He suggests, for example, that each folded time series be smoothed with the Nadaraya–Watson kernel or another local regression technique (Section 6.6.1). Another approach involves regression with sinusoidal functions; this can treat multiple and nonstationary periodicities.

11.6.3 Statistical significance of periodogram peaks

The significance of a single peak in a Schuster periodogram from classical Fourier analysis can be formally calculated using a χ^2 distribution assuming homoscedastic normal (white) noise with known variance, no periodic signal, and an infinite, evenly spaced dataset. However, the calculation is quite different if the variance is not known in advance but must be estimated from the data; this is the case seen in astronomical data. It is also often difficult to establish the degrees of freedom of the problem when the periodogram has been modified by smoothing and/or tapering. The significance level is also incorrect if the time series has any trend or autocorrelation, or if multiple periodicities are present. Even if conditions are appropriate for this significance test, the probability applies to a single frequency and not to the entire spectrum, so that the significance must be corrected by a large factor based on the false alarm rate of testing many potential periods. The difficulties of evaluating the significance of spectral peaks in Fourier analysis are discussed by Percival & Walden (1993).

For the Lomb–Scargle periodogram with unevenly spaced data, Horne & Baliunas (1986) argue that the significance of a single peak can be expressed as the sum of gamma functions which simplify to an exponential distribution with density

$$f(z) = \frac{1}{\sigma^2} e^{-z/\sigma^2}, \tag{11.44}$$

where σ is the total variance of the data including any signals. When searching for the existence of an unknown periodicity over N_p independent periods, the probability that some

spectral peak is higher than z_0 is

$$P(Z > z_0) = 1 - (1 - e^{-z_0})^{N_p}. \tag{11.45}$$

Although this result is widely used in astronomy to evaluate the significance of features in the LSP, it has a number of difficulties. First, like the exponential distribution for the standard Fourier periodogram, it assumes that the time series consists only of a single periodic signal in homoscedastic white (Gaussian) noise, $\epsilon \sim N(0, \sigma_0^2)$, and that σ_0 is known in advance. But if the variance is estimated from the data, which is almost always the case, the F distribution must be used rather than the exponential distribution, and the resulting probabilities are substantially different (Koen 1990). The exponential distribution of the LSP also requires that $\sum_{j=1}^{n} \cos(wt_j) \sin(wt_j) = 0$, a condition that depends on the spacing of the observation times. If autocorrelated and quasi-periodic behavior, either noise or astrophysical in origin, is present, there is no analytic estimator for the statistical significance of a peak in the Lomb–Scargle periodogram (or any other periodogram for unevenly spaced data, Section 11.6 below). The problem of significance estimation in Fourier spectra is exacerbated here because the signals can interact nonlinearly with gaps in the unevenly spaced datasets, giving spurious peaks and strong aliasing.

As analytical approaches are ineffective under these circumstances, numerical methods are essential for investigating the significance of a periodogram peak. First, periodograms of permutations of the data among the fixed observation times can be examined to establish the probability distribution of the periodogram in the presence of uncorrelated noise. Second, periodic test signals can be superposed on permuted datasets to study any peak frequency displacements, broadening and alias structures. Third, autocorrelated but aperiodic noise can be superposed on the permuted datasets to examine whether spurious spectral peaks readily appear. Reegen (2007) and others recommend other procedures for estimating LSP significance levels (Section 11.6.1).

11.6.4 Spectral analysis of event data

The methods described so far consider time series of the form (x_i, t_i) where a real-valued intensity is measured at unevenly spaced times. A related problem arises in high-energy astrophysics – X-ray, gamma-ray and neutrino astronomy – where individual events are detected. If the source is emitting at a constant flux, then the arrival times follow a Poisson distribution with an exponential distribution of intervals between adjacent events. A Poissonian noise component may also be present. X-ray and gamma-ray sources are often variable and, particularly if they arise from isolated or accreting rotating neutron stars, will exhibit periodicities. Mathematically, these would be classified as **nonstationary Poisson processes** with periodic components. Some methods for spectral analysis have been developed to treat this problem.

The most commonly used periodogram for event data is based on the Rayleigh probability distribution function (Leahy *et al.* 1983, Protheroe 1987),

$$f(x, \sigma) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)} \tag{11.46}$$

which can be generated from the square root of the quadratic sum of two normally distributed random variables. This statistic can be used as a test for uniformity in angle around a circle.

It is a special case ($m = 1$) of the $\boldsymbol{Z_m^2}$ **statistic** which can be written

$$Z_m^2 = \frac{2}{n}\sum_{k=1}^{m}\left[\left(\sum_{i=1}^{n}\sin(2\pi k\phi_i)\right)^2 + \left(\sum_{i=1}^{n}\cos(2\pi k\phi_i)\right)^2\right] \tag{11.47}$$

where ϕ_i are the data points in the interval 0–2π after folding with a trial period. The Rayleigh test is sensitive to a broad, sinusoidal signal while higher orders of Z_m^2 are sensitive to periodic structures with smaller duty cycles. Both are preferable to the traditional use of Pearson's χ^2 statistic that requires grouping of the folded events into bins of arbitrary origin and width. At high count rates, the periodograms based on the Rayleigh statistic approach those based on the Fourier series. The use of these statistics for periodicity searches with event data in gamma-ray astronomy is reviewed by Buccheri (1992).

The ***H* test** based on the Z_m^2 statistic was developed by de Jager *et al.* (1989) to find the optimal order m that minimizes the mean integrated square error (MISE; see Section 6.4.1) for the given dataset. This criterion gives

$$H = \max(Z_m^2 - 4m - 4) \tag{11.48}$$

where the maximization is calculated over choices of m. The H test has excellent performance for both narrow and broad pulses in simulated data. It is commonly used today by researchers searching for high-energy pulsars using ground-based (e.g. MAGIC, HESS, HEGRA) and space-based (e.g. AGILE, Fermi) gamma-ray telescopes.

A periodogram based on the **Kuiper statistic** has occasionally been used instead of the Rayleigh statistic on event data folded with trial periods (Paltani 2004). This statistic is similar to the Kolmogorov–Smirnov statistic, measuring the supremum distance between the empirical distribution function of folded event arrival times and a model based on the null hypothesis of no variation. The Kuiper test has better performance on simulated data with narrow pulses superposed on noise, while the Rayleigh statistic shows higher sensitivity on simulated data with broad pulses.

Swanepoel *et al.* (1996) propose a periodogram for event data based on a simple non-parametric statistic applied to epoch folded event data. The statistic is the minimum phase interval that contains a specified number of events. The statistic is unbiased and the bootstrap is used to estimate confidence levels.

11.6.5 Computational issues

The range of frequencies for calculating a spectrum is not obvious for unevenly spaced or event data. The low-frequency cutoff can be scaled inversely to the observation length so that a few periods will be present. The high-frequency cutoff for evenly spaced data is the **Nyquist frequency** $\nu_N = 1/(2\Delta t)$ where Δt is some time interval between observations. Periodic structure in the time series at higher frequencies will be aliased into the calculated spectrum. However, the Nyquist frequency is defined for unevenly spaced data, and the estimated spectrum has reduced sensitivity to periodicities at some range of high frequencies.

When data spacings are very uneven and the time series duration is long, the folded phase diagrams can differ considerably for even very small changes of trial period. It thus may

be necessary to compute the power spectrum, using the Lomb–Scargle or other method, at many frequencies above the Nyquist frequency limit appropriate for evenly spaced data (Pelt 2009).

It is well-known that a completely random sequence of observing times removes spurious aliased spectral peaks when the period of the variation is near the observation interval Δt (Shapiro & Silverman 1960). But even small deviations from random sequences can introduce aliases. A particularly insidious problem arises when the underlying time series with strong autocorrelation, but no strict periodicity, is observed at nonrandom unevenly spaced observation times. Then large, spurious peaks in the spectrum can appear, giving the illusion of strong periodic behavior. The possible presence of spurious peaks can be tested with permutations and simulations of the dataset, but a definitive conclusion requires new observations to test predictions. The problem of reliable inference of periodicity from unevenly spaced time series has beset the field of variable and binary-star research for decades (e.g. Morbey & Griffin 1987). NASA's Kepler mission is now providing denser and more evenly spaced light curves for ∼150,000 stars (Appendix C.15).

11.7 State-space modeling and the Kalman filter

Regression and stochastic parametric models (Sections 11.3.3 and 11.3.4) are often solved by likelihood methods. A general likelihood-based formulation of time series models called **state-space modeling** was developed in the mid-twentieth century for engineering applications where signals, often from complex dynamic mechanical systems, must be monitored for changes and adjustment of the system. The **Kalman filter** is an efficient algorithm for updating the likelihood as new data arise; here the term "filter" has meaning closer to "smoothing" and "forecasting" than the alterations of time series like a "high-pass filter".

We model the development of a multivariate time series $\mathbf{X}_t$ in two equations, somewhat analogously to the errors-in-variable models discussed in Section 7.5:

$$\begin{aligned} \mathbf{X}_t &= \mathbf{h}_t \mathbf{S}_\mathbf{t}(\theta) + \epsilon_t \\ \mathbf{S}_\mathbf{t}(\theta) &= \mathbf{G}_t \mathbf{S}_{t-1}(\theta) + \eta_t . \end{aligned} \tag{11.49}$$

In the measurement equation, $\mathbf{X}$ is a vector of observed values of a collection of variables at time t. $\mathbf{S}$ is the **state vector** that depends on the parameters θ, and $\mathbf{h}$ is the matrix that maps the true state space into the observed space. In the state equation, $\mathbf{G}$ is the **transition matrix** that defines how the system changes in time. The ϵ and η terms are uncorrelated and stationary normal errors. The system has a short-term memory.

A very simple state-space model would be a univariate system with both h and G as constant. This reproduces the random walk model of Section 11.3.3. A more complex nonstationary ARIMA-type linear growth model can be written as

$$\begin{aligned} X_t &= Y_t + \epsilon_t \\ Y_t &= Y_{t-1} + \beta_{t-1} + \eta_t \\ \beta_t &= \beta_{t-1} + \zeta_t \end{aligned} \tag{11.50}$$

where β represents the random walk component. A linear regression model relating X to some covariate Z commonly written as $X_t = \alpha_t + \beta_t Z_t + \epsilon_t$ can be formulated as the state-space model (11.49) where θ is the two-element vector $\theta^T = [\alpha_t, \beta_t]$, $\mathbf{h}_t^T = [1, z_t]$, $\mathbf{G}_{11} = \mathbf{G}_{22} = 1$, and η with zero variance.

The advantage of a state-space formulation for time series modeling is its capability of treating complex models with multiple input time series, linear and nonlinear deterministic dependencies, periodicities, autoregressive stochastic processes, known and unknown noise terms with heteroscedasticity, and so forth. The goals of the calculation might be interpolation and extrapolation of observed time series, estimating confidence bands around time series, detecting change-points in system behavior, or estimating the θ parameters to understand the underlying dynamical process.

The Kalman filter is a recursive procedure for updating the state vector $\mathbf{S}_{t-1}$ by incorporating the new data $\mathbf{X}_t$ under a least-squares or maximum likelihood criterion. It uses the current estimate of a fixed transition matrix $\mathbf{G}$ and the variance–covariance matrix of $\mathbf{S}$ to construct a gain matrix for correcting $\mathbf{S}_{t-1}$ to $\mathbf{S}_t$. In a least-squares context, the update equations were easily computable even before computers as they are linear and do not depend on past states of the system. Today, MLE and Bayesian approaches are commonly applied using numerical techniques such as the EM algorithm and Markov chain Monte Carlo. The Kalman filter is a simple example of **dynamic Bayesian networks**, and Equations (11.49) can be reformulated in terms of **hidden Markov models**. State-space models can be recovered from sparse or noisy data using the Kalman filter with techniques from **compressive sensing**.

Methodology for state-space modeling has been developed for 50 years and has great flexibility in treating sophisticated problems (Durbin & Koopman 2001). The methods can either identify or incorporate changes in behavior (nonstationarity). Hierarchical models can link variables observed at the telescope with intrinsic state variables determined by astrophysical theory. State-space methods can treat models which are nonlinear and noise which is non-Gaussian (including Poissonian noise). Bayesian methods can give parameter posterior distribution functions using importance sampling techniques. Model goodness-of-fit can be evaluated using the Akaike and Bayesian information criteria.

The volume by Kitagawa & Gersch (1996) illustrates the power of state-space methods for problems encountered in geosciences and astronomy such as seismological signals of earthquakes, annual temperature variations, Poissonian measurements of rapid X-ray variations in black-hole binary systems, and quasi-periodic variations in sunspot number. They adopt a quasi-Bayesian approach to state-space analysis involving **smoothing priors**. Several of their analysis procedures may be useful in astronomy. In particular, they formulate a likelihood for a state-space model of a continuous process observed at unevenly spaced intervals (Kitagawa & Gersch 1996, Sections 5.3 and 7.3). The likelihood can be used either for estimation of model parameters, or for smoothing the dataset into a continuous curve.

State-space models are rarely used in astronomy today, but may be valuable for complicated problems with dense evenly spaced data streams such as X-ray timing studies of accreting black holes (Vio *et al.* 2005). Consider, for example, the search for gravitational wave signals from a laser interferometric observatory like LIGO. Here one has a stream of primary

signals representing positions of the fringes, ancillary signals representing conditions of the instrument, sinusoidal variations of instrumental origin (LIGO's violin modes), and stochastic noise which may be correlated with instrument conditions. The state-space model would then have terms representing all known effects, and test data would be used to establish the state vector. Likelihood theory could then be used to search for change points (such as chirps from a binary star in-spiral) or astrophysical signals (such as periodicities from rotating neutron stars). Analogous treatments could be made for study of X-ray accretion disk systems, radio pulsars, asteroseismology and exoplanetary, transit detection. State-space models and Kalman filters can also be applied to problems with $1/f$-type long-term memory behavior (Section 11.9).

11.8 Nonstationary time series

As outlined in Section 11.1, astronomical sources exhibit a great variety of nonstationary variability where the mean levels change with time. These include explosions and transients (X-ray bursters, supernovae, gamma-ray bursts, eruptive variable stars, magnetic flaring on stellar surfaces), high and low states (accretion systems like cataclysmic variables and X-ray binaries) and episodic ejections (BL Lac objects, Herbig–Haro objects). There are also many statistical approaches to the analysis of nonstationary time series. Several strategies commonly used in astronomical research, or with promising prospects, are outlined here.

1. Regression for deterministic behavior If a regression model can justifiably be formulated – polynomial, autoregressive or nonlinear based on astrophysical theory – then best-fit parameters can be estimated by least-squares, maximum likelihood or Bayesian methods. Either standard regression methods (Section 11.3.4) or the sophisticated state-space approach (Section 11.7) can be applied. The model may lead to astrophysical insights, or be used as a phenomenological fit to remove long-term trends to allow analysis of the stationary residual time series.

2. Bayesian blocks and bump hunting This is a semi-parametric regression method designed for signal characterization and change-point detection in nonstationary astronomical Poisson time series (Scargle 1998, Dobigeon *et al.* 2007). Here the data consist of a sequence of events, typically photon arrival times from an X-ray or gamma-ray telescope. This procedure constructs a likelihood assuming a sequence of constant flux levels with discontinuous jumps at an arbitrary number of change-points at arbitrary times. The segmentation is conducted by using a hierarchical Bayesian approach with Gibbs sampling to jointly estimate the flux levels and change-point times. The method avoids the limitations of common methods relying on grouping events into bins and then applying methods that rely on Gaussian statistics. Bayesian blocks are widely used to study variable X-ray sources. A similar model can be applied to Gaussian data.

3. Change-point analysis When the behavior of a nonstationary time series changes abruptly, the boundary between the two behaviors is called a change-point. A host of time-domain methods is available to detect change-points (Brodsky & Darkhovsky 1993,

Chen & Gupta 2000). When likelihood methods are used to establish the model of the initial state, the likelihood ratio test can be used to find changes in mean, variance, trend, autocorrelation or other model characteristics. Nonparametric tests (Section 5.3.1), and the engineer's **CUSUM** (cumulative sum) control chart, are also widely used to find structural changes in time series. Change-point analysis tools are currently rarely used in astronomy.

4. Wavelet analysis The wavelet transform, like the Fourier transform, translates the information of a time series into another space through convolution with an orthonormal basis function (Percival & Walden 2000, Mallat 2009). But unlike the Fourier transform, the wavelet transform is well-suited to nonstationary phenomena, particularly if the variations occur on a range of temporal scales or over small portions of the time series. The wavelet transform of X_t at scale s and time u is

$$W_X(u,s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} X_t g\left(\frac{t-u}{s}\right) dt \tag{11.51}$$

where g is a simple basis function (such as the Gaussian, Mexican hat, Daubechies, or cubic spline wavelet function) which rises and falls from zero with some shape. Selection of wavelet coefficients and performing an inverse transform can filter the time series to enhance features at certain scales. This can reduce high-frequency noise, low-frequency trends, or provide efficient data compression (such as the JPEG 2000 standard). Wavelet analysis provides a flexible and powerful toolbox for analyzing nonstationary signals with nonlinear trends, $1/f$-type noise (Section 11.9), change-points, transient signals or signals that are most evident in time-frequency displays. Figure 11.9 illustrates a discrete wavelet transform applied to an astronomical dataset.

Wavelets have been used in hundreds of astronomical studies over 20 years for the study of gamma-ray bursts, solar phenomena, multiperiodic variable stars and a wide variety of spatial analyses. Foster (1996) investigates the performance of wavelet analysis in the analysis of unevenly spaced time series of variable stars, and finds that standard wavelet transforms can give unreliable results. He recommends a wavelet weighted by the local density and variance.

5. Singular spectrum analysis Singular spectrum analysis (SSA) is an approach to time series analysis with roots in multivariate statistics that has become popular in geosciences (including meteorology, oceanography and climatology), dynamical physics and signal processing applications (Elsner & Tsionis 1996). It is a nonparametric, matrix-oriented procedure designed to differentiate and extract broad classes of structures such as trends, oscillatory behavior and noise. SSA first obtains the eigenvectors of the **trajectory matrix** with the lag-covariance $\sum_{t=1}^{n-|i-j|} X_t X_{t+|i-j|}$ as the (i, j)-th element. The time series is then projected onto these orthogonal functions to give temporal principal components. SSA is useful for smoothing time series with complex behavior, extracting nonsinusoidal oscillating signals, interpolating gaps in the time series and locating structural changes in nonstationary time series. In astronomy, it has been used in a few studies of variable solar phenomena.

6. Gamma test The gamma statistic has recently been proposed as a nonlinear curve-fitting technique for temporal and spatial signals based on a minimum-variance criterion

using neighboring values within a specified window (Jones 2004). Local variances are estimated by fitting a simple model to the local structure, calculating the variance for a range of neighbors around each point, and extrapolating these values to zero distance. This is repeated for each data point in the time series. The result is a filtering of the dataset that suppresses noise and enhances localized signals. In astronomical applications, it has been shown to be effective in revealing extended, low-surface-bright features in autocorrelated noise (Boyce 2003). The gamma statistic computation is unusually efficient with $O(n \log n)$ operations and thus may be appropriate for megadatasets.

7. Hidden Markov models Hidden Markov models (HMMs) are a sophisticated and flexible mathematical framework that can simultaneously model stationary stochastic processes (such as red noise or periodicities), predefined structures of various shapes and durations (such as eclipses or outbursts) and their temporal relationships (such as change-points between variability states). As with state-space models (Section 11.7), HMMs are regression techniques where the variables which determine the state of the system are not directly observable, but are hidden behind some stochastic behavior. In Fraser's (2009) phrasing: "Trying to understand a hidden Markov model from its observed time series is like trying to figure out the workings of a noisy machine from looking at the shadows its moving parts cast on a wall, with the proviso that the shadows are cast by a randomly-flickering candle." These models can either be considered alone or in combination with artificial neural networks, state-space models, Kalman filters, Bayesian methods, decision trees and other methods within a likelihood context.

HMMs have been widely applied in speech recognition, image processing, bioinformatics and finance. The developments and applications of HMMs are extensive with hundreds of monographs available; Fraser (2009) provides an accessible introduction. HMMs have not yet been applied to astronomical problems, but are planned for automated classification of millions of spectra to be obtained with the Chinese LAMOST survey telescope (Chen 2006). HMMs could be effective in treating signal extraction problems such as exoplanetary transits and supernova events in the billions of photometric light curves expected from the LSST and other survey telescopes.

11.9 $1/f$ noise or long-memory processes

While many time series in the social sciences are dominated by short-term correlation behavior that is reproduced by ARMA-type models discussed in Section 11.3.3, time series in astrophysical and other physical systems are often dominated by long-term autoregressive behavior (Press 1978). It is common that the variance increases with characteristic time-scale, and the power spectrum can be fit by a power-law function at low frequencies. The physicist and astronomer calls this **$1/f$ noise**, **flicker noise** or **red noise** from an inverse correlation of power against frequency f in a Fourier power spectrum. The statistician calls time series of this type **long-memory** (also called **persistent**, **long range** and **fractional**) noise processes, in contrast to **short-memory** processes that can be understood with low-order ARMA models. A few results from the statistical and econometric literature are

summarized here; details are presented in the volumes by Beran (1989), Robinson (2003) and Palma (2007). Some early results were derived by B. Mandelbrot in his well-known studies of fractal processes.

Long-memory processes can arise in several ways (Beran 1989, Chapter 1). One possibility is the combination of many short-memory subprocesses. The sound of a bubbling brook has a $1/f$-type noise structure because it aggregates the events associated with different interactions between the moving water and obstructions in the flow. In astronomy, the observed emission may appear to arise from an unresolved source, but in reality is the aggregation of variations at different locations on (say) the surface of a star or accretion disk. Another possibility is that a single process with self-similar hierarchical variation is involved, perhaps regulated by stochastic differential equations. A turbulent interstellar cloud may have this property, although the variations we study are spatial not temporal. A third possible source of long-memory variations are instrumental and observational problems: calibration drifts, changes in background conditions during a night or satellite orbit, or interactions between stages in a complex data reduction procedure.

A long-memory process is often modeled with an autocorrelation function that decays as a power law at large time lags $k \to \infty$,

$$ACF_{LM}(k) \propto k^{2d-1} \tag{11.52}$$

for $0 < d < 1/2$. **Hurst's self-similarity parameter** $H = d + 1/2$ is sometimes used. In the context of ARMA-type models, this is called a **fractional ARIMA process (FARIMA or ARFIMA)**. For example, the variability of a FARIMA(0,d,0) model is intermediate between those of a stationary AR(1) and a nonstationary random walk model. The corresponding Fourier spectral density function at low frequencies is

$$f_{LM}(\nu) \propto |\nu|^{-2d}. \tag{11.53}$$

The $1/f$-noise case corresponds to $d = 1/2$ which is marginally nonstationary.

Long-memory processes in general can be modeled using likelihood methods. These include maximizing the likelihood with the **Durbin–Levinson algorithm**, solving the state-space model with the Kalman filter, deriving Bayesian posterior distributions for assumed priors and loss functions, and constructing the periodogram using Whittle's approximate likelihood. The **Whittle (1962) likelihood** is a modified MLE applied in the Fourier domain designed for autoregressive processes. When the variance of the noise variations changes with time, then heteroscedastic models are needed. These likelihood calculations can be technically difficult, asymptotic properties of the statistics can be complicated, and small-sample behavior may be unknown.

For the problem of estimating the slope d in a Gaussian fractional noise process, and discriminating long-memory processes ($d > 0$ in Equation 11.52) from a random walk, three solutions have been widely used.

First, an estimate for the slope d can be obtained by least-squares regression on the periodogram (Robinson 1995). Given the periodogram from the discrete Fourier transform

$$I(\nu) = \left| \frac{\sum_{t=1}^{n} X_i e^{it\nu}}{\sqrt{2\pi n}} \right|^2 \tag{11.54}$$

and defining

$$a_j = \ln(4 \sin^2(\nu_j/2)), \tag{11.55}$$

the least-squares slope estimator is

$$\hat{d}_L S(l) = \frac{\sum_{j=l+1}^{m} (a_j - \bar{a}) \ln(I(\nu_j))}{\sum_{j=l+1}^{m} (a_j - \bar{a})^2} \tag{11.56}$$

where $0 \leq l < m < n$ and $\bar{a}$ is the mean of the time series. This method is often called the **Geweke–Porter-Hudak (GPH) estimator** (Geweke & Porter-Hudak 1983). However, its mathematical properties (e.g. asymptotic bias and variance) are not strong. In particular, it is difficult to decide the precise frequency range for computing the linear regression; recommendations for the optimal number of periodogram values include $m \simeq n^{1/2}$ and $m \simeq n^{4/5}$ where n is the number of data points in the time series (Hurvich *et al.* 1998). The regression can be performed on a smoothed or multi-tapered periodogram; a similar method of regression on the binned periodogram was independently developed by astronomers Papadakis and Lawrence (1993). Other modifications to the GPH estimator have also been suggested (Andersson 2002).

Second, a least-squares time-domain procedure for estimating the slope d, called **detrended fluctuation analysis**, originally developed to study DNA sequences, has become popular in a variety of fields (Peng *et al.* 1994). The time series x_t is grouped into k bins each with $m = n/k$ data points. A linear regression is performed within each bin and the residual variance is computed:

$$\sigma_k^2 = \frac{1}{m} \sum_{i=1}^{m} (X_t - \hat{\alpha}_k - \hat{\beta}_k t)^2. \tag{11.57}$$

Note that the regression lines will be discontinuous at the bin boundaries. The average of these variances, $F = (1/k) \sum_{j=1}^{k} \sigma_k^2$, is related to the slope of the long-memory process d as $F(k) \propto k^{d+1/2}$ for a Gaussian fractional noise process. The slope d can then be estimated by another linear regression between $\log F$ and $\log k$. The method is described by Palma (2007, Section 4.5.4) and is applied to an astronomical dataset in Figure 11.8.

Third, an estimator for d has been developed based on the **discrete wavelet transform** of the time series. A vector of squared wavelet amplitudes summed over different time-scales is distributed as χ^2 with dependency on d. The slope is then estimated from a linear regression involving this vector and the original time series. In general, wavelet analysis of nonstationary $1/f$-type processes can be effective.

In general, statistical evaluation of $1/f$-type (fractional ARMA) processes encounters many pathologies (Beran 1989, Chapters 8 and 10; Robinson 2003, Chapter 10). While the sample mean of a fractional ARMA process is unbiased, its convergence to the true mean can be very slow and standard confidence intervals are incorrect. The standard sample variance is biased, and some long-memory processes have infinite variance. Both the spectral density at zero frequency and the area under the ACF over all lags can be infinite. Caution is also needed when seeking relationships between a long-memory random variable and covariates, as these time series show excursions from the mean over various time-scales and short samples often show spurious trends. Standard statistics for hypothesis testing,

like the likelihood ratio test, no longer follow a χ^2 distribution. Nonparametric e.d.f.-based goodness-of-fit tests, like the Kolmogorov–Smirnov test, can be very inaccurate: even true models may be rejected due to long-memory variations. While some of these problems arise with all autocorrelated processes, they are particularly exacerbated with long-memory processes.

Clearly, simple statistical evaluation of $1/f$-type time series is fraught with difficulties. For time series with mild variations, or with very long durations, cautious use of standard methods may be tractable. Variable transformation, such as considering the differenced time series $x_i - x_{i-1}$ or the inverse of the periodogram, sometimes helps. A variety of sophisticated mathematical approaches to these problems have been developed over the past several decades with associated theorems and convergence calculations (Beran 1989, Robinson 2003, Palma 2007). Research on long-memory processes is rapidly developing with applications in genetics, finance, human physiology, Internet studies, geology, meteorology, ecology and other fields. There are many new methods available for astronomers to consider when treating $1/f$-type processes in astronomical time series.

11.10 Multivariate time series

As mentioned in Section 11.1, astronomical time series are sometimes measured in several wavelength bands which may show similar, lagged or different variations depending on the astrophysical situation. This is the domain of multivariate time series analysis which is well-developed for problems encountered often in fields like econometrics and meteorology. For example, temporal variations in the prices of fertilizer, corn and pork may be correlated with time lags associated with the agricultural process of growing grain for livestock consumption. Analogously, the peak emission from supernova explosions or gamma-ray burst afterglows may evolve from short to long wavelengths as the expanding fireball cools or becomes transparent to different wavelengths of light.

For a bivariate stochastic temporal process measured in evenly spaced observations of the form (X_i^1, X_i^2) for $i = 1, \ldots, n$, the sample second-order moments are the two individual autocorrelation functions and the **cross-correlation function**

$$CCF(k) = \frac{1}{n} \frac{\sum_{i=1}^{n-k} (X_t^1 - \bar{X}^1)(X_t^2 - \bar{X}^2)}{\sqrt{Var(X^1)Var(X^2)}} \tag{11.58}$$

for lags $k \geq 0$. Here the numerator is called the cross-covariance function and the denominator is the product of the two sample standard deviations. This estimator of the true cross-correlation function is consistent and asymptotically unbiased, but its variance may be large and its distribution difficult to derive as it convolves the variations of the individual time series. The CCF is thus often calculated on time series residuals after periodic, trend or autocorrelated signals are removed by pre-whitening, regression or ARMA-type modeling. If the CCF shows a single strong peak at some lag k, then this may reflect similar structure in the two time series with a delay. However, the CCF may show structure for other reasons, so a careful check of this interpretation is needed.

The bivariate cross-correlation function can be readily generalized to multivariate time series using a correlation matrix function where the (i, j)-th element of the matrix is the bivariate CCF for variables x^i and x^j. In econometrics and other fields, one often seeks explanatory and predictive relationships between a single dependent variable and a suite of covariates, all of which vary in time and may be correlated with each other with different lag-times. A large literature exists to address these dynamic regression, **distributed lag** or **vector autoregressive (VAR)** models. But astronomers can rarely acquire evenly spaced time series data in many variables for a single object, so this approach will not be generally useful in astronomy.

11.11 Remarks

Time series analysis is a vast field with a myriad methods that might be applied. In some cases, the typical astronomical analysis of time series data is well-motivated. When sinusoidal periodicities are present or suspected, Fourier analysis is an excellent approach. Astronomers have led innovations in spectral analysis for finding periodicities in unevenly space data and nonsinusoidal signals. Many of these methods are reasonable, and the interpretation by users is often simplistic. The Lomb–Scargle periodogram (perhaps the most highly cited astrostatistical method in the astronomical literature) has a strong mathematical foundation, but its performance is sometimes imperfect and the statistical significance of its peaks is difficult to establish. When secular trends are present, then regression is a correct procedure; indeed, here the independent variable t is fixed, and this model is more appropriate to time series data than to regression between two random variables (Chapter 7).

Astronomers have actively generated non-Fourier periodograms to treat common problems of unevenly spaced data points, nonsinusoidal recurrent patterns, and heteroscedastic measurement errors, and Poisson rather than Gaussian processes. There is ample need for statisticians to evaluate these methods, establishing their mathematical properties and limitations.

However, astronomers have habitually used spectral analysis, rather than time-domain modeling, to analyze aperiodic and stochastic phenomena. Although all information in the original time series is contained in the Fourier coefficients (as it is in the autocorrelation function), the periodogram does not concentrate the signal well for aperiodic autoregressive signals. For $1/f$-type noise, for example, a FARIMA model might capture the essential behavior in a handful of parameters which requires dozens of components of a Fourier spectrum.

It is important to recognize that when any kind of autocorrelated structure is present – periodicities, secular trends or aperiodic autocorrelated variations – then standard statistics may not apply; in particular, variances and standard deviations are often underestimated. This is relevant to many astronomical problems. Consider, for example, the estimation of the equivalent width of a faint spectral line in a univariate spectrum dominated by continuum emission. Here the wavelength is a fixed time-like variable so time series methods apply.

The common procedure is to evaluate the variance of the signal around the putative line, compute the signal-to-noise ratio at the wavelength of interest, and evaluate its significance from the normal distribution. But the continuum signal often has autocorrelation due either to structure in the source continuum emission or to instrumental effects. In such cases, the assumption of i.i.d. underlying the standard computation of the variance is wrong and the equivalent width significance can be overestimated. The problem is particularly exacerbated when the autocorrelation function does not converge as in some $1/f$-type noise behaviors, the statistician's long-memory processes.

One reason astronomers often prefer Fourier to time-domain analysis is the astrophysical interpretation of the parameters. There is often a lack of a clear relationship between the derived model parameters of an ARMA-like model and the scientific goals of the study. In economic or meteorological studies, the meaning of the parameter values of an FARIMA or ARCH model is not as important as obtaining a successful prediction of future values. But the astronomer has less interest in predicting the future X-ray luminosity of an accretion disk than in understanding the physical processes underlying the variations. Astrophysical theories today are rarely designed to predict autoregressive behavior. One unusual success in linking advanced time series model parameters to a physical process is the interpretation of the variations of the X-ray source Sco X-1 which exhibits red noise and quasi-periodic oscillations on time-scales from milliseconds to hours. Using Fourier and wavelet decompositions, the empirical time series model has been linked to a physical model known as the **dripping rail** (Scargle *et al.* 1993). We encourage future astrophysical theoretical studies to make predictions of autoregressive properties to match the statistical benefits of time-domain time series modeling of non-periodic phenomena.

Astronomers can benefit from other statistical approaches to time series analysis, particularly for nonstationary phenomena. Wavelet analysis and Bayesian blocks are examples of methods gaining popularity for treating multi-resolution features that are present only at certain times. State-space and hidden Markov modeling are sophisticated and flexible approaches having considerable potential for understanding complex astronomical time series problems.

The statistical findings summarized here are not restricted to time series analysis, but will be relevant to any astronomical situation where the data are a function of a single ordered time-like variable. This includes astronomical spectra where intensity is a function of wavelength, and images where astronomical spectra are a function of pixelated position on a detector. We encourage experimentation with time series methodology for the analysis of astronomical spectra and images.

11.12 Recommended reading

Due to the wide range of time series methodologies, we list below only broad-scope texts. Recommended topical books include Warner (1998, elementary) and Percival & Walden (1993, advanced) for spectral analysis, Kitagawa & Gersch (1996) and Durbin & Koopman (2001) for state-space methods, Fan & Yao (2003) for nonlinear models, Mallat (2009) and

Percival & Walden (2000) for wavelet analysis, and Beran (1989) and Palma (2007) for long-memory processes ($1/f$-type noise).

Chatfield, C. (2004) *The Analysis of Time Series: An Introduction*, 6th ed., Chapman & Hall/CRC, Boca Raton
An excellent elementary presentation of time series analysis. Coverage includes stationary stochastic processes, autocorrelation, autoregressive modeling, spectral analysis, state-space models, multivariate modeling and brief introductions to advanced topics.

Cowpertwait, P. S. P. & Metcalfe, A. V. (2009) *Introductory Time Series with R*, Springer, Berlin
A short and clear elementary treatment of times series analysis with **R** applications covering autocorrelation, forecasting, stochastic models, regression, stationary and non-stationary models, long-memory processes, spectral analysis, multivariate models and state-space models.

Shumway, R. H. & Stoffer, D. S. (2006) *Time Series Analysis and Its Applications with R Examples*, 2nd ed., Springer, Berlin
A well-written intermediate-level text with useful descriptions of **R** implementation. Topics include exploratory data analysis, regression, ARMA-type models, spectral analysis and filtering, long-memory and heteroscedastic processes, state-space models and cluster analysis.

Wei, W. W. S. (2006) *Time Series Analysis: Univariate and Multivariate Methods*, 2nd ed., Pearson, Harlow
A more detailed text on time series. Coverage includes autocorrelation, ARMA-type models, regression, model selection, Fourier and spectral analysis, vector time series, state-space models and the Kalman filter, long-memory and nonlinear processes, aggregation and sampling.

11.13 R applications

Excellent tutorials on time series analysis using **R** are provided by Shumway & Stoffer (2006) and Cowpertwait & Metcalfe (2009). Elementary methods are reviewed by Crawley (2007).

We use the variability of Galactic X-ray source GX 5-1 to illustrate many of the time series capabilities of **R**. GX 5-1 is an X-ray binary-star system with gas from a normal companion accreting (falling) onto a neutron star. Highly variable X-rays are produced in the inner accretion disk, often showing stochastic red noise and quasi-periodic oscillations of uncertain origin. The dataset, described in Appendix C.12, consists of 65,536 measurements of photon counts in equally spaced 1/128-second bins obtained with the Japanese Ginga satellite during the 1980s. Although strictly a Poisson process, the count rates are sufficiently high that they can be considered to be normally distributed.

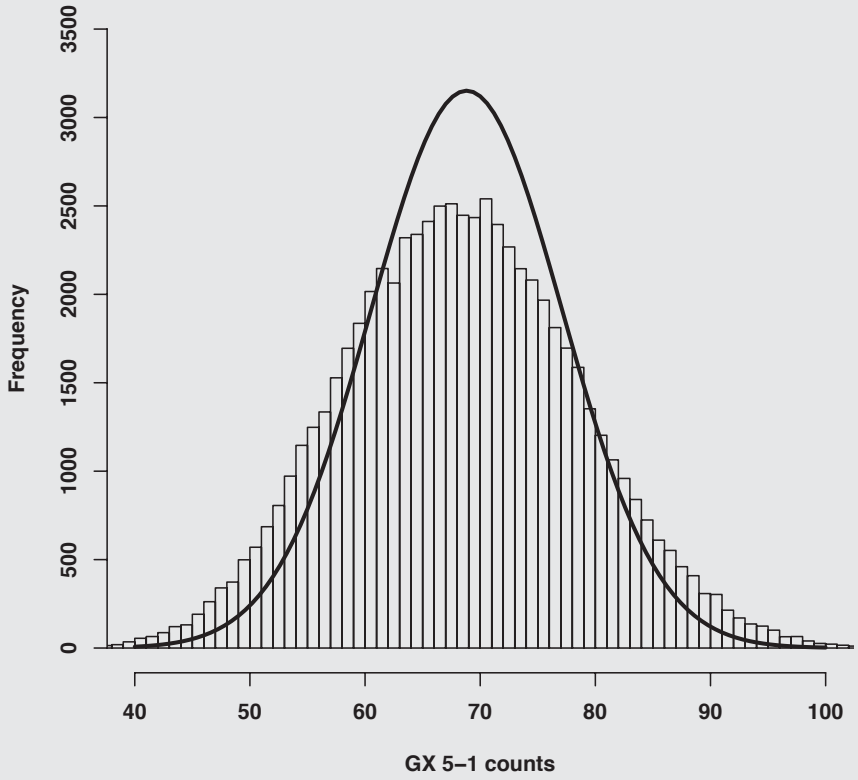

Fig. 11.1 Histogram comparing the distribution of X-ray counts in 1/128-second time increments to a normal distribution.

Using methods that are rarely considered in astronomy, but which are standard in many other applications of time series analysis, we have arrived at some unusual conclusions with respect to standard astronomical interpretations of rapid X-ray variations in accretion binary systems (van der Klis 2000). The structure can be attributed to high-order autoregressive behavior plus a very-low-frequency component. While red noise is evident, we find its interpretation as a simple power-law $1/f^{\alpha}$ component to be inaccurate. While a quasi-periodic component is evident, we find it is entirely attributable to autoregressive behavior with no strict periodicities involved.

11.13.1 Exploratory time series analysis

We start with visual exploration of the dataset in a variety of displays. The histogram of counts in each bin shows that the variance is 24% larger than expected from a white noise process, and asymmetry about the mean is present (Figure 11.1). Examination of the time series in its raw form (Figure 11.2) and after various smoothing operations (Figure 11.3) does not reveal obvious nonstationary structure to account for the extra variance, although

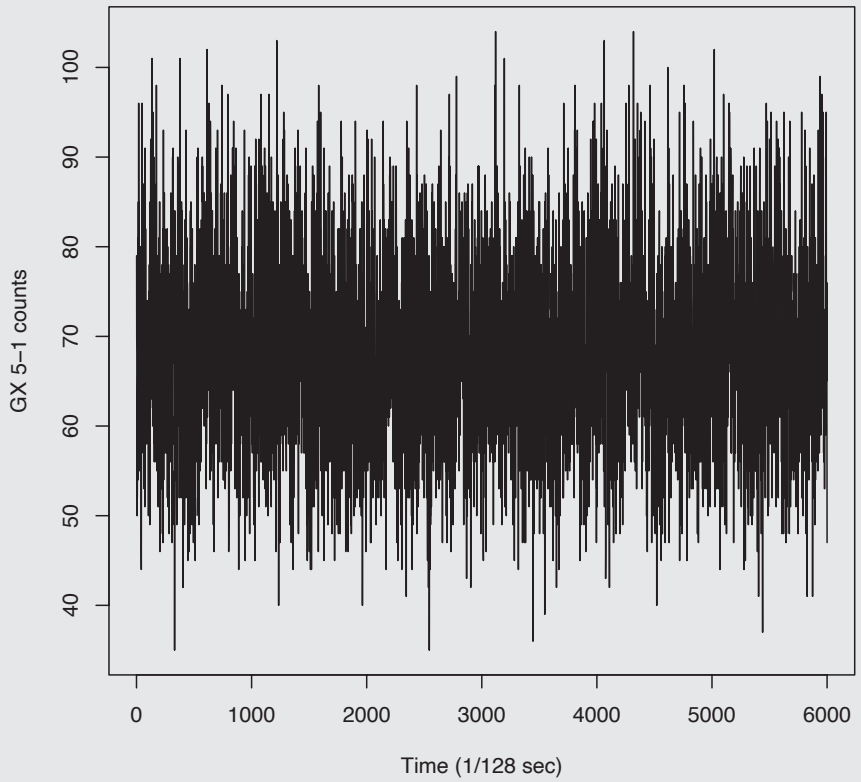

Fig. 11.2 Raw time series of the X-ray emission from the Galactic X-ray binary system GX 5-1 measured with the Ginga satellite observatory. This plot show the first few percent of the dataset.

some of the smoothers give a hint of autocorrelated variations with a characteristic timescale around 20–50 seconds. GX 5-1 thus appears to exhibit stochastic, possibly autocorrelated and quasi-periodic, variability that can now be investigated in detail.

In the **R** code below, we start by reading this ASCII dataset using the *scan* function rather than *read.table*, as we want the values to be placed into a single vector rather than into a tabular matrix. The counts and time stamps are collected into an object with **R** class "time series" using the *ts* function; this allows a wide variety of time series functions to operate on the dataset. The normal function superposed on the histogram uses the function *curve* because *hist* does not permit superposition using *lines*. In Figure 11.3, we display five smoothers offset vertically. Two are based on the GX.ts time series object and use the *filter* and *kernapply* functions. Three operate on the original counts and time stamp vectors and use kernel density estimation, a nearest-neighbor super-smoother and the *lowess* local regression fit described in Section 6.6.2. In three cases, we chose a bandwidth of seven time bins, but for the super-smoother we use the cross-validation bandwidth default and for *lowess* we choose a span of 0.05. Examination of different smoothers and bandwidths is encouraged during the exploratory stage of a time series analysis.

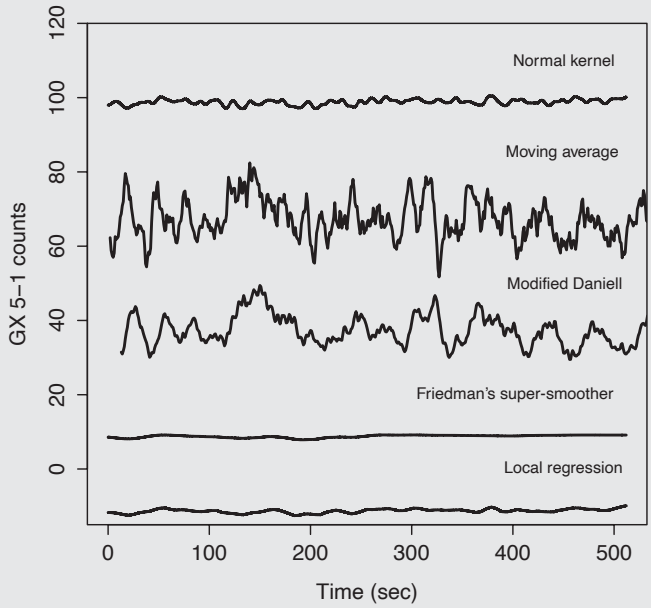

Fig. 11.3 The full GX 5-1 time series subject to various smoothing procedures: normal kernel density estimator, moving average estimator, modified Daniell (boxcar) smoother, Friedman's cross-validation super-smoother, and Cleveland's local regression model. Except for the moving average smoother, the curves are displaced vertically for clarity.

```
# Read in GX 5-1 data and create time series

GX.dat <- scan("http://astrostatistics.psu.edu/MSMA/datasets/GX.dat")
GX.time <- seq(from=0, to=512, length.out=65536)
GX.ts <-  ts(GX.dat, GX.time) ; GX.ts.offset <- ts(GX.dat-30, GX.time)

# Compare histogram of counts to normal distribution

hist(GX.dat, breaks=100, xlim=c(40,100), ylim=c(0,3500), xlab='GX 5-1 counts',
   font=2, font.lab=2, main='')
curve(dnorm(x,mean=mean(GX.dat), sd=sqrt(mean(GX.dat)))*65536, lwd=3, add=T)
sd(GX.dat) / sqrt(mean(GX.dat))  # result is 1.24

# Examine raw and smoothed time series

plot.ts(GX.ts[1:6000], ylab='GX 5-1 counts', xlab='Time (1/128 sec)',
   cex.lab=1.3, cex.axis=1.3)
plot(GX.time,GX.dat, ylim=c(-10,115), xlab='Time (sec)', ylab='GX 5-1 counts',
   cex.lab=1.3, cex.axis=1.3, type='n')
lines(ksmooth(GX.time, GX.dat+30, 'normal', bandwidth=7), lwd=2)
```

```
text(450, 110, 'Normal kernel')
lines(filter(GX.ts, sides=2, rep(1,7)/7), lwd=2)
text(450, 85, 'Moving average')
lines(kernapply(GX.ts.offset, kernel('modified.daniell', 7)), lwd=2)
text(450, 50, 'Modified Daniell')
lines(supsmu(GX.time, GX.dat-60), lwd=2)
text(400, 20, 'Friedman's super-smoother')
lines(lowess(GX.time, GX.dat-80, 0.05), lwd=2)
text(450, 0, 'Local regression')
```

11.13.2 Spectral analysis

Many astronomical time series analyses emphasize harmonic or spectral analysis. The **R** script below starts with a manual construction of the Schuster periodogram based on the script provided by Shumway & Stoffer (2006, Chapter 2), and then uses the automated function *spec.pgram* (Figure 11.4, top panel). Note that entries in the raw periodogram can be very uncertain; for this spectrum, the 500th spectral value is 119 with 95% range [32,4702] assuming an asymptotic χ^2 distribution.

We then proceed to experiment with various smoothers and tapers to reduce the variance of the raw periodogram. A taper has little effect in this case; they are more important for shorter datasets. Smoothing, however, has a dramatic effect on the periodogram appearance and much is learned about the excess variance of the time series (Figure 11.4). It appears to rise from two distinct processes: a noise component below ~0.05 rising towards lower frequencies, and a strong but broadened spectral peak around 0.17–0.23. These are the red noise and quasi-periodic oscillations addressed in hundreds of studies of accretion binary-star systems like GX 5-1 (van der Klis 2000).

```
# Raw periodogram

f <- 0:32768/65536
I <- (4/65536) * abs(fft(GX.ts) / sqrt(65536))^2
plot(f[2:60000], I[2:60000], type="l", ylab="Power", xlab="Frequency")

Pergram <- spec.pgram(GX.ts,log='no',main='')
summary(Pergram)
Pergram$spec[500] # value of 500th point
2*Pergram$spec[500] / qchisq(c(0.025,0.975),2)

# Raw and smoothed periodogram

par(mfrow=c(3,1))
spec.pgram(GX.ts, log='no', main='', sub='')
spec.pgram(GX.ts, spans=50, log='no', main='', sub='')
spec.pgram(GX.ts, spans=500, log='no', main='', sub='')
```

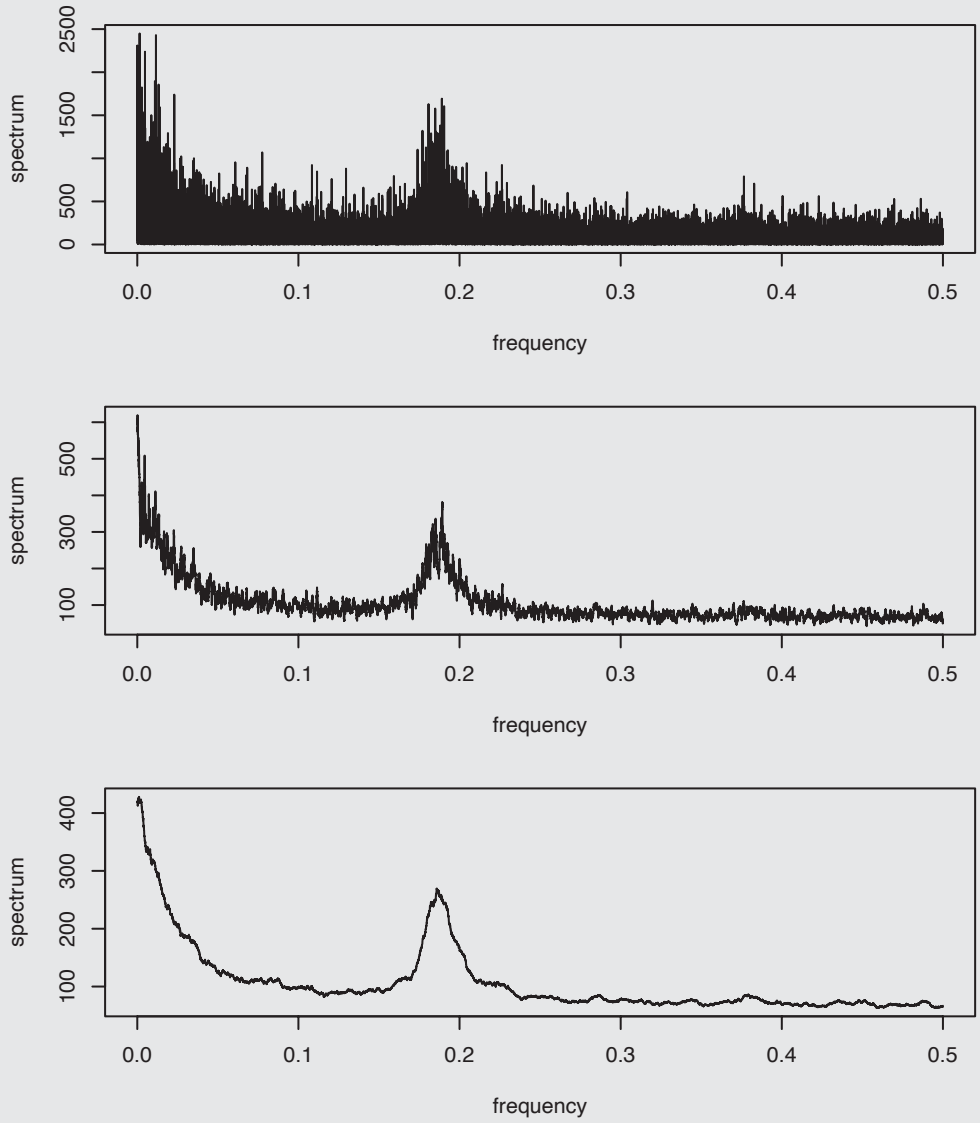

Fig. 11.4 Raw and smoothed (modified Daniell with $w = 50$ and 500 bandwidths) periodograms for GX 5-1.

11.13.3 Modeling as an autoregressive process

The red noise and (quasi-)periodic structure in the GX 5-1 data are clearly seen in plots of the (partial) autocorrelation function (Figure 11.5). The PACF is more informative: the strongest predictor of a current value are the values separated by 4–6 time increments, and oscillations have periods around 5–6 increments. The envelope of significant autocorrelation extends to lags of 20–30 increments.

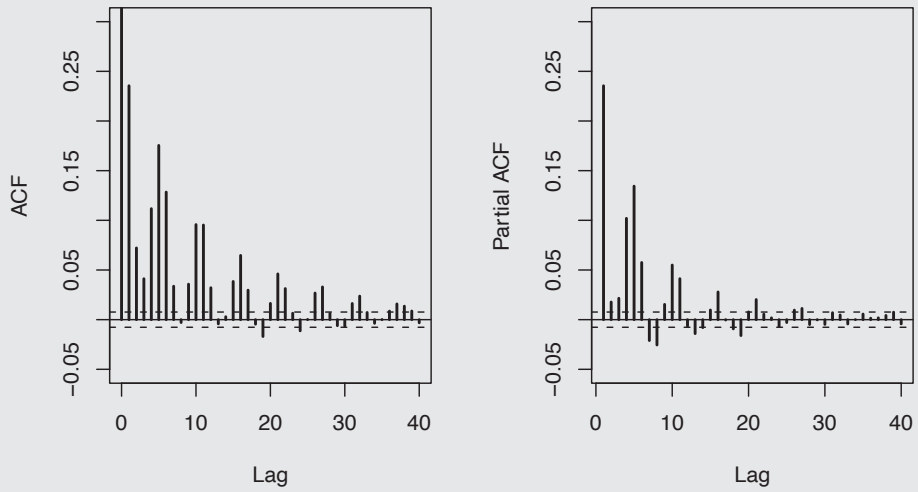

Fig. 11.5 Autocorrelation and partial autocorrelation functions for GX 5-1 time series.

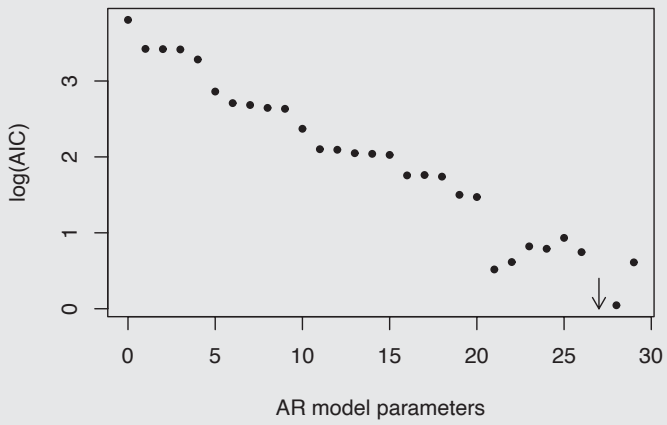

Fig. 11.6 Values of the Akaike information criterion as a function of model complexity for autoregressive AR(p) models of the GX 5-1 time series.

The presence of autocorrelation and (from the smoothed time series) absence of changes in the average level motivates an AR (rather than ARMA or ARIMA) model. The *ar* function in **R** selects a model using the minimum Akaike information criterion. Several computational options are provided; we choose ordinary least squares. AR(27) is the best model, although any model with order $p > 22$ is adequate (Figure 11.6). The largest positive coefficients are at lags $= 1$, 4, 5 and 6 and negative coefficients at lags 10 and 11, again indicating oscillation with period around five increments. Parameter uncertainties can be based on asymptotic theory in the function *ar* or by bootstrap resampling of the model

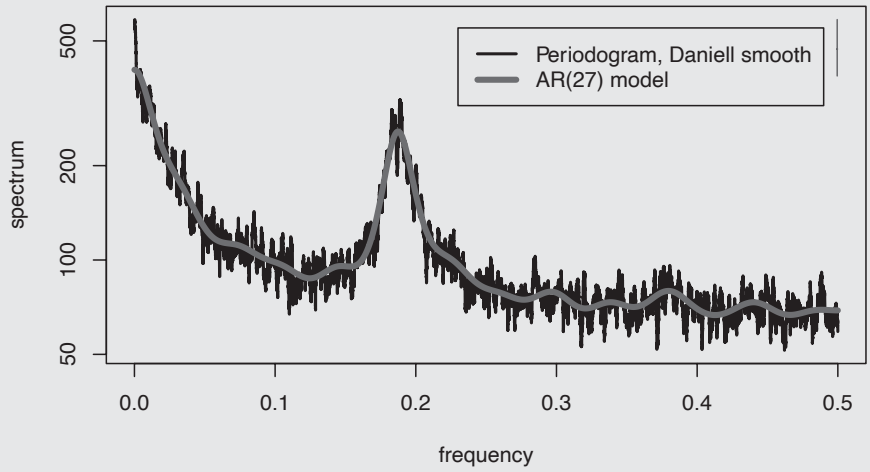

Fig. 11.7 Smoothed periodogram of the GX 5-1 time series and its AR(27) model. For a color version of this figure please see the color plate section.

residuals. An **R** script for the bootstrap resampling procedure is presented by Shumway & Stoffer (2006, Chapter 3).

Since the AR model is stochastic, we cannot meaningfully compare its time series to the observed GX 5-1 data. However, the Fourier spectrum of the model can be compared to the GX 5-1 spectrum. The function *ar.spec* in **R** runs *ar*, chooses the model with the minimum AIC, and computes the Fourier transform of the model. This program found that AR(27) is the best model, as it gives a slightly stronger spectral signal around frequency 0.19 than the AR(12) model. Figure 11.7 shows that the spectral properties of GX 5-1 are remarkably well-reproduced by the model: the steep red noise signal at frequencies below 0.05, the shallow red noise around 0.05−0.1, the height and width of the QPO peak around 0.19, the slight asymmetry around the peak, the weak harmonic around 0.38, and the spectral noise level at high frequency.

```
# Autocorrelation functions of the GX 5-1 time series

par(mfrow=c(1,2))
acf(GX.ts, 40, main='', ci.col='black', ylim=c(-0.05,0.3), lwd=2)
pacf(GX.ts, 40, main='', ci.col='black', ylim=c(-0.05,0.3), lwd=2)

# Autoregressive modeling

ARmod <- ar(GX.ts, method='ols')
ARmod$order # model selection based on AIC
ARmod$ar # best-fit parameter values
ARmod$asy.se.coef$ar # parameter standard errors
```

```
plot(0:29, log10(ARmod$aic[1:30]), xlab='AR model parameters',
    ylab='log(AIC)', pch=20)
arrows(27, 0.4, 27, 0.0, length=0.1)

# Spectrum of AR model

ARspec <- spec.ar(GX.ts, plot=F)
GXspec <- spec.pgram(GX.ts, span=101, main='', sub='', lwd=2)
lines(ARspec$freq, ARspec$spec, col='red', lwd=4)
legend(0.23, 550, c('Periodogram, Daniell smooth', 'AR(27) model'),
   lty=c(1,1), lwd=c(2,4), col=c('black','red'))
```

11.13.4 Modeling as a long-memory process

The only spectral feature in Figure 11.7 not accounted for by the AR model is the excess signal at the very lowest frequency. This indicates that the GX 5-1 behavior has an additional long-term memory component. As discussed in Section 11.9, we can treat long-memory processes in both the time domain and frequency domain. We can generalize the AR model above to nonstationary ARIMA and FARIMA models. The *arima* function in **R** calculates the likelihood in a state-space formulation using the Kalman filter (Section 11.7). However, in this case with 27 AR parameters, it is computationally expensive to calculate a range of ARIMA models with additional differencing and moving averaging components. Trials with fewer AR parameters suggests that models such as ARIMA(12,1,1) provide little benefit over the corresponding AR model.

Several estimators of the long-memory parameter d are provided by **CRAN** packages based on both spectral and time-domain techniques. Recall from Section 11.9 that $d = \alpha/2$ and the Hurst parameter is $H = d + 1/2$, where the spectral power is $P \propto (1/f)^{\alpha}$. The function *fdGPH* in **CRAN** package *fracdiff* (Fraley 2009) calculates the Geweke–Porter-Hudek estimator from linear regression on the raw log periodogram; this method gives $d = 0.10 \pm 0.04$. The function *fdSperio* uses Reisen's (1994) regression on the log periodogram smoothed with a lag Parzen window and obtains the same value with higher precision, $d = 0.10 \pm 0.01$. These values indicate a relatively weak long-memory process with noise power scaling roughly as $1/f^{0.2}$. Similar calculations are provided by function *hurstSpec* in package *spectral*.

In the time domain, a FARIMA model calculation obtained from the **CRAN** package *fracdiff* gives a best-fit value of $d = 0.06$. The **CRAN** package *fractal* has several functions which compute the long-memory Hurst parameter from the time-domain data (Constantine & Percival 2010). These results are noticeably inconsistent. We obtain $H = 0.03$ using detrended fluctuation analysis; 0.12 from an arc-hyperbolic-sine transformation of the autocorrelation function; and 0.66–1.00 from various blockings of the time series. Clearly, it is difficult to establish a consistent value for the $1/f^{\alpha}$-noise spectral shape. The *fractal* package also has useful capabilities for simulating several types of fractal time series for

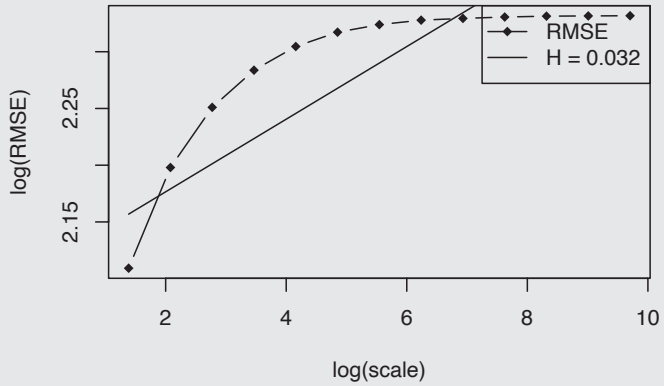

Fig. 11.8 Log-log plot of the root mean square error against scale from the detrended fluctuation analysis of the GX 5-1 time series. The slope of the linear fit gives the long-range memory Hurst parameter.

comparison with a real dataset: a $1/f^{\alpha}$ process; a fractionally differenced process and fractional Gaussian noise; and fractional Brownian motion.

The likely cause of these discrepant results is that the red noise in GX 5-1 is not truly fractal, such as a $1/f^{\alpha}$ process where the scaling of the variance on different scales would be a power law. A plot from the detrended fluctuation analysis shows a poor fit of a power-law model to the data (Figure 11.8), so that different estimation procedures of the fractal dimension arrive at different results.

```
# Spectral estimates of the long-range memory parameter d

install.packages('fracdiff') ; library(fracdiff)
d.FARIMA <- fracdiff(GX.ts, nar=27, nma=1, ar=ARmod$ar) ; d.FARIMA$d
d.GPH <- fdGPH(GX.ts)
d.Reisen <- fdSperio(GX.ts)

# Time domain estimates of the long-range memory parameter H=d+1/2

install.packages('fractal') ; library(fractal)
d.DFA <- DFA(GX.ts) ;  d.DFA
plot(d.DFA)
d.ACVF <- hurstACVF(GX.ts) ; d.ACVF
d.block <- hurstBlock(GX.ts)  ; d.block
```

11.13.5 Wavelet analysis

In examination of the GX 5-1 time series, we did not see any short-lived structures such as rapid increases or decreases in emission. However, it is valuable to make a wavelet decomposition of the time series to assist in visualization of the dataset at different time-scales. Here we use **CRAN**'s package *waveslim* (Whitcher 2010) to construct the discrete

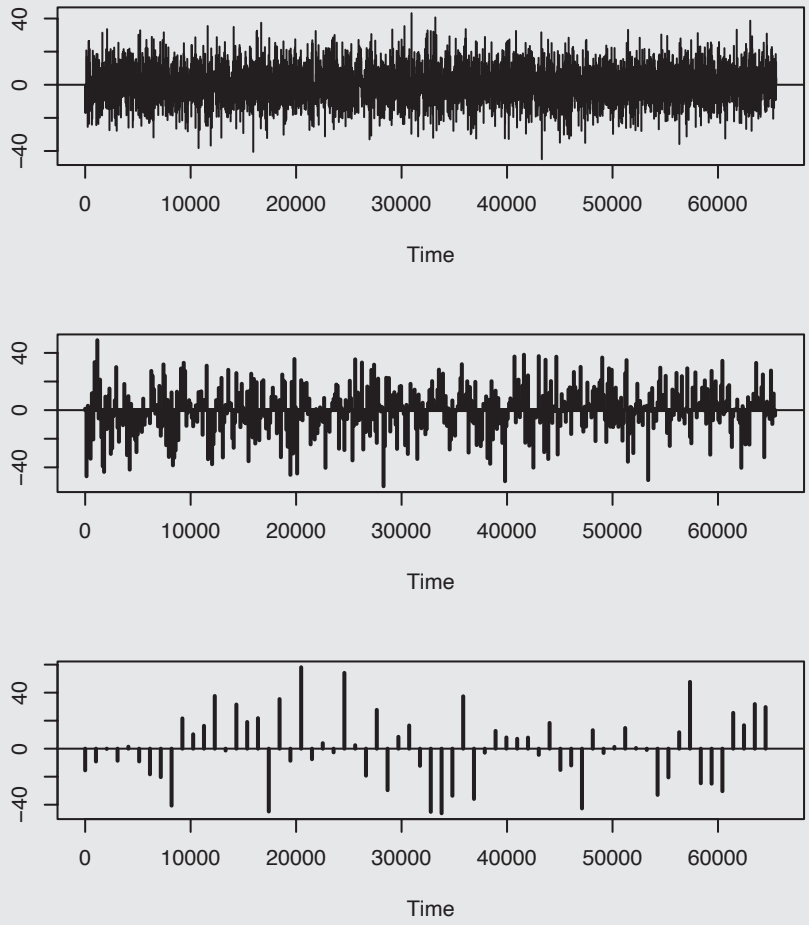

Fig. 11.9 Wavelet coefficients during the GX 5-1 time series at three time resolutions.

wavelet transform using Mallat's pyramid algorithm and Daubechies' orthonormal basis functions at ten different temporal scales. This package implements a variety of wavelet methods for one-, two-, and three-dimensional data as described in the volume by Gencay *et al.* (2001). Figure 11.9 shows the wavelet coefficients for the time series at three scales. As expected, no obvious structure is seen.

```
# Discrete wavelet transform

install.packages('waveslim') ; library(waveslim)
GX.wav <-  dwt(GX.ts,n.levels=10)
par(mfrow=c(3,1))
plot.ts(up.sample(GX.wav[[4]],2^4),type='h',axes=F,ylab='') ; abline(h=0)
plot.ts(up.sample(GX.wav[[7]],2^7),type='h',axes=F,ylab='',lwd=2) ; abline(h=0)
plot.ts(up.sample(GX.wav[[10]],2^{10}),type='h',axes=F,ylab='',lwd=2) ; abline(h=0)
par(mfrow=c(1,1))
```

11.13.6 Scope of time series analysis in R and CRAN

R supports the class *ts* for evenly spaced time series data. Basic analysis includes autocorrelation functions (*acf* and *pacf*) and linear filters with the function *filter*. Modeling includes univariate (*arima*) and multivariate (*VAR*) autoregressive models, maximum likelihood structural regression models (*StructTS*) and local regression models (*stl*).

CRAN's *zoo* package (with its extension *xts*) creates the *zoo* class for ordered observations such as irregularly spaced time series (Zeileis & Grothendieck 2005). Editing, merging, plotting, summaries, computing lags and windows, aggregation into bins, interpolation between values, treatment of *NA* values, conversion from other **R** classes, and other infrastructure is provided, but support for duplicate observation times is limited. At the present time, there are no time series analysis applications.

CRAN has dozens of time series packages for evenly spaced data; an overview is provided at http://cran.r-project.org/web/views/TimeSeries.html. While many cover similar material primarily oriented towards econometrics and finance, a number of packages may be useful to astronomers. If outliers are present, and need to be downweighted or removed, the *robfilter* package provides a variety of iterated filters and regressions based on robust statistics (Section 7.3.4). The *RTisean* package provides a number of functions for high-pass filtering, noise reduction, polynomial modeling and more. Several very capable packages for wavelet analysis are available including *wavelets*, *wavethresh* and the *wmtsa* package accompanying the monograph by Percival & Walden (2000). The package *sde* implements methods for stochastic differential equations discussed in the monograph by Iacus (2008). Bayesian approaches are implemented in *bspec* for autocorrelation and spectral analysis, *ArDec* for autoregressive decomposition, and *MSBVAR* for dynamic vector autoregressive modeling.

An **R** script for the Lomb–Scargle periodogram, designed for genetics research problems, is available outside of **CRAN** at http://research.stowers-institute.org/efg/2005/LombScargle.