Chapter 18
Time Series

18.1 Introduction

In the analysis of neural data, time is important. We experience life as evolving, and neurophysiological investigations focus increasingly on dynamic features of brain activity. If we wish to understand the signals produced by nervous system processes we must use an analytical framework that is built for time-varying observations.

From a mathematical point of view, time is a number with an arbitrarily-chosen origin, the value $t = 0$ typically representing an experimental or behavioral marker such as the onset of a visual cue. We may work backward in time by taking t to be negative. Although measurements are always made with some resolution of temporal accuracy, often determined by a sampling rate (such as 20 KHz, giving a precision of $\Delta t = .05$ ms), mathematically we allow t to be any real number, such as $t = \frac{\pi}{2}$ s. When measurements depend on time we may think of them as functions of time, as in $y = f(t)$, and when we acknowledge that the measurements are noisy we might write

$$Y = f(t) + \varepsilon$$

where ε is a random variable representing noise and Y is written as a capital letter to emphasize that it, too, is a random variable. Given n observation pairs $(t_1, y_1), \ldots, (t_n, y_n)$ we might write

$$Y_i = f(t_i) + \varepsilon_i, \tag{18.1}$$

and this returns us to the usual nonparametric regression model of Chapter 15, in which the variables $\varepsilon_1, \ldots, \varepsilon_n$ are assumed independent. While at first glance (18.1) may seem natural, this kind of formulation does not yet go far enough in dealing with measurements that vary across time because it does not take account of the sequential nature of the argument t. In (18.1) the values $i = 1, 2, \ldots, n$ are generally no longer arbitrary labels but rather important and meaningful indications of temporal ordering with $t_1 < t_2 < \cdots < t_n$. If time matters, then even the noise variables $\varepsilon_1, \ldots, \varepsilon_n$ may

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_18,

be related to one another, and thus no longer independent. In this case, specialized methods can produce powerful results. The term *time series*, refers both to data collected across time and to the large body of theory and methods for analyzing such data.

Let us switch over to the general notation for random variables and write a theoretical sequence of measurements as $X_1, X_2, \ldots$, and a generic random variable in the sequence as X_t. Another way to say the X_t variables are dependent is that knowing $X_1, X_2, \ldots, X_{t-1}$ should allow us to predict, at least up to some uncertainty, X_t. Predictability plays an important role in time series analysis.

Example 2.2 (continued from p. 27) On p. 27 we displayed several EEG spectrograms taken under different stages of anesthesia. We noted earlier that both the roughly 10 Hz alpha rhythm and the 1–4 Hz delta rhythm are visible in the time series plot. In this scenario we can say a lot about the variation among the EEG values based on their sequence along time: in the time bin at time t the EEG voltage is likely to be close to that at time $t - 1$ and from the voltage in multiple time bins preceding time t we could produce a good prediction of the value at time t. □

The spectrograms in Example 2.2 display the rhythmic, wave-like features of the EEG signals contrasting them across phases of anesthesia. They do so by decomposing the signal into components of various frequencies, using one of the chief techniques of time series analysis. The decompositions are possible in this context because the EEGs may be described with relatively simple and standard time series models, but this is not true of all time series. The EEG series are, in a sense, very special because their variation occurs on a time scale that is substantially smaller than the observation interval. By contrast, if we go back to Fig. 1.5 of Example 1.6 we see another time series where the variation is on a longer time scale. The EPSC signal drops suddenly, and only once, shortly after the beginning of the series, then recovers slowly throughout the remainder of the series. In other words, the variation in the EPSC takes place on a time scale roughly equal to the length of the observation interval. Another way to put this is that the EEG at time x_t may be predicted reasonably well using only the preceding EEG values $x_{t-1}, x_{t-2}, \ldots, x_{t-h}$, going back h time bins, where h is some fairly small integer, but a prediction of the EPSC at x_t based on earlier observations would require nearly the entire previous series and still might not be very good. The most common time series methods, those we describe here, assume predictability on relatively short time scales.

So far we have said that the EEG at time x_t may be predicted using the preceding EEG values $x_{t-1}, x_{t-2}, \ldots, x_{t-h}$, but we did not specify which value of t we were referring to. Part of the point is that it doesn't much matter. In other words, it is possible to predict almost *any* x_t using the preceding h observations. (We say "almost" any x_t because we have to exclude the first few x_t observations, with $t \leq h$, where there do not exist h preceding observations from which to predict.) Furthermore, the formula we concoct to combine $x_{t-1}, x_{t-2}, \ldots, x_{t-h}$ in order to predict x_t may be chosen independently of t. This is a very strong kind of predictability, one that is stable across time, or *time-invariant*. The notion of time invariance is at the heart of time series analysis.

We now begin to formalize these ideas. Let X_t be the measurement of a series at time t, with $t = 1, \ldots, n$. Let $\mu_t = E(X_t)$ and $\Sigma_{ij} = Cov(X_i, X_j)$. As soon as we contemplate estimation of this mean vector and covariance matrix we are faced with a serious difficulty. For simplicity consider time t and the problem of estimating μ_t and $\sigma_t^2 = \Sigma_{tt}$. If we have many replications of the measurements at time t (as is usually the case, for example, with evoked potentials) we can collect all the observations across replications at time t and compute their sample mean and sample variance. However, if we have only one time series, and therefore one observation at t, we do not have a sample from which to compute the sample mean and variance. The only way to apply any kind of averaging is by using observations at other values of time. Thus, we can only get meaningful estimates of mean and covariance by making assumptions about the way X_t varies across time. Let us introduce a theoretical time series, or *discrete-time stochastic process* $\{X_t; t \in \mathcal{Z}\}$, $\mathcal{Z}$ being the set of all integers. We are now in a position to define the kinds of time invariance we will need. We say that the series X_t is *strictly stationary* if it is time-invariant in the sense that the joint distribution of each set of variables $\{X_t, X_{t+1}, \ldots, X_{t+h}\}$ is the same as that of the variables $\{X_s, X_{s+1}, \ldots, X_{s+h}\}$ for all t, s, h. Because the time index takes all possible integer values it is an abstraction (no experiment runs indefinitely far into the past and future) but it is an extremely useful one. A standard notation in the time series context is $\gamma(s, t) = \Sigma_{st}$. The function $\gamma(s, t)$ is called the *autocovariance function* and the *autocorrelation function* (ACF) is defined by

$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}}.$$

The prefix "auto," which signifies here that we are considering dependence of the time series on itself, is a hint that one might instead consider dependence across multiple time series, where we would instead have "cross-covariance" and "cross-correlation" functions (which we discuss in Section 18.5). A time series is said to be *weakly stationary* or *covariance stationary* if (i) μ_t is constant for all t and (ii) $\gamma(s, t)$ depends on s and t only through the magnitude of their difference $|s-t|$. This weaker sense of stationarity is all that is needed for many theoretical arguments. Under either form of stationarity we follow the convention of writing the autocovariance function in terms of a single argument, $h = t - s$, in the form $\gamma(h) = \gamma(t - h, t)$. Note that $\gamma(0) = V(X_t)$. It is not hard to show that $\gamma(0) \geq |\gamma(h)|$ for all h, and $\gamma(h) = \gamma(-h)$. In the stationary case the autocorrelation function becomes

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{18.2}$$

Illustration: The 3-point moving average process

$$X_t = \frac{1}{3}(U_t + U_{t-1} + U_{t-2})$$

where the U_t variables are independent, with $E(U_t) = 0$ and $V(U_i) = \sigma_U^2$, is a stationary process with autocovariance and autocorrelation

$$\begin{aligned}
\gamma(0) &= \frac{\sigma_U^2}{3} \\
\gamma(\pm 1) &= \frac{2\sigma_U^2}{9} \\
\rho(\pm 1) &= \frac{2}{3} \\
\gamma(\pm 2) &= \frac{\sigma_U^2}{9} \\
\rho(\pm 2) &= \frac{1}{3} \\
\gamma(\pm h) &= \rho(h) = 0, \text{ for } |h| \geq 3.
\end{aligned}$$

□

Having defined what it means for a process to be stationary, and also having defined the autocorrelation function, let us return to the distinction we were trying to draw between the EEG and EPSC time series. The EEG series may be modeled as stationary, and furthermore its variation is consistent with what is called *short-range dependence*. A theoretical time series exhibits short-range dependence when its correlation function $\rho(h)$ vanishes quickly as h becomes infinite. For the most common time series models the correlation function vanishes exponentially fast (i.e., there is a positive number a for which $\rho(h)e^{a|h|} \to 0$ as $h \to \pm\infty$). On the other hand, it is questionable whether one would want to model the EPSC time series as stationary and, if so, it would be necessary to use a model that assumes long-range dependence, where the correlation function dies out slowly as h becomes infinite. Time series analysis is concerned with variation across time while being cognizant of the role of stationarity. Much time series theory explicitly assumes stationarity. There is also considerable interest in non-stationary series, but the theoretical developments involve particular kinds of non-stationarity or modifications of methods that apply to stationary series. In contrast, nonparametric regression does not consider time-invariance arguments at all. In (18.1) the usual nonparametric assumption is $E(\varepsilon_t) = 0$, and we have $\mu_t = E(Y_t) = f(t)$. In other words, instead of a constant mean required by stationarity, the nonparametric problem focuses on the evolution of the mean as a function of time. In fact, many investigations involve a mix of these two possibilities: there is a stimulus that produces a time-varying mean component of the response, but there is also a wave-like time-invariant component of the response. From a practical point of view, it is very important to consider these components separately.

Example 15.2 (continued) For illustrative purposes we analyze here a small record of an LFP, which was recorded for 30 s (seconds) and sampled at 1 KHz as part of the experiment described briefly on p. 421. We confine our attention to the first second and the last second (each consisting of 1,000 observations), and will consider whether the signal appears consistent across these two time periods in the sense of containing

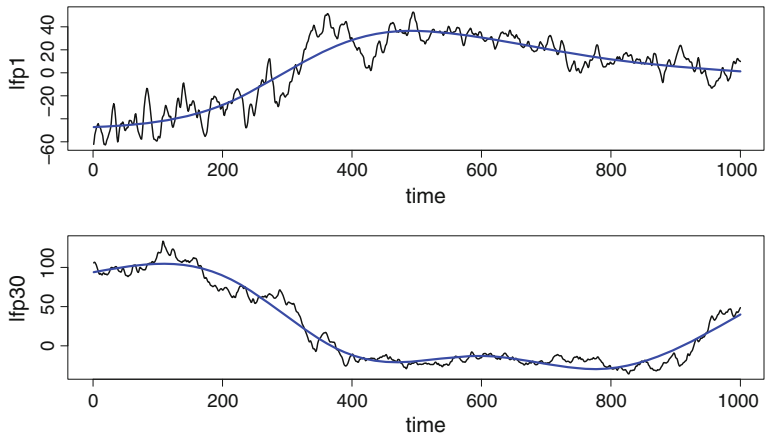

Fig. 18.1 LFP and smoothed versions representing slowly-varying trends. *Top* First second of average LFP. *Bottom* Last (thirtieth) second of average LFP. Smoothing was performed using regression splines with a small number of knots, as described on p. 421.

the same delta-wave content. Figure 18.1 displays these two time series, together with smoothed versions of the average LFP in these two periods. When we focus on a single second of observation time, the slow-wave activity shows up as slowly-varying mean signals, or trends, represented by the smoothed versions of the two LFP traces in the figure. Even though the slowly-varying trends could be considered roughly oscillatory on a longer time scale, at this time scale they can not be represented as oscillatory and are, instead, sources of long-range dependence or non-stationarity akin to that in Fig. 1.5. In order to capture the higher-frequency, stationary activity in these plots (with short-range dependence) we must first remove the slow trends. We analyze these data further in subsequent sections. □

In motivating stationarity we brought up the problem of estimating the mean and covariance functions, pointing out that in the absence of replications some assumptions must be made. Under stationarity the value of the constant mean $\mu_t = \mu$ may be estimated by the sample mean and an obvious estimator of the autocovariance function is the *sample autocovariance function*

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h} - \overline{x})(x_t - \overline{x}) \tag{18.3}$$

for $h = 0, 1, \ldots, n-1$ and then $\hat{\gamma}(-h) = \hat{\gamma}(h)$.We then have the *sample autocorrelation function* (sample ACF),

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \tag{18.4}$$

which is an estimator of the autocorrelation function (18.2).

In this chapter we provide an overview of key concepts in time series analysis. Section 18.2 describes the two major approaches to time series analysis. Section 18.3 gives some details on methods used to decompose time series into frequencies, as in Example 2.2. There are several important subtleties, and we discuss these as well. Section 18.4 discusses assessing uncertainty about frequency components, and Section 18.5 reviews the way these methods are adapted to assess dependence between pairs of simultaneous time series.

18.2 Time Domain and Frequency Domain

In discussing Example 2.2, on p. 514, we alluded to the decomposition of the signal into frequency-based components. In general, time series analysis relies on two complementary classes of methods. As the name indicates, *time domain* methods view the signal as a function of time and use statistical models that describe temporal dependence. *Frequency domain* methods decompose the signal into frequency-based components, and describe the relative contribution of these in making up the signal. In this section we provide a brief introduction to these two approaches, starting with frequency-based analysis. Here are two examples.

Example 18.1 Gamma oscillations in MEG during learning Cortical oscillatory activity in the gamma band (roughly 30–120 Hz) has been associated with many cognitive functions. Chaumon et al. (2009) used MEG imaging to investigate the role of gamma oscillations during unconscious learning. They used a paradigm in which subjects were to find the letter "T" within a set of distractors and determine its orientation. On some trials, which they called "predictive," the distractors were repeated and the location of the "T" remained the same. On other trials, which they called "nonpredictive," the distractors changed configurations and the location of the "T" changed. The subjects were shown many blocks containing 12 trials of each type. Although they remained unaware of the information provided by the configuration type, their reaction time decreased faster across blocks for the predictive trials than for the nonpredictive trials. The authors were interested in whether this unconscious learning was associated with changes in gamma band activity recorded with MEG. □

Example 18.2 fMRI BOLD signal and neural activity To investigate the neural basis of the fMRI BOLD signal, Logothetis et al. (2001) recorded local field potential (LFP) and multi-unit activity (MUA) together with fMRI from a region in primary visual cortex across 29 experimental sessions using 10 macaque monkeys. The stimulus involved rotating checkerboard patterns. In examining the relationship between LFP and BOLD, the authors focused on gamma band activity from 40 to 130 Hz. □

We now introduce another example, which we will use repeatedly in several parts of this chapter to demonstrate analytical techniques.

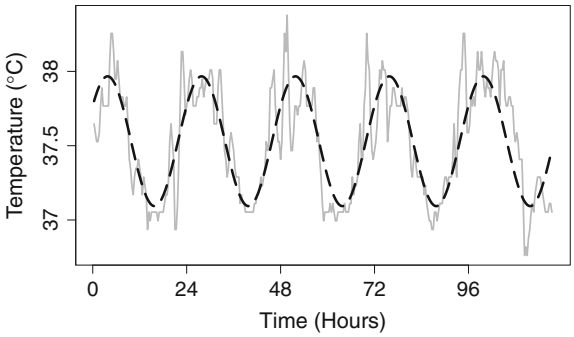

Fig. 18.2 Core temperature on a human subject, recordings taken every 20 min; *y*-axis in units of degrees Celsius (data shown with a *solid line*). Overlaid on the data is the least-squares fit of a cosine (shown with a *dashed line*), having a period of 24 h (hours).

Example 18.3 The circadian rhythm in core temperature Human physiology, like that of other organisms, has adapted to the cycle of changing environmental conditions, and resulting levels of activity, across each day and night. The result is a clear day/night pattern in hormone levels in the blood, and other indicators of the body's attempt to maintain homeostasis. In a study of methodology used to characterize circadian rhythms, Greenhouse et al. (1987) analyzed core temperatures of a human subject measured every 20 min across several days. Figure 18.2 displays the data. There is an obvious daily cycle in the temperatures. Figure 18.2 also shows a cosine curve, with a 24 h period, that has been fitted to the data using ordinary least-squares regression. □

The cosine curve in Fig. 18.2 was obtained by applying linear regression. We discussed fitting a cosine curve previously, in Example 12.6, in the context of directional tuning. Here, we begin with a cosine function $\cos(2\pi\omega_1 t)$, where ω_1 is the frequency (in cycles per unit time), then introduce an amplitude R_{amp}, an offset average value μ_{avg}, and a phase ϕ to put it in the functional form

$$f(t) = \mu_{avg} + R_{amp}\cos(2\pi(\omega_1 t - \phi)). \tag{18.5}$$

Details: The function $R_{amp}\cos(2\pi\omega_1 t)$ varies between a minimum of $-R_{amp}$ and a maximum of R_{amp}, and its average on $[0, 1]$ is 0. Adding the constant μ_{avg} makes the cosine oscillate around μ_{avg} with minimum $\mu_{avg} - R_{amp}$ and maximum $\mu_{avg} + R_{amp}$. It is also worth mentioning that the regression in Example 12.6 was set up slightly differently because the explanatory variable of interest was not time but rather the angle $\theta = 2\pi(\omega t - \phi)$. □

Based on (18.5) the statistical model for observations $y_1, \ldots, y_n$ at time points $t_1, \ldots, t_n$ is then

$$Y_i = \mu_{avg} + R_{amp} \cos(2\pi(\omega_1 t_i - \phi)) + \varepsilon_i$$

where, for the core temperature data, $\omega_1 = 1/72$ cycles per 20 min is the frequency corresponding to 1 cycle per day (a 24 h period). To simplify fitting, this model may be converted to a linear form, i.e., a form that is linear in the unknown parameters. Using

$$\cos(u - v) = \cos u \cos v + \sin u \sin v \tag{18.6}$$

with $u = 2\pi\omega_1 t_i$ and $v = 2\pi\phi$ we have

$$R_{amp} \cos(2\pi(\omega_1 t_i - \phi)) = A \cos(2\pi\omega_1 t_i) + B \sin(2\pi\omega_1 t_i) \tag{18.7}$$

where $A = R_{amp} \cos(2\pi\phi)$ and $B = R_{amp} \sin(2\pi\phi)$. We may therefore rewrite the statistical model as

$$Y_i = \mu_{avg} + A \cos(2\pi\omega_1 t_i) + B \sin(2\pi\omega_1 t_i) + \varepsilon_i, \tag{18.8}$$

which has the form of a linear regression model, and may be fitted using ordinary linear regression. Specifically, we do the following:

1. Assume the data $(t_1, \ldots, t_n)$ and $(y_1, \ldots, y_n)$ are in respective variables `time` and `temp`.
2. Define

$$\texttt{cosine} = cos(2\pi\texttt{time}/72)$$
$$\texttt{sine} = sin(2\pi\texttt{time}/72).$$

3. Regress `temp` on `cosine` and `sine`.

For future reference we note that the squared amplitude of the cosine function in (18.7) is

$$R_{amp}^2 = A^2 + B^2 \tag{18.9}$$

and the phase is

$$\phi = \frac{1}{2\pi} \arctan(\frac{B}{A}). \tag{18.10}$$

In the core temperature data of Example 18.3 there is a clear, dominant periodicity, which is easily described by a cosine function using linear regression. We may do a bit better if we allow the fitted curve to flatten out a little, compared to the cosine function. This is accomplished by introducing a second frequency, $\omega_2 = 2\omega_1$ to produce the model

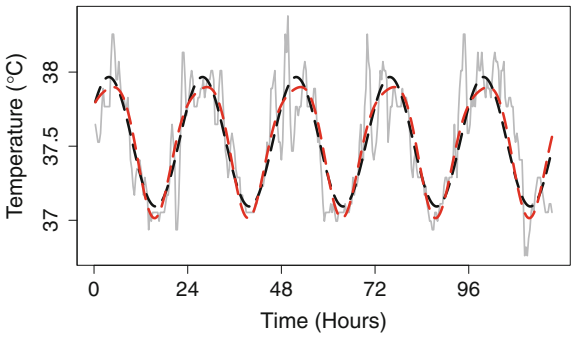

Fig. 18.3 Plot of core temperature, as in Fig. 18.2, together with fit of (18.8), shown in the *black dashed line*, using the fundamental frequency $\omega_1 = 1/72$ (one oscillation every 72 data points, i.e., every 24 h), and fit of (18.11), shown in *red dashed line*. The latter improves the fit somewhat in the peaks and troughs.

$$Y_i = \mu_{avg} + A_1 \cos(2\pi\omega_1 t_i) + B_1 \sin(2\pi\omega_1 t_i) + A_2 \cos(2\pi\omega_2 t_i) + B_2 \sin(2\pi\omega_2 t_i) + \varepsilon_i. \tag{18.11}$$

Example 18.3 (continued from p. 519) Least-squares regression using model (18.11) yields a highly significant effect for the second cosine–sine pair ($p < 10^{-6}$) and Fig. 18.3 displays a modest improvement in fit. □

Model (18.8) was modified in (18.11) by introducing the additiona1 cosine–sine pair corresponding to the frequency ω_2. In principle this process could be continued by introducing frequencies of the form $\omega_k = k\omega_1$ for $k = 3, 4, \ldots$. Here, ω_1 is called the *fundamental frequency*, the additional frequencies ω_k are *harmonic frequencies*, and the resulting regression model is often called *harmonic regression*. For the core temperature data it turns out that $k = 2$ is a satisfactory choice (see Greenhouse et al. 1987) but, in general, one might use linear regression to fit many harmonics and ask how much variation in the data is explained by each cosine–sine pair. For this purpose one might use contributions to R^2, which is the germ of the idea behind one of the main topics in time series, *spectral analysis*. Spectral analysis can be a very effective way to describe wave-like behavior, as seen in the EEG signals of Example 2.2.

18.2.1 *Fourier analysis is one of the great achievements of mathematical science.*

Spectral analysis, otherwise known as Fourier analysis,[1] decomposes an oscillatory signal into trigonometric components. Because many physical phenomena may be described by applying this technique (and it is at the heart of quantum mechanics), the physicist Richard Feynman called[2] the ability to create such decompositions "probably the most far-reaching principle in mathematical physics." From a practical point of view, our world has been changed dramatically by applications of Fourier analysis.

The argument may be broken into several steps.

1. The signal may be represented by a smoothly varying function $f(t)$, for values of t (usually thought of as time) in a suitable interval $[a, b]$, which, for convenience, we may take[3] to be $[0, 1]$.
2. If we pick n values of t spaced evenly across the interval, say, $t_1, t_2, \ldots, t_n$, then $f(t)$ may be determined to a close approximation by its values at these points, i.e., by $f(t_1), f(t_2), \ldots, f(t_n)$, for sufficiently large n. That is, if $f(t)$ varies smoothly then, for practical purposes, interpolation will suffice to reproduce it from its values $f(t_1), f(t_2), \ldots, f(t_n)$.
3. The cosine and sine functions $\cos(2\pi t)$ and $\sin(2\pi t)$ are periodic, completing a single cycle on $[0, 1]$, and thus having frequency 1 (per unit time). This is the fundamental frequency and the corresponding harmonic frequencies are $2, 3, 4, \ldots$. The cosine and sine functions at harmonic frequencies may be considered primitive functions—meaning building blocks of other functions—on $[0, 1]$. When we evaluate a sufficiently large number of primitive functions at $t_1, t_2, \ldots, t_n$, and take linear combinations of them, we are able to reproduce $f(t)$ at the values $t_1, t_2, \ldots, t_n$, which, according to step 2, suffices for reconstructing $f(t)$ throughout $[0, 1]$. That is, we can decompose $f(t)$ into harmonic trigonometric components. This has the potential to provide the appealing interpretation that $f(t)$ is "made up" of particular harmonic components in particular amounts, according to the linear combinations.
4. In order to have this interpretation make sense, the "particular amount" of each component given by the decomposition in step 3 must not depend on the number of components being considered, for that would make the interpretation self-contradictory. In non-orthogonal decompositions the amount, or weight, given to a particular component *does* depend on the other components being considered, but for orthogonal decompositions it does not. (See the discussion in Chapter 12,

[1] The term "spectral analysis" sometimes connotes statistical analysis, rather than purely mathematical analysis, but for now we are ignoring any noise considerations.

[2] Feynman et al. (1963 Volume I, p. 49–1).

[3] The argument we sketch here makes the most sense for functions that are periodic on $[0, 1]$, meaning that they satisfy $f(0) = f(1)$. In Section 18.3.6 we discuss what happens when this condition fails to hold.

p. 351.) Harmonic trigonometric functions are orthogonal, so the interpretation is internally consistent.

These steps all involved major conceptual breakthroughs for mathematics.[4] Taken together they suggest that a signal represented by a smoothly varying function $f(t)$ may be decomposed into cosine and sine harmonic components. This is what Fourier analysis accomplishes.

To be a little more specific, suppose that $f(t)$ is a function on the interval [0, 1] and let us consider time points $t_j = \frac{j}{n}$ for $j = 1, 2, \ldots, n$ where, for simplicity, we assume n is odd so that $(n-1)/2$ is an integer. If we evaluate $f(t)$ at the time points t_j we get an n-dimensional vector

$$y = (f(t_1), f(t_2), \ldots, f(t_n))^T. \tag{18.12}$$

Now define the harmonic trigonometric functions $f_k(t) = \cos(2\pi kt)$ and $g_k(t) = \sin(2\pi kt)$, for $k = 1, 2, \ldots, (n-1)/2$. By evaluating these functions at $t_1, t_2, \ldots, t_n$ we form vectors $f_k = (f_k(t_1), f_k(t_2), \ldots, f_k(t_n))^T$ and $g_k = (g_k(t_1), g_k(t_2), \ldots, g_k(t_n))^T$ and, it turns out, the collection of vectors

$$1_{vec}, f_1, \ldots, f_{(n-1)/2}, g_1, \ldots, g_{(n-1)/2}$$

are orthogonal, where $1_{vec} = (1, 1, \ldots, 1)^T$. (This follows from straightforward algebraic manipulation, together with properties of sines and cosines, see Bloomfield 2000). They therefore form an orthogonal basis for R^n (see Section A.9), which means that any vector y, such as in (18.12), may be written in the form

$$\begin{aligned} y = \mu_{avg} 1_{vec} + A_1 f_1 + \cdots + A_{(n-1)/2} f_{(n-1)/2} \\ + B_1 g_1 + \cdots + B_{(n-1)/2} g_{(n-1)/2}. \end{aligned} \tag{18.13}$$

If we define

$$\begin{aligned} p_n(t) = \mu_{avg} + A_1 f_1(t) + \cdots + A_{(n-1)/2} f_{(n-1)/2}(t) \\ + B_1 g_1(t) + \cdots + B_{(n-1)/2} g_{(n-1)/2}(t) \end{aligned} \tag{18.14}$$

[4] The first requires the notion of function, which emerged roughly in the 1700s, especially in the work of Euler (the notation $f(x)$ apparently being introduced in 1735). The second may be considered intuitively obvious, but a detailed rigorous understanding of the situation did not come until the 1800s, particularly in the work of Cauchy (represented by a publication in 1821) and Weierstrass (in 1872). The notion of harmonics was one of the greatest discoveries of antiquity, and is associated with Pythagoras. The third and fourth steps emerged in work by D'Alembert in the mid-1700s, and by Fourier in 1807. Along the way, representations using complex numbers were used by Euler (his famous formula, given below, appeared in 1748), but they were considered quite mysterious until their geometric interpretation was given by Wessel, Argand, and Gauss, the latter in an influential 1832 exposition. A complete understanding of basic Fourier analysis was achieved by the early 1900s with the development of the Lebesgue integral. Recommended general discussions may be found in Courant and Robbins (1996), Lanczos (1966), and Hawkins (2001).

then we have

$$f(t) = p_n(t) \tag{18.15}$$

for $t = t_j$ for $j = 1, \ldots, n$ and, by interpolation we get the approximation

$$f(t) \approx p_n(t), \tag{18.16}$$

for all $t \in [0, 1]$, which may be considered a decomposition of $f(t)$ into trigonometric components based on the n data values $f(t_1), f(t_2), \ldots, f(t_n)$. The constants $\mu_{avg}, A_1, \ldots, A_k, B_1, \ldots, B_k$ are called the *Fourier coefficients* of $f(t)$. By analogy with the approximate representation of functions by polynomials, the expression $p_n(t)$ in (18.14) is often called a *trigonometric polynomial*. With reference to (18.7), we may say that $A_k f_k$ and $B_k g_k$ together determine the component of $f(t)$ having frequency k.

We now consider the magnitude of y. Using the orthogonality of the component vectors, Eq. (18.13) gives

$$\begin{aligned} ||y||^2 = ||\mu_{avg} 1_{vec}||^2 + ||A_1 f_1||^2 + \cdots + ||A_{(n-1)/2} f_{(n-1)/2}||^2 \\ + ||B_1 g_1||^2 + \cdots + ||B_{(n-1)/2} g_{(n-1)/2}||^2 \end{aligned}$$

and re-writing this we get

$$||y||^2 = ||\mu_{avg} 1_{vec}||^2 + \sum_{k=1}^{(n-1)/2} ||A_k f_k||^2 + ||B_k g_k||^2. \tag{18.17}$$

Equation (18.17) decomposes the squared magnitude of y into magnitudes corresponding to its trigonometric components. Using (18.15) we say that any vector of function evaluations may be written in terms of the trigonometric basis vectors, and its squared length is equal to the sum of squares of its trigonometric components. From (18.16) we see that an analogous statement should hold for functions on [0,1].

We can also use (18.17) to give a nice interpretation of the Fourier decomposition in terms of least-squares regression. We begin by considering (18.13) to be a noiseless regression equation. If we regress y on the variables $f_1, \ldots, f_{(n-1)/2}, g_1, \ldots, g_{(n-1)/2}$ we obtain the coefficients $A_1, B_1, \ldots, A_{(n-1)/2}, B_{(n-1)/2}$. Furthermore, because the trigonometric vectors are orthogonal, the coefficient found by regressing y on all the variables $f_1, \ldots, f_{(n-1)/2}, g_1, \ldots, g_{(n-1)/2}$ is the same as the coefficient of f_k (or g_k) in the regression of y on f_k (or g_k) alone. Thus, it makes sense to say that $A_k f_k$ and $B_k g_k$ together uniquely represent the component of y corresponding to frequency k. Because (18.13) provides an exact fit of y, if we regress y on all the variables $f_1, \ldots, f_{(n-1)/2}, g_1, \ldots, g_{(n-1)/2}$ we get $R^2 = 1$. The regression of y on 1_{vec} gives $\mu_{avg} = \bar{y}$ and $\mu_{avg} 1_{vec} = \bar{y} 1_{vec}$ has squared length $n\bar{y}^2$ so that (18.17) may be rewritten in terms of the total sum of squares

$$||y||^2 - n\bar{y}^2 = \sum_{k=1}^{(n-1)/2} ||A_k f_k||^2 + ||B_k g_k||^2$$

and, dividing both sides by $||y||^2 - n\bar{y}^2$ while using $R^2 = 1$ we get

$$R^2 = \sum_{k=1}^{(n-1)/2} R_k^2, \tag{18.18}$$

where

$$R_k^2 = \frac{||A_k f_k||^2 + ||B_k g_k||^2}{||y||^2 - n\bar{y}^2}, \tag{18.19}$$

which is the proportion of variation in y, and therefore $f(t)$, at frequency k. In other words, this trigonometric representation, using sines and cosines at harmonic frequencies, has the wonderful property that it decomposes the variability of the function $f(t)$ into frequency-based components, the magnitudes of which add to the total variation in $f(t)$. The decomposition (18.18) into components (18.19) is the starting point for spectral analysis.

18.2.2 The periodogram is both a scaled representation of contributions to R^2 from harmonic regression and a scaled power function associated with the discrete Fourier transform of a data set.

We now apply to data $x_1, x_2, \ldots, x_n$ the spectral analysis decomposition discussed in Section 18.2.1. We write $y = (x_1, x_2, \ldots, x_n)$ and use (18.13). We may get a rough idea of the relative contributions to the variability in the data due to the harmonic frequency components simply by plotting R_k^2, defined in Eq. (18.19), against the frequency k. A scaled plot of R_k^2 against frequency is known as the *periodogram*, with the precise definition appearing in Eq. (18.25). The periodogram, together with some important modifications of it, is enormously useful in practice.

Example 18.3 (continued from 521) The periodogram for the core temperature data (introduced on p. 519) is shown in Fig. 18.4. Note the dominant contribution to R^2 corresponding to the roughly daily cycle. □

The coefficients A_k and B_k in (18.13) and (18.19) turn out to be

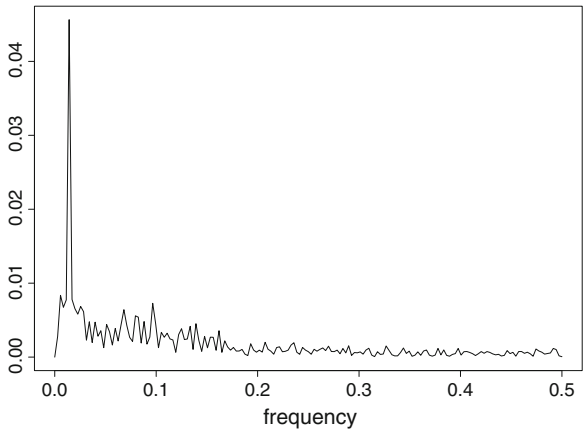

Fig. 18.4 Periodogram of core body temperature data. There is a peak at the frequency representing, very nearly, daily oscillation and this peak is much higher than the remainder of the periodogram.

$$\mu_{avg} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$A_k = \frac{2}{n}\sum_{j=1}^{n} x_j \cos(2k\pi j/n) \tag{18.20}$$

$$B_k = \frac{2}{n}\sum_{j=1}^{n} x_j \sin(2k\pi j/n) \tag{18.21}$$

for $k = 1, \ldots, (n-1)/2$. Because the cosine and sine terms always occur in pairs, it is often simpler to represent expressions (18.20) and (18.21) instead in exponential form via Euler's formula,

$$e^{i\theta} = \cos\theta + i\sin\theta, \tag{18.22}$$

which is also Eq. (A.31) in the Appendix. This formula is extremely helpful in Fourier analysis. On the one hand, it provides a kind of "book-keeping" of cosine and sine terms within a complex exponential while, on the other hand, it simplifies many manipulations because multiplication becomes addition of exponents. Applying Euler's formula (18.22), we have

$$\sum_{j=1}^{n} x_j \cos(2k\pi j/n) + i\sum_{j=1}^{n} x_j \sin(2k\pi j/n) = \sum_{j=1}^{n} x_j e^{2k\pi ij/n}$$

and then (18.20) and (18.21) may be replaced with

$$A_k + iB_k = \frac{2}{n}\sum_{j=1}^{n} x_j e^{2\pi ikj/n}$$

for $k = 1, \ldots, (n-1)/2$. By convention the equivalent form

$$A_k - iB_k = \frac{2}{n}\sum_{j=1}^{n} x_j e^{-2\pi ikj/n} \tag{18.23}$$

for $k = 1, \ldots, (n-1)/2$, is used instead. Aside from the multiplier, the right-hand side of (18.23) is the *discrete Fourier transform*. Specifically, for a data sequence $x_1, \ldots, x_n$, we let

$$\omega_j = j/n$$

denote frequency, for $j = 0, \ldots, n-1$. Then the discrete Fourier transform (DFT) is given by

$$d(\omega_j) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t e^{-2\pi i\omega_j t} \tag{18.24}$$

and the periodogram is

$$I(\omega_j) = |d(\omega_j)|^2. \tag{18.25}$$

Here we are interested only in the first $(n-1)/2$ frequencies (if n is odd; otherwise, the first $n/2$ frequencies). From (18.23) we have $d(\omega_j) = \frac{\sqrt{n}}{2}(A_j - iB_j)$, and because $||A_j + iB_j||^2 = A_j^2 + B_j^2$, we get

$$|d(\omega_j)|^2 = \frac{n}{4}(A_j^2 + B_j^2).$$

According to the definition in Eq. (18.19), $A_j^2 + B_j^2$ is proportional to R_j^2 (meaning that the constant multiple does not depend on j) and so we arrive at

$$I(\omega_j) \propto R_j^2,$$

which justifies the interpretation of the periodogram we gave on p. 525. Algorithms for computing the DFT are based on the *fast Fourier transform*, which had a huge impact on signal processing following a 1965 publication of the method by James Cooley and John Tukey. The DFT also has an interpretation using the terminology of signal processing. If we return to the interpretation of $x_1, \ldots, x_n$ as function values $f(t_1), \ldots, f(t_n)$ as in Eq. (18.16), then $||y||^2 = ||(f(t_1), \ldots, f(t_n))||^2$ is

(approximately, by (18.16)), the *power* of the function $f(t)$ on $[0, 1]$ and $I(\omega_j)$ is (approximately[5]) proportional to the power of $f(t)$ at frequency ω_j.

Unfortunately, in spectral analysis, the various notational conventions that get invoked are not consistent across authors. In particular, we have introduced the *Fourier frequencies* $\omega_j = j/n$ for $j = 0, 1, \ldots, n-1$. Because we divided the harmonic integers by n, the Fourier frequencies are restricted to the interval $[0, 1]$. In fact, because we use only the first $(n-1)/2$ frequencies (if n is odd and the first $n/2$ frequencies if n is even) they are restricted to $[0, \frac{1}{2}]$. In some texts $j = 1, \ldots, n$ is used. Furthermore, the multiplier of the complex exponential sum we used in (18.24) to define the DFT is also not universal. For some purposes one must pay attention to the definitions being used by a particular book or piece of software.

It is also important to notice that the Fourier frequencies we have defined on $[0, 1]$ (or $[0, \frac{1}{2}]$) have units of cycles per observation. If the units of time (such as seconds) involve m observations (such as m observations per second) then $m\omega_j$ will be in cycles per unit time. See the legend to Fig. 18.6.

With some additional mathematics, these concepts carry over to infinite-dimensional vector spaces with inner products. The infinite-dimensional representation is analogous: periodic functions (actually, square-integrable periodic functions) form a vector space for which the harmonic trigonometric functions provide an orthogonal basis. The resulting infinite-dimensional harmonic trigonometric expansion is called a Fourier expansion, and the coefficients are the Fourier coefficients.[6] In mathematics, Fourier analysis concerns infinite-dimensional function spaces, but in statistics and engineering these terms are also applied, as here, to the finite-dimensional setting involving data.

The DFT and its inverse are finite versions of the usual Fourier transform and its inverse, which is used extensively in mathematical analysis and signal processing, including theoretical studies of stationary time series. We discuss stationary time series in Section 18.3.1. We also discuss, in the remainder of Section 18.3, several practical issues that arise when using and interpreting the periodogram. We have already mentioned one of these in our discussion of Example 15.2.

Example 15.2 (continued from p. 421) Fig. 18.5 displays the log periodogram for the first second of average LFP, which was plotted previously in the top portion of Fig. 18.1. In Section 18.3.6 we explain why the log transform is used. The point, for now, is that the periodogram does not have a peak corresponding to delta range or other frequencies. This is quite common in series that have slowly varying trends. In contrast, after we remove the trends seen in Fig. 18.1 from the two series (by subtraction, so that the residuals are analyzed instead) the peaks of interest become visible, as seen in Fig. 18.6. □

---

[5] The approximation becomes exact when $f(t)$ is periodic, $f(t)^2$ has a finite integral, and the expansion involves all of the infinitely many harmonics.

[6] With appropriate mathematics (especially the theory of Lebesgue integration) it may be shown that every square-integrable function on [0,1] may be represented, equivalently, by its set of Fourier coefficients, and its integrated squared magnitude is equal to the sum of squares of the coefficients.

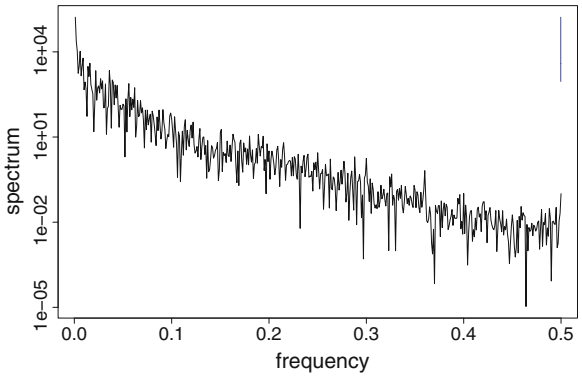

Fig. 18.5 Log periodogram for the first second of average LFP data in Example 15.2.

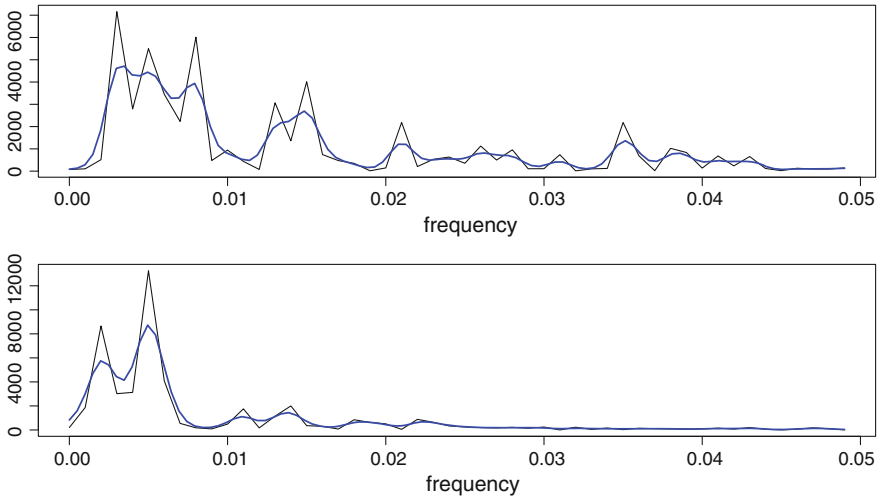

Fig. 18.6 Periodograms and smoothed periodograms from LFP detrended series. *Top* First second of average LFP. *Bottom* Last second of average LFP. Notice that the frequency units are cycles per observation. To get cycles per second (Hz) we must multiply by the number of observations per second, which is 1,000. Thus, the first peak of power in these plots is centered roughly at .005, which corresponds to 5 Hz.

The contrast between Figs. 18.5 and 18.6 illustrates the importance of checking time series for slowly-varying trends, and removing them from the data before performing spectral analysis. This is often called *detrending* the series.

18.2.3 Autoregressive models may be fitted by lagged regression.

As we have indicated, time series are special among kinds of data because of their serial dependence, e.g., the value of X_t is likely to depend on the value of X_{t-1}. The simplest form of dependence is linear dependence, as in the *autoregressive model* given by

$$X_t = \phi X_{t-1} + \epsilon_t.$$

This says that X_t has a regression on X_{t-1}, and otherwise is determined by noise. For consistency with later notation let us write the noise variables as[7] W_t:

$$X_t = \phi X_{t-1} + W_t. \tag{18.26}$$

The natural generalization,

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + W_t, \tag{18.27}$$

is called an *autoregressive model of order p*, written $AR(p)$. The W_t variables are usually assumed to be i.i.d. $N(0, \sigma^2)$. Model (18.26) then becomes the standard $AR(1)$ model. The parameter ϕ in (18.26) is usually assumed to satisfy $|\phi| < 1$, and analogous, but more complicated constraints are assumed for the parameters in (18.27).

> *Some details:* It may be shown that the case of (18.26) with $\phi = 1$, known as a *random walk* model (confer p. 126), is non-stationary. This makes it unsuitable for most auto-regressive modeling methodology. $\phi = -1$ is also non-stationary. The case $|\phi| > 1$ is somewhat more subtle, and it turns out to be non-causal in the sense that X_t depends on W_{t+i} for $i > 0$. The condition $|\phi| < 1$ restricts the $AR(1)$ so that it is neither non-stationary nor non-causal. Additional explanation is provided in time series texts such as Shumway and Stoffer (2006). □

Because the $AR(p)$ model (18.27) has the form of an ordinary linear regression model, we may apply it to data $x = (x_1, \ldots, x_n)$ using ordinary least squares regression after first defining suitable *lagged* variables. In the simplest case, with $p = 1$, we begin by defining a pair of variables y and x_{B1}, each of length $n - 1$:

[7] W is often used to represent time series noise out of deference to Norbert Wiener, a major figure in the development of time series theory.

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{pmatrix} = \begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}$$

$$x_{B1} = \begin{pmatrix} x_{B1,1} \\ x_{B1,2} \\ \vdots \\ x_{B1,n-1} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \end{pmatrix}.$$

We use the subscript $B1$ for "back 1" because $x_{B1,t} = y_{t-1}$ (x_{B1} "lags" behind y and is often called the lag-1 version of y). We then fit the $AR(1)$ model (18.26) by performing least-squares regression of y on x_{B1}, without using an intercept. The resulting regression coefficient becomes the estimate $\hat{\phi}$ of the $AR(1)$ parameter ϕ.

More generally, to fit an $AR(p)$ model using ordinary least squares we begin by defining $y_{n-p} = x_n$, $y_{n-p-1} = x_{n-1}, \ldots, y_1 = x_{n-p+1}$ and then also defining x_{B1} to be the lag-1 version of y, x_{B2} to be the analogous lag-2 version of y, etc., until we reach x_{Bp}. We then regress y on the variables $x_{B1}, x_{B2}, \ldots, x_{Bp}$.

It is often unclear what order p should be used in the $AR(p)$ model. Sometimes the model selection criteria AIC or BIC are used (see Section 11.1.6). One simple idea is to pick a relatively large value of p, perform the regression, and examine the coefficients from first to last to see when they become non-significant. A similar idea is to use the sample autocorrelation function (ACF), which was defined in (18.4), and the partial autocorrelation function (PACF). Under fairly general conditions, if $X_1, \ldots, X_n$ are i.i.d. with finite variance, and the sample ACF is computed for the random variables X_t, then

$$\sqrt{n}\hat{\rho}(h) \xrightarrow{\mathcal{D}} N(0, 1).$$

Based on this result, the sample ACF is usually plotted together with horizontal lines drawn at $\pm 2/\sqrt{n}$. If the series were i.i.d., then roughly 95 % of the sample autocorrelation coefficients would fall between theses lines. The ACF coefficients outside these lines are considered significant, with $p < .05$, approximately, for large n. This is illustrated for Example 18.3 below.

A difficulty with the sample ACF plot, however, is that it is based on the individual correlations of each lagged variable with the original data. That is, its results come from many single-variable regressions, of y on x_{Bk} for various values of k. A significant regression of y on x_{B2}, for example, could be based on the correlation between x_{B1} and x_{B2} and may reflect a relationship between y and x_{B1}. An alternative is to perform the multiple linear regression of y on *both* x_{B1} and x_{B2} and examine whether the coefficient of x_{B2} is significant, which assesses the explanatory power of x_{B2} after including x_{B1} in the model. The sample PACF at lag h is the sample partial correlation, defined by (5.22), between the time series and itself at lag-h given the lag-1 through lag-$h-1$ series. The lag-h partial autocorrelation coefficient measures

the lag-h correlation after adjusting for the effects of lags 1 through $h-1$, adjusting as in multiple linear regression. It may be computed as the normalized lag-h regression coefficient found from an $AR(h)$ model, normalized by dividing the series by the sample variance $\hat{\gamma}(0)$.

A detail: Suppose X_t is a mean-zero stationary Gaussian series. Then the theoretical PACF is given by $\phi_{11} = Cor(X_t, X_{t+1})$ and for $h \geq 2$,

$$\phi_{hh} = Cor(X_t, X_{t+h} | X_{t+1}, X_{t+2}, \ldots, X_{t+h-1}).$$

More generally, for any mean-zero stationary series let $X_t^{h-1} = \sum_{j=1}^{h-1} \beta_j X_{t-j}$ where the coefficients $\beta_1, \ldots, \beta_{h-1}$ minimize $E((X_t - \sum_{j=1}^{h-1} \alpha_j X_{t-j})^2)$ over the α_js. Then, for $h \geq 2$,

$$\phi_{hh} = Cor(X_t - X_t^{h-1}, X_{t+h} - X_{t+h}^{h-1}). \qquad \square$$

Once again, using large-sample theory, horizontal lines may be drawn on the sample PACF to indicate where the coefficients stop being significant. The sample PACF is often used to choose the order of the autoregressive model.

Example 18.3 (continued from p. 525) Let us consider an $AR(p)$ model for the core temperature residuals following the cosine regression reported on p. 519, and then detrending (using BARS, see Section 15.2.6). We take $p = 22$. The fitted coefficients are plotted in Fig. 18.7. Here is an abbreviated table of coefficients:

Variable	Coefficient	Std. Err.	t-ratio	p-value
x_{B1}	.906	.057	15.9	$< 10^{-15}$
x_{B2}	−.205	.077	−2.7	.008
x_{B3}	−.147	.078	−1.9	.06
x_{B4}	.005	.078	.1	.95
x_{B5}	−0.154	.078	−1.9	.05
x_{B6}	.115	.078	.9	.35
...				
x_{B21}	−.031	.076	−.4	.69
x_{B22}	.011	.057	−.2	.84

Only the first two lagged variables have large t statistics, so it appears that only the first two lagged variables are likely to be helpful in predicting the response variable. Also shown in Fig. 18.7 is the sample ACF, together with horizontal lines drawn at $\pm 2/\sqrt{n}$. The PACF in Fig. 18.7 has nonzero lag-1 and lag-2 coefficients, but the remaining coefficients are not distinctly different from zero relative to statistical uncertainty. Using an $AR(2)$ fit to the residuals added to the fitted 24 h cycle produces the overall fit to the temperature data shown in Fig. 18.8. $\square$

In general, autoregressive models may be fit by maximum likelihood. We now connect ML estimation with lagged least-squares regression (p. 531), by writing down the

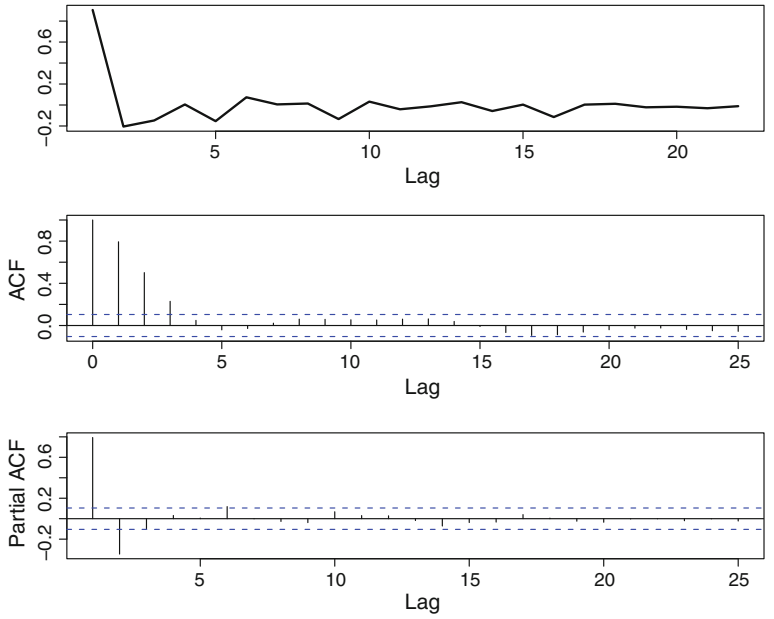

Fig. 18.7 Autoregressive model of order $p = 22$ for core temperature residuals. *Top* Coefficients $\hat{\phi}_i$ as a function of lag i. *Middle* The sample autocorrelation function. *Bottom* The sample partial autocorrelation function.

likelihood function for the $AR(1)$ model, assuming X_t is Gaussian with mean zero and $|\phi| < 1$. We have $X_1 \sim N(0, \sigma_1^2)$ where

$$\sigma_1^2 = \sigma_W^2/(1-\phi^2). \tag{18.28}$$

We also have $X_t|X_{t-1} = x_{t-1} \sim N(\phi x_{t-1}, \sigma_W^2)$ for $t = 2, \ldots, n$. The joint pdf is

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) &= f_{X_1}(x_1) f_{X_2|X_1}(x_2|X_1 = x_1) \cdots f_{X_n|X_{n-1}}(x_n|X_{n-1} = x_{n-1}) \\ &= \frac{1}{\sigma_1} f_Z(\frac{x_1}{\sigma_1}) \prod_{t=2}^{n} \frac{1}{\sigma_W} f_Z(\frac{x_t - \phi x_{t-1}}{\sigma_W}) \end{aligned}$$

where $f_Z(z)$ is the $N(0, 1)$ pdf. The factors in the product above may be written

$$\begin{aligned} \frac{1}{\sigma_W} f_Z(\frac{x_t - \phi x_{t-1}}{\sigma_W}) &= \frac{1}{\sqrt{2\pi}\sigma_W} \exp(-\frac{(x_t - \phi x_{t-1})^2}{2\sigma_W^2}) \\ &= \frac{1}{\sqrt{2\pi}\sigma_W} \exp(-\frac{(y_{t-1} - \phi x_{B1,t-1})^2}{2\sigma_W^2}). \end{aligned}$$

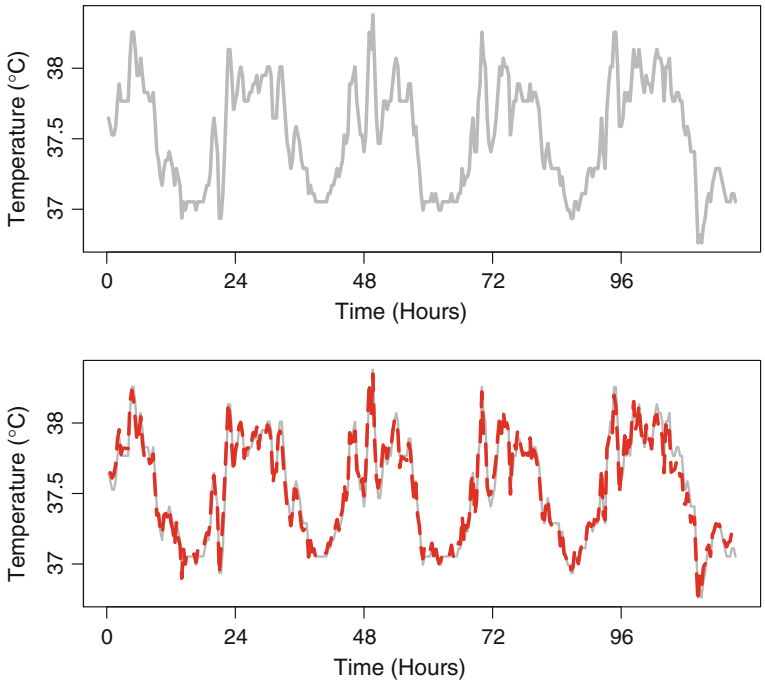

Fig. 18.8 Core temperature data together with fit. *Top* plot of temperature data. *Bottom* Plot of temperature data together with fit (*in red*) based on the sum of an $AR(2)$ fit to residuals and the fitted 24 h cycle.

This final form of each factor is the same as would appear in the likelihood for the regression of y on x_{B1}, with no intercept. Thus, if we ignore x_1, maximizing the likelihood $L(\phi, \sigma_W)$ amounts to solving the ordinary least-squares problem in the regression of y on x_{B1}. This maximization is called *conditional maximum likelihood* because we act as if the distribution of X_1 is given, i.e., it involves no unknown parameters. Because σ_1 in (18.28) is a function of ϕ and σ_W, when we include the factor due to X_1, which is $f_Z(x_1/\sigma_1)/\sigma_1$, the maximization problem changes and it is no longer solvable by least squares. Thus, the MLE must be found by an iterative method, but it is likely to be very close to the conditional MLE. Similar considerations hold also for $AR(p)$ models: the likelihood is nonlinear in the autoregressive parameters, but if we condition on the first p values then ML estimation reduces to ordinary least squares lagged regression. Statistical software for fitting autoregressive models typically either uses ML estimation, or a method that is very nearly equivalent. (The Kalman filter, described in Section 16.2.5, is sometimes used to obtain ML estimates in time series models.) For large samples, the fitted coefficients are essentially the same as those obtained using lagged regression.

The fit to the core temperature data in the bottom panel of Fig. 18.8 combines the fitted 24 h cycle and the $AR(2)$ fit to the residuals. This is an example of *regression*

with time series errors. As mentioned on p. 346, a general approach to regression with time series errors may be based on weighted least squares. Specifically, the model (12.64) may be used with the variance matrix R defined by the $AR(p)$ process and a fit, together with confidence intervals and significance tests, may be obtained[8] from the following steps:

1. Fit the regression variables X to the response variable Y using ordinary least squares;
2. Fit an $AR(p)$ model to the residuals from step 1;
3. Re-fit the regression variables X to the response variable Y using weighted least squares (see p. 345), based on the estimated R matrix found from the fitted auto-regressive model in step 2.

In practice, steps 1-3 may be adequate but, in addition, steps 2 and 3 could be iterated, or ML estimation could be applied once the $AR(p)$ model is determined in Step 2 (e.g., Greenhouse et al. 1987). Statistical software for regression with time series errors is usually based on ML estimation.

18.3 The Periodogram for Stationary Processes

18.3.1 The periodogram may be considered an estimate of the spectral density function.

The DFT is relatively easy to use without thinking about its continuous analogue. However, to understand the way the DFT behaves, and to derive statistical assessments of uncertainty, we must consider the analogous object defined for a theoretical stationary time series $\{X_t; t \in \mathcal{Z}\}$.

Assume $\sigma_t^2 = V(X_t) < \infty$ and let $\mu_t = E(X_t)$. Recall that the autocovariance function is given by

$$\gamma(h) = E((X_t - \mu_t)(X_{t+h} - \mu_{t+h})).$$

Under the summability condition

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty \tag{18.29}$$

general results give the existence of a *spectral density function* $f(\omega)$ for which

[8] The fit in Fig. 18.8 avoided step 3, and would not change very much if step 3 were included, but the statistical inferences involving confidence intervals and significance tests do require step 3.

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i\omega h} f(\omega) d\omega \tag{18.30}$$

and

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i\omega h}. \tag{18.31}$$

From (18.31) it follows immediately that the spectral density is positive, $f(\omega) = f(-\omega)$, $f(\omega)$ is periodic with period 1, and

$$\gamma(0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\omega) d\omega. \tag{18.32}$$

Equation (18.32) says that the total variability $V(X_t)$ is the integral of the spectral density function. This is a continuous analogue of the discrete decomposition (18.18).

Note that (18.29) rules out pure sinusoids. Signals that have purely periodic (composite sinusoidal) components have "mixed" spectra consisting of "line spectra" representing the pure sinusoids and spectral densities representing everything else.

Returning to the periodogram, defined in Equation (18.25), some manipulations (which we omit) show that it may be written in the form

$$I(\omega_j) = \sum_{h=-(n-1)}^{n-1} \hat{\gamma}(h) e^{-2\pi i\omega_j h} \tag{18.33}$$

where $\hat{\gamma}(h)$ is the sample autocovariance function defined in (18.3). Comparing (18.33) with (18.31), we see that the periodogram may be considered an estimator of the spectral density. In addition, using $\hat{\gamma}(-h) = \hat{\gamma}(h)$, Equation (18.33) shows that the periodogram is proportional to the DFT of the sample covariance function.

Further manipulations show that the periodogram may also be written as

$$I(\omega_j) = \frac{1}{n} \sum_{h=-(n-1)}^{n-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \mu)(x_t - \mu) e^{-2\pi i\omega_j h}$$

for $j \neq 0$ and if we replace x_t and $x_{t+|h|}$ with their theoretical counterparts X_t and $X_{t+|h|}$, and then take the expectation, we get

$$E(I(\omega_j)) = \sum_{h=-(n-1)}^{n-1} \left(\frac{n - |h|}{n}\right) \gamma(h) e^{-2\pi i\omega_j h}.$$

Let us consider what happens[9] when $\omega_j \to \omega$ as $n \to \infty$. Assuming the summability condition (18.29) holds we get

$$E(I(\omega_{j_n})) \to \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h},$$

that is,

$$E(I(\omega_{j_n})) \to f(\omega). \tag{18.34}$$

This result forms a connection between the data-based periodogram and the theoretical spectral density: when the periodogram is considered an estimator of the spectral density, for large samples it is approximately unbiased. However, as we will see in Section 18.3.3, the periodogram only becomes a reasonable estimator after smoothing is applied.

18.3.2 For large samples, the periodogram ordinates computed from a stationary time series are approximately independent of one another and chi-squared distributed.

In Section 18.3.1 we showed that the periodogram may be considered an estimator of the spectral density function, but we ended with the remark that it only becomes reasonable after smoothing. We develop this important observation in Section 18.3.3. Here we first review some basic results on the large-sample distribution of the DFT and periodogram. These allow us to get confidence intervals for quantities based on the periodogram, including smoothed periodograms.

The starting point is to imbed the data $x_1, \ldots, x_t$ in a hypothetical infinite sequence of random variables X_t, where t is taken to run through all integers, including negative integers. The assumptions needed for the distributional results are (1) the time series $\{X_t\}$ is stationary; (2) for sufficiently large h, the variables $\{X_t, t < t_0\}$ are nearly independent of the variables $\{X_t, t > t_0 + h\}$ (for any, and therefore—under stationarity—every, t_0); and (3) the spectral density $f(\omega)$ exists. These conditions allow application of the Central Limit Theorem(CLT) to the sum that defines the DFT. We are being deliberately vague in the statement of (2). For technical discussion see Lahiri (2003a).

To get asymptotic variances and covariances, and the asymptotic distribution of the periodogram, let us replace x_t by X_t in (18.20) and (18.21) and consider the large-sample distribution of the coefficients

[9] To get a sequence of Fourier frequencies ω_j that converge to ω, define $\omega_{j_n} = j_n/n$ with j_n a sequence of integers for which $j_n/n \to \omega$.

$$A_k = \frac{2}{n}\sum_{j=1}^{n} X_j \cos(2k\pi j/n)$$
$$B_k = \frac{2}{n}\sum_{j=1}^{n} X_j \sin(2k\pi j/n).$$

To simplify a little, let us write

$$d_c(\omega_k) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_j \cos(2k\pi j/n)$$
$$d_s(\omega_k) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_j \sin(2k\pi j/n).$$

We assume that the expectation of X_t is zero (if not, we can subtract $E(X_t)$ from each variable). By the CLT, $d_c(\omega_j)$ and $d_s(\omega_j)$ are approximately normally distributed. In addition, we have $E(d_c(\omega_k)) = E(d_s(\omega_k)) = 0$ and, it turns out, for the large-sample variances we have

$$V(d_c(\omega_k)) \approx \frac{1}{2} f(\omega_k) \tag{18.35}$$
$$V(d_s(\omega_k)) \approx \frac{1}{2} f(\omega_k) \tag{18.36}$$

while the covariances are approximately zero: for $j \neq k$,

$$Cov(d_c(\omega_j), d_c(\omega_k)) \approx 0 \tag{18.37}$$
$$Cov(d_s(\omega_j), d_s(\omega_k)) \approx 0 \tag{18.38}$$

and for all j, k,

$$Cov(d_c(\omega_j), d_s(\omega_k)) \approx 0. \tag{18.39}$$

The asymptotic independence in (18.37)–(18.39) greatly simplifies statistical inference based on the DFT.

The periodogram is related to $d_c(\omega_k)$ and $d_s(\omega_k)$ by

$$I(\omega_k) = d_c(\omega_k)^2 + d_s(\omega_k)^2.$$

From the CLT together with (18.35) and (18.36), we have

$$\sqrt{\frac{2}{f(\omega_k)}} d_c(\omega_k) \xrightarrow{\mathcal{D}} N(0,1)$$
$$\sqrt{\frac{2}{f(\omega_k)}} d_s(\omega_k) \xrightarrow{\mathcal{D}} N(0,1).$$

By (18.39) these two random variables are approximately independent. Recalling that if $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$, independently, then $Z_1^2 + Z_2^2 \sim \chi_2^2$ we therefore have

$$\frac{2I(\omega_k)}{f(\omega_k)} \text{is approximately } \chi_2^2 \tag{18.40}$$

which we may also write as

$$I(\omega_k) \text{ is approximately } \frac{f(\omega_k)}{2}\chi_2^2.$$

Furthermore, from (18.37)–(18.39), we have that $I(\omega_j)$ and $I(\omega_k)$ are approximately independent for $j \neq k$.

The limiting distribution in (18.40) is a beautifully convenient result, making it relatively easy to get confidence intervals for quantities derived from the periodogram. We describe the methods in Section 18.4.1.

18.3.3 Consistent estimators of the spectral density function result from smoothing the periodogram.

As we discussed in Chapter 8, in large samples the distribution of an estimator T should become concentrated near the quantity θ it is estimating. While (18.40) gives a nice way to assess uncertainty about the periodogram, it also shows that the large-sample distribution of the periodogram does *not* become concentrated around the spectral density: its variance does not decrease with the sample size. In statistical parlance, the periodogram is not a consistent estimator. However, under conditions analogous to those used for consistency of linear smoothers in nonparametric regression, as discussed in Section 15.3.3, smoothed versions of the periodogram will be consistent. This is strong theoretical motivation for smoothing the periodogram.

In the statistical and neuroscientific literatures there are five main approaches to smoothing the periodogram. The first is to apply a smoother, such as a Gaussian kernel smoother to the sequence of values $I(\omega_k)$. Kernel smoothers were discussed in Section 15.3.1 in the context of nonparametric regression and Section 15.4.1 in the context of density estimation. Because kernel smoothers compute linear combinations of the data they are linear smoothers or *linear filters*. We make some further comments about linear filters in Section 18.3.4. When applied to time series Gaussian kernel smoothers are usually called *Gaussian filters*.

The second method of smoothing a periodogram is to split the time domain into a set of many long intervals (long enough to capture low frequencies of potential interest), estimate the spectral density within each interval, and average the resulting estimates. With this method it may be shown that it is advantageous to allow the intervals to have some overlap (Welch 1967). The estimator based on such averaging is sometimes known by the acronym WOSA for *weighted overlapping segment averaging* or *Welch's method*.

The third approach applies a simple generalized linear model based on the asymptotic distribution of the periodogram in (18.40). Recall that the $\chi^2_2/2$ distribution is the same as the standard exponential distribution $Exp(1)$. We may then write

$$I(\omega_k) \stackrel{\cdot}{\sim} f(\omega_k) Exp(1) \tag{18.41}$$

or

$$I(\omega_k) \stackrel{\cdot}{\sim} Exp(\lambda_k) \tag{18.42}$$

where

$$\lambda_k = \frac{1}{f(\omega_k)}.$$

This says that the periodogram ordinates form, approximately, a generalized linear model and therefore may be smoothed using the technology in Section 15.2.3, adapted for exponential regression. The likelihood function based on (18.42) is called the *Whittle likelihood*.

The fourth class of methods for smoothing the periodogram again uses the asymptotic distribution in the form of (18.41) but instead deals with the log ordinates. Letting $Y_k = \log I(\omega_k)$, (18.41) may be written

$$Y_k \approx \log f(\omega_k) + \epsilon_k \tag{18.43}$$

where the ϵ_k variables are independently distributed as $\log X$ where $X \sim Exp(1)$. This provides a standard nonparametric regression model, and the log of an exponential random variable is reasonably close to being normal. However, $E(\epsilon_k) \neq 0$, so there is some bias introduced into the estimation process. Nonetheless, in many cases the bias is small relative to the variation in the log periodogram.

The fifth way to smooth a periodogram is to assume the data follow an autoregressive model, and then use the resulting form of the spectral density. Specifically, calculations show that the $AR(p)$ model (18.27) has spectral density

$$f_X(\omega) = \frac{\sigma_W^2}{|1 - \phi_1 e^{-2\pi i\omega} - \phi_2 e^{-4\pi i\omega} - \cdots - \phi_p e^{-2p\pi i\omega}|^2}.$$

In addition, a more concise class of models, known as *autoregressive moving average* or *ARMA* models, is often used, and these too have closed-form expressions for their spectral densities.

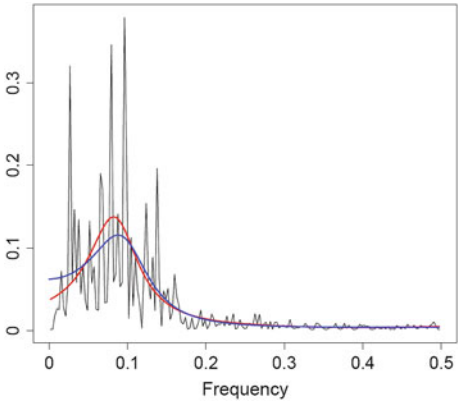

Fig. 18.9 Spectral density estimates for the BARS-detrended residuals from the core body temperature data, after removing the fitted 24 h cycle. The tapered periodogram is highly variable; the Whittle smoothed version is overlaid in *blue*; and the estimate from the $AR(3)$ model is overlaid in *red*.

Example 18.3 (continued from p. 525) We obtained smooth versions of the periodogram for the core temperature data after first removing the trend. (Recall our discussion of Example 15.2 on p. 528; to fit the trend we used the nonparametric regression methodas described briefly in Chapter 15). The $AR(3)$ spectral density estimate is shown in Fig. 18.9. Note that it is very smooth. (An $AR(2)$ based estimate gives similar results.) The Whittle smoothed periodogram is shown for comparison, and agrees reasonably well. There appears to be a peak near $\omega_j = .1$. To interpret this, we need units. The temperature was sampled every 20 min, and there were 352 observations. If $\omega_j = .1$, then the frequency is .1 per time unit (or 35.2 per 352 time units). To get frequency per day we multiply by 72 and get roughly 7. There appears to be a roughly oscillatory component with a period of about 3.5 h. □

We elaborate briefly on linear smoothing in Section 18.3.4 but otherwise omit details on smoothing periodograms.[10] Smoothing is typically handled in spectral analysis software. Regardless of the method used, the most important point is that *some* smoothing is essential.

18.3.4 Linear filters can be fast and effective.

We indicated in Section 18.3.3 that kernel smoothers are linear filters. In this section we say what we mean by a linear filter, and indicate why linear filters are widely applied.

[10] A reference advocating methods three and four, above, is Fan and Kreutzberger (1998).

Suppose we have time series data $x_1, \ldots, x_n$. A linear filter is a set of numbers (coefficients) $\{a_r, a_{r+1}, \ldots, a_s\}$ and its application to the series x_t results in the filtered series

$$y_t = \sum_{h=r}^{s} a_h x_{t-h} \tag{18.44}$$

where, typically, $s - r$ is much less than n. For example, the result of applying the five-point filter with coefficients $(1, 2, 3, 2, 1)/9$ would be

$$y_t = \frac{1}{9}(x_{t-2} + 2x_{t-1} + 3x_t + 2x_{t+1} + x_{t+2}) \tag{18.45}$$

for $t = 3, \ldots, n - 2$. A Gaussian filter would be similar but would instead use a normal (Gaussian) pdf to define the coefficients.

It may be shown that the DFT of $\{y_t\}$ is related to the DFT of $\{x_t\}$ according to

$$d_y(\omega) = \sqrt{n} d_a(\omega) d_x(\omega) \tag{18.46}$$

where $d_a(\omega)$ is the Fourier transform of $\{a_r, a_{r+1}, \ldots, a_s, 0, 0, \ldots, 0\}$, with the zeroes being added to fill up the rest of the n data values. (This is called "padding" the sequence.) The quantity $\sqrt{n} d_a(\omega)$ is called the *transfer function* and its squared magnitude is the *power transfer function*. Expression (18.46) makes it possible to analyze easily the effects of linear filters. This, coupled with their simplicity and the high speed with which they may be computed makes them a very common method of choice for smoothing a time series and the resulting periodogram.

Example 18.3 (continued) We applied the 5-point linear filter described above to the residuals from the core temperature data following simple harmonic regression, yielding a series of the form (18.45). The top panel of Fig. 18.10 shows the residual series and the middle panel shows the power transfer function. The power transfer function decreases to nearly zero as the frequency increases so that high-frequency components have been essentially eliminated from the filtered series. The resulting series is shown in the bottom panel of Fig. 18.10. The filtered series is smoother than the original series. This 5-point linear filter is predominantly a high frequency filter but, as the middle panel of Fig. 18.10 shows, its effects are not restricted to the highest frequencies: there is a gradual squelching of middle-range frequencies as well. □

We have just found that the 5-point linear filter used in (18.45), and applied above to the data from Example 18.3, acts mostly as a high-frequency filter but also displays some gradual mid-range filtering. This might be considered undesirable and one might consider trying to use an ideal high-frequency (or *low-pass*) filter that has a power transfer function of the form

$$H(\omega) = \begin{cases} 1 & \text{for } 0 \le |\omega| \le \omega_c \\ 0 & \text{for } \omega_c < |\omega| \le \frac{1}{2} \end{cases}$$

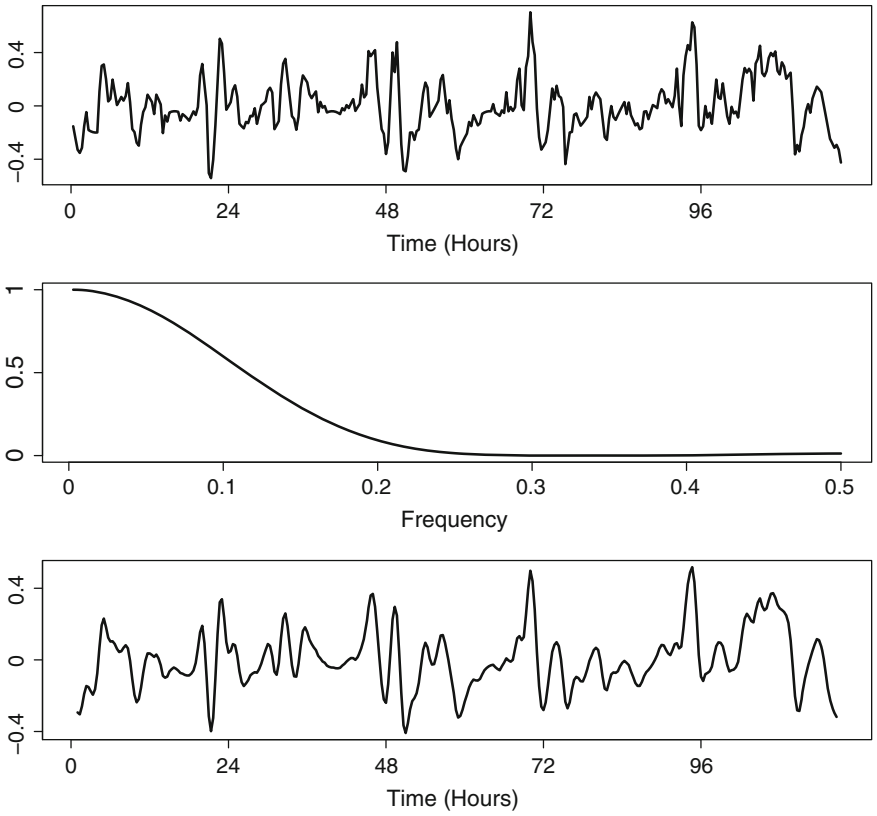

Fig. 18.10 *Top* Core temperature data after removing dominant 24 h effect, i.e., the residuals after simple harmonic regression. *Middle* The power transfer function of the five-point linear filter with coefficients (1, 2, 3, 2, 1)/9, showing a strong diminution of the higher frequency components. *Bottom* Core temperature data after applying the five-point linear filter with coefficients (1, 2, 3, 2, 1)/9.

which would remove all components with frequencies $\omega > \omega_c$ and leave all other components of the series unchanged. One might then, in principle, try to find a filter that corresponds to this power transfer function. This approach turns out to introduce certain technical problems associated with Fourier transforms of discontinuous functions. In practice, time series software typically provides some option for low-pass filtering based on a linear filter, or a combination of linear filters, which aims to approximate the effect of the ideal power transfer function. Similarly, most software provides options for *high-pass* filtering, which approximates an ideal filter that would remove frequencies $\omega < \omega_c$ for some ω_c, and *band-pass* filtering, which approximates an ideal filter that would remove frequencies outside some interval (ω_a, ω_b); the range (ω_a, ω_b) then becomes the frequency band that is retained by the band-pass filter. We illustrated a form of high-pass filtering when we detrended the LFP series

in Example 15.2, with our discussion surrounding Fig. 18.6 (see p. 528), and then again filtered the data in Example 18.3 before fitting the auto-regressive model on p. 532. In the latter case, the detrending method was nonlinear. The advantage of linear filters in practice is the speed with which results may be computed.

All of these remarks about linear filters have theoretical counterparts.

Some details: Suppose $\{X_t; t \in \mathcal{Z}\}$ is a stationary process with spectral density $f_X(\omega)$ and the series $\{a_h; h \in \mathcal{Z}\}$ satisfies

$$\sum_{h=-\infty}^{\infty} |a_h| < \infty.$$

If we let

$$A(\omega) = \sum_{h=-\infty}^{\infty} a_h e^{-2\pi i \omega h},$$

then the filtered process $\{Y_t; t \in \mathcal{Z}\}$ defined by

$$Y_t = \sum_{h=-\infty}^{\infty} a_h X_{t-h}$$

is stationary with spectral density

$$f_Y(\omega) = |A(\omega)|^2 f_X(\omega).$$

Here, the series of coefficients $\{a_h; h \in \mathcal{Z}\}$ is known as the *impulse response function*. □

18.3.5 *Frequency information is limited by the sampling rate.*

While the Fourier frequencies $\omega_k = k/n$ are defined for $k = 1, \ldots, n$, the resulting cosine functions are constrained by the important restriction

$$\cos(2\pi \frac{k}{n} t) = \cos(2\pi \frac{n-k}{n} t) \tag{18.47}$$

for every integer t.

Details: In (18.6) put $u = 2\pi t$ and $v = 2\pi \frac{k}{n} t$ to get

$$\cos(2\pi \frac{n-k}{n} t) = \cos(2\pi t) \cos(2\pi \frac{k}{n} t) + \sin(2\pi t) \sin(2\pi \frac{k}{n} t)$$

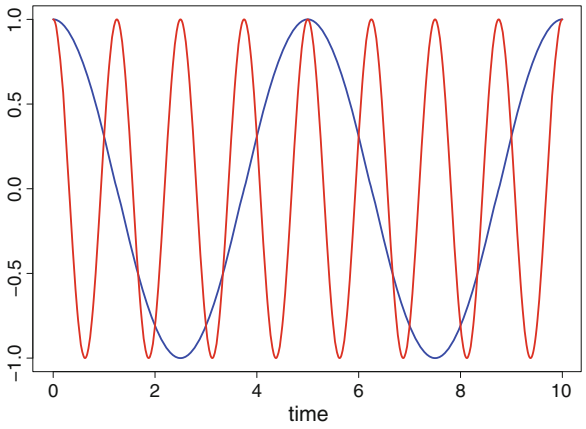

Fig. 18.11 A plot illustrating aliasing of two frequencies for $n = 10$. Two cosine functions are plotted: $\cos(2\pi\omega_1 t)$ and $\cos(2\pi\omega_2 t)$ for $\omega_1 = 2/10$ and $\omega_2 = 8/10$. At all the values $t = 1, \ldots, 10$ these cosine functions agree, so that the frequencies ω_1 and ω_2 are aliased. Note that the time interval between peak and trough corresponding to the second frequency is less than the sampling interval of 1 (equivalently, $\omega_2 > 1/2$) so that, in a sense, the second cosine is oscillating too fast to be determined at this sampling rate. Simple harmonic regression fits for any data sampled at $t = 1, \ldots, 10$ will be the same using ω_2 as using ω_1.

and when t is an integer $\sin(2\pi t) = 0$ while $\cos(2\pi t) = 1$. □

Thus, any cosine with a frequency $\frac{1}{2} < \omega_k < 1$ will have precisely the same values at all integers t as the cosine with frequency $1 - \omega_k$. This is known as *aliasing*: it is not possible to distinguish a cosine function having frequency $\omega^* > \frac{1}{2}$ from another cosine with a frequency in $(0, \frac{1}{2})$. By sampling $x_t = \cos(2\pi\omega t)$ at points $t = 1, \ldots, n$, the fastest visible oscillations occur at the frequency $\omega = \frac{1}{2}$, for which $x_t = \cos(\pi t) = (-1)^t$. (When multiplied by n to get back to the original units of time, this fastest visible frequency of oscillation is called the *Nyquist frequency.*) The situation is illustrated in Fig. 18.11. Corresponding to (18.47) we also have

$$\sin(2\pi\frac{k}{n}t) = -\sin(2\pi(\frac{n-k}{n})t).$$

These aliasing relations have analogues in the DFT. They imply that[11] the second half of the components of the DFT, those for which $\omega_k > \frac{1}{2}$, are redundant with the first. Plots of the periodogram therefore correspond to frequencies only up to $\omega_k = \frac{1}{2}$.

[11] This assumes the data are real numbers. It is occasionally useful, instead, to examine data that consist of complex numbers.

18.3.6 *Tapering reduces the leakage of power from non-Fourier to Fourier frequencies.*

The intuitive description of Fourier analysis in Section 18.2.1 left out an important fact. If we consider the fundamental cosine and sine functions $\cos(2\pi t)$ and $\sin(2\pi t)$, these are functions not only on $[0, 1]$ but on the whole real line. They and all of the resulting cosine and sine functions at harmonic frequencies, i.e., the functions $\cos(2\pi kt)$ and $\sin(2\pi kt)$ for $k = 1, 2, \ldots$, will be periodic on the interval $[0, 1]$. So that all of these functions satisfy

$$f(0) = f(1). \tag{18.48}$$

The rough arguments we gave in Section 18.2.1 make the most sense for functions that satisfy (18.48). When this constraint does not hold, it turns out that the Fourier approximation (18.16) suffers from a failure to adequately represent $f(t)$, which is known as the *Gibbs phenomenon.* The corresponding effect when applying the DFT to data is known as *leakage*.

To describe the problem of leakage, let us consider the periodogram of the cosine function $x_t = \cos(2\pi\omega t)$, for $t = 1, \ldots, n$. Calculation shows that this periodogram (for each Fourier frequency ω_j) is given by

$$I(\omega_j) = n|D_n(\omega - \omega_j)|^2 \tag{18.49}$$

where

$$D_n(\phi) = \frac{\sin(\pi n\phi)}{n\sin(\pi\phi)}$$

is known as the *Dirichlet kernel.* If ω is a Fourier frequency, then $I(\omega_j)$ has a single spike at $\omega_j = \omega$ and is zero at all other Fourier frequencies ω_j. In other words, in this case the periodogram correctly finds the sole cosine component.

> *Details:* Note that as $\phi \to 0$, $D_n(\phi) \to \frac{1}{n}$ (by L'Hopital's rule), so $D_n(\phi)$ at $\phi = 0$ is defined to be $D_n(0) = \frac{1}{n}$. Thus, when $\omega_j = \omega$ we have $I(\omega_j) = \frac{1}{n}$. If ω is a Fourier frequency then $\omega - \omega_j$ has the form $\frac{k}{n}$ for some integer k and $D_n(\omega - \omega_j) = 0$ for all j except when $\omega_j = \omega$. □

On the other hand, when ω is not a Fourier frequency the Dirichlet kernel creates "side lobes," as shown in Fig. 18.12, where $D_n(\omega - \omega_j)$ will be nonzero even for frequencies ω_j that are not immediately non-adjacent to ω. As a consequence, the power at frequency ω will "leak" to other frequencies in the periodogram, so the periodogram indicates misleadingly that those other frequencies are present in the data.

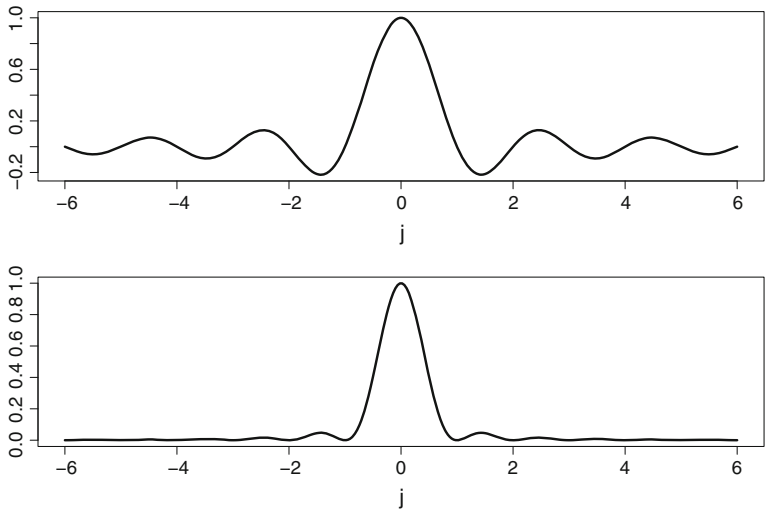

Fig. 18.12 *Top* The Dirichlet kernel $D_{100}(j/100)$, here plotted for values of j ranging from -6 to 6. A continuous *curve* was generated by taking non-integer values of j. *Bottom* The periodogram $I(j/100) = 100|D_{100}(j/100)|^2$, after scaling by dividing by 100.

The problem of leakage is very dramatic when the *dynamic range* of the data is large. Dynamic range refers to the ratio of the largest to smallest positive periodogram values (usually measured on the $\log_{10}$, or decibel, scale).

Illustration: As an illustration, consider

$$x_t = 20\cos(2\pi\omega_1 t) + \cos(2\pi\omega_2 t) \tag{18.50}$$

where $n = 100$, $\omega_1 = .05$ and $\omega_2 = .15$. Its periodogram is shown in the top panel of Fig. 18.13. To see the second frequency it is necessary to use a log scale to plot the periodogram, as shown in the bottom panel of Fig. 18.13. Log periodogram plots are used as defaults in many contexts. Now consider the leakage-prone variant where we take $\omega_1 = 1/22$ rather than $1/20$. Its periodogram is shown in Fig. 18.14. In this case leakage obscures the second peak almost entirely, and if the periodogram were noisy (as it is with real data) it would be extremely difficult to see the second peak at all. □

Leakage is also a problem when there are trends, which cause large low-frequency coefficients in the periodogram.

Example 15.2 (continued from p. 528) We previously showed the log periodogram for the LFP data in Fig. 18.5. The very low frequency trends cause leakage, which obscures the higher frequencies of interest. □

The standard solution to the problem of leakage is to force the data to satisfy (18.48) by applying *tapering*. Tapering decreases bias due to leakage in spectral density

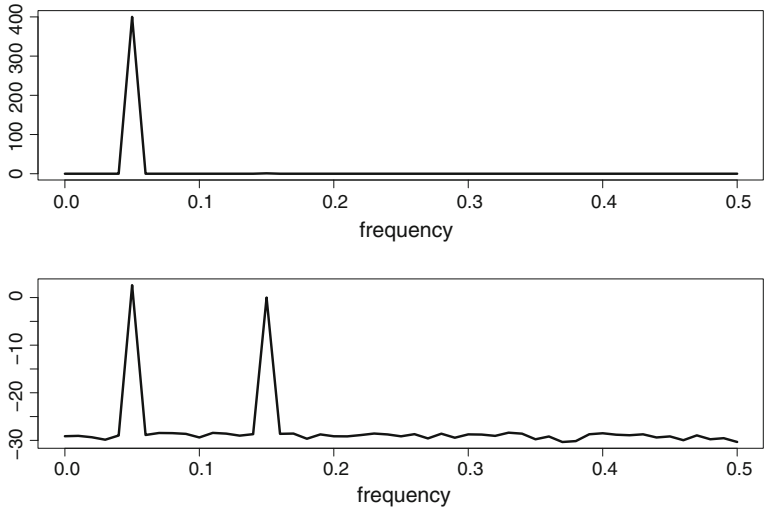

Fig. 18.13 *Top* Periodogram of $x_t = 20\cos(2\pi\omega_1 t) + \cos(2\pi\omega_2 t)$, where $n = 100$, $\omega_1 = .05$ and $\omega_2 = .15$. *Bottom* Log periodogram of x_t. In the log scale the second peak becomes visible.

estimation by damping down the ends of the series toward zero, forcing the series to have period equal to its length (and thus satisfying (18.48)). This is accomplished in standard spectral analysis software. Because the beginning and end of the tapered series have values close to zero, however, this reduces the effective sample size of the series and therefore loses some information. It has been shown that the use of the mean of multiple tapers can recover this information.[12] Multi-taper estimation is used as a default in some software.

18.3.7 Time-frequency analysis describes the evolution of rhythms across time.

Up until this point, Section 18.3 has presented powerful methods for spectral analysis of time series under the assumption of stationarity. We have emphasized that time series should not be considered stationary when there are slowly varying trends, as displayed in Fig. 1.5 of Example 1.6 and Fig. 18.1 of Example 15.2. In many cases, however, a different kind of non-stationarity is present and, in fact, may be of great interest: the frequency content of a signal may change across time.

Example 2.2 (continued from p. 514) The spectrograms in Fig. 2.2 on p. 27 displayed nicely some changes in the frequency content of EEGs across the course of

[12] See Mitra and Pesaran (1999), Percival and Walden (1993), and Thomson (1982).

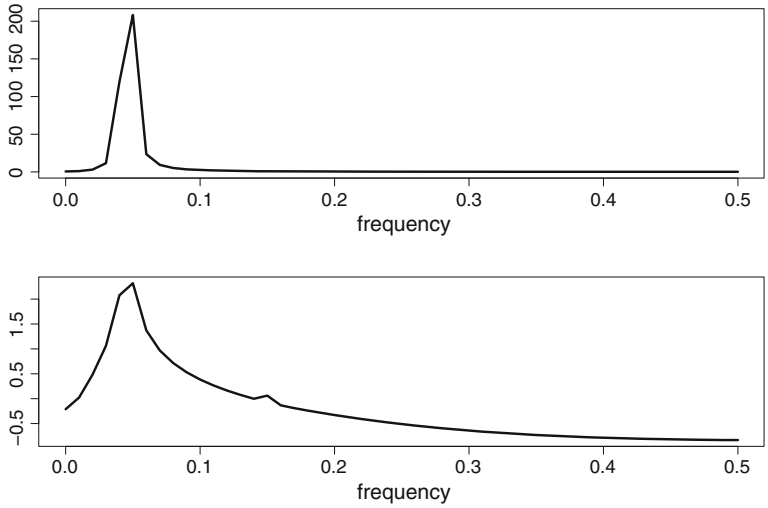

Fig. 18.14 *Top* Periodogram of $x_t = 20\cos(2\pi\omega_1 t) + \cos(2\pi\omega_2 t)$, where $n = 100$, $\omega_1 = 1/22$ and $\omega_2 = .15$. *Bottom* Log periodogram of x_t. Due to leakage, the second peak is obscured.

the experiment. Specifically, the alpha rhythm appeared during an epoch in which the subject's eyes closed, and during induction of anesthesia. □

Spectrograms, such as that in Example 2.2, may be created by segmenting the observation time interval $[0, T]$ into a set of subintervals $[0, T_1], [T_1, T_2], \ldots, [T_k, T]$, and then computing spectral density estimates within each interval. The estimated spectrum is then plotted on the y-axis for every time interval, with time labeled along the x-axis. The intervals must be chosen to be long enough so that there are substantial series from which to estimate the spectrum, yet short enough that the series may be considered stationary within each interval. Some spectrogram software takes as a default 512 observations per interval (with corrections to this to allow for T not being divisible by 512). Some smoothing (and tapering) of the spectral density estimates across time is often incorporated. One way to smooth across time, which is available as an option in most spectrogram software, is to choose the analysis intervals to be overlapping. In some experiments there are repeated trials, in which case the spectrograms may be averaged across trials.

Example 18.2 (continued from p. 518) To display the LFP response to the stimulus Logothetis et al. (2001) used a spectrogram that incorporated tapering and was averaged across trials and across subjects. It showed strong power in the gamma range after onset of the stimulus. □

Time-frequency analysis is often performed using wavelets (Section 15.2.8). Because of the scaling property (the narrowing range) in the definition (15.9), wavelet regression provides a representation that is localized in both time and frequency, with frequency here defined by the scale of the wavelets. See Percival and Walden (2000).

Example 18.1 (continued from p. 518) In their study of MEG oscillatory activity during learning, Chaumon et al. (2009) used Morlet wavelets (see p. 429) to decompose MEG sensor signals across time and frequency. They analyzed the log-transformed power within a 30–48 Hz, band at time 100–400 ms after target onset, from one group of sensors over the occipital lobe and another group of sensors over the frontal lobe. They found that during the learning phase (the first few blocks) of the experiment this gamma band power in the sensors over the occipital lobe was higher for the predictive trials than for the nonpredictive trials ($p < .005$ based on an across-subject paired t-test, using 16 subjects) with the power for the predictive trials being elevated above baseline. On the other hand, during the same learning period, the gamma band power in the sensors over the frontal lobe was depressed for the nonpredictive trials ($p < .0001$), but not for the predictive trials (with the predictive and nonpredictive gamma band power being different, $p < .01$). □

18.4 Propagation of Uncertainty for Functions of the Periodogram

18.4.1 Confidence intervals and significance tests may be carried out by propagating the uncertainty from the periodogram.

The large-sample result described by (18.41) together with the approximate independence of $I(\omega_j)$ and $I(\omega_k)$, for $j \neq k$, provide uncertainty about the estimate of the spectral density and also make it easy to propagate this uncertainty. Importantly, this result holds in the same form for periodograms computed with suitable tapers. (See the brief discussion in Percival and Walden (1993, p. 190), which cites Brillinger (1981, p. 107).)

Now suppose we have computed some feature of the periodogram and we want a 95 % confidence interval associated with that feature. For example, we may have smoothed the periodogram and may want bands to represent our uncertainty. Let $m = (n-1)/2$ if n is odd; $n/2$ if n is even. For a range of ω values, write the smoothed version at frequency ω in the form $g_\omega(I(\omega_1), \ldots, I(\omega_m))$. That is, the operation that produced the smooth value at frequency ω is being written as a function g_ω of the periodogram values. We would say that $g_\omega(I(\omega_1), \ldots, I(\omega_m))$ is an estimator of $f(\omega)$. To apply propagation of error we do the following.

1. For $j = 1$ to J:

   For $i = 1, \ldots, m$:
   generate observations Y_i from an $Exp(1)$ distribution;
   define $U_i^{(j)} = \hat{f}(\omega_i)Y_i$, where $\hat{f}(\omega_i)$ is an estimate of $f(\omega_i)$ (based on a smoothed periodogram).
   Compute $W^{(j)} = g_\omega(U_1^{(j)}, U_2^{(j)}, \ldots, U_m^{(j)})$.

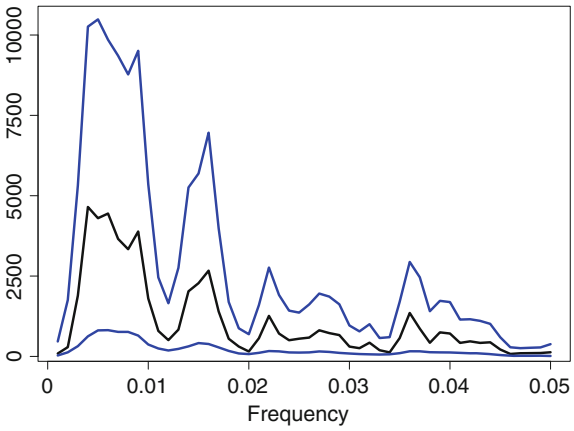

Fig. 18.15 Smoothed periodogram and approximate, pointwise 95 % confidence bands, from the beginning-period LFP detrended series.

2a. Set $\overline{W} = \frac{1}{J}\sum W^{(j)}$ and then $SE^2 = \frac{1}{J-1}\sum(W^{(j)} - \overline{W})^2$ is the squared standard error of $g_\omega(I(\omega_1), \ldots, I(\omega_m))$.

2b. Let $W_{.025}$ and $W_{.975}$ be .025 and .975 quantiles in the sample $W^{(1)}, \ldots, W^{(J)}$. Then $(W_{.025}, W_{.975})$ is an approximate 95 % confidence interval (for $f(\omega)$) associated with $g_\omega(I(\omega_1), \ldots, I(\omega_m))$.

In practice, we would compute a whole set of $W^{(j)}$ values for different g_ω functions, corresponding to different values of ω. This would give us approximate pointwise[13] confidence bands on the smoothed periodogram.

In step 1 of the algorithm above an estimate $\hat{f}(\omega_i)$ (based on the smoothed periodogram) is used in place of $f(\omega_i)$, because the latter is unknown and so can't be computed. This is usually called a bootstrap, analogously to the bootstrap procedures in Chapter 9.

Example 15.2 (continued from p. 528) Returning to the pair of 1 s average LFP recordings, we noted previously, in Figs. 18.1 and 18.5, the need to detrend the time series before looking for periodicities under the assumption of stationarity. Figure 18.6 displayed the smoothed periodograms of the detrended series. Pointwise 95 % confidence bands together with the smoothed periodogram for the first period, obtained by propagation of uncertainty, are shown in Fig. 18.15.

We next consider whether the first and last periods have the same spectral density (an indication of stationarity). Figure 18.16 shows the two smoothed periodograms overlaid. A significance test may be based on the integrated squared difference between the two smooth curves. Specifically, if $\hat{f}_1(\omega)$ and $\hat{f}_2(\omega)$ are the two spectral

[13] By pointwise we mean that at any given frequency ω the bands would provide an approximate 95 % confidence interval. An alternative is to compute approximate *simultaneous* confidence bands, meaning bands that provide approximate 95 % confidence simultaneously for all ω. This may be accomplished with a suitable adaptation of the algorithm.

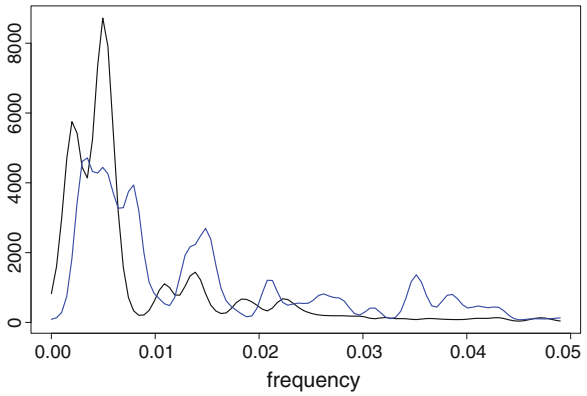

Fig. 18.16 Smoothed periodograms from beginning and end periods, overlaid.

density estimates, then we use

$$t_{obs} = \sum_k (\hat{f}_1(\omega_k) - \hat{f}_2(\omega_k))^2$$

as the test statistic. To compute a p-value under $H_0 = f_1(\omega) = f_2(\omega)$ for all ω, we take as a "pooled" estimate

$$\hat{f}(\omega_k) = \frac{1}{2}(\hat{f}_1(\omega_k) + \hat{f}_2(\omega_k))$$

for $k = 1, \ldots, m$. We then generate a pseudo-sample of pairs of periodograms using $\hat{f}(\omega)$ as the spectral density, and for each generated pair of periodograms, apply smoothing and compute t. We then see what fraction of the generated t values is greater than t_{obs}. This is our approximate p-value. In this case, we obtained $p = 0.53$, indicating no evidence that the spectra from the two recording intervals are different. □

18.4.2 Uncertainty about functions of time series may be obtained from time series pseudo-data.

The method above propagates the uncertainty from the asymptotic distribution of the periodogram to anything computed from it. If, however, an analytical technique bypasses the periodogram a different method must be used to propagate uncertainty. A more general idea is to use the approximate normal distributions on the coefficients, in order to propagate the uncertainty from the DFT itself. In other words, one may begin with the uncertainty in the DFT obtained from the data, and then apply an

inverse DFT to generate time series that behave the same as the original series in the sense of having (approximately) the same spectrum. The resulting time series pseudo-data are sometimes called *surrogate data*.

An efficient method of carrying out such simulations (based on "circulant embedding") is described in Percival and Constantine (2006). Code by these authors is available in the CRAN library of R packages, within the package `fractal`. See below. As described in the Percival and Constantine paper, the method is closely related to *surrogate time series*, e.g., Schreiber and Schmitz (2000). Additional "bootstrap" resampling methods for spectral analysis, with an emphasis on theoretical results, are discussed in Chapter 9 of Lahiri (2003b). We omit detailed discussion of this topic and note only that the pseudo data generated by this approach are normal (Gaussian), and so do not reflect any sources of uncertainty arising from substantial non-normal variation in the data.

18.5 Bivariate Time Series

Suppose $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are sequences of observations made across time, and the problem is to describe their sequential relationship. For example, an increase in y_t may tend to occur following some increase or decrease in a linear combination of some of the preceding x_t values. This is the sort of possibility that bivariate time series analysis aims to describe.

Example 18.4 Beta oscillations during a sensorimotor task. Brovelli et al. (2004) recorded local field potentials from multiple sites simultaneously while a subject (a rhesus monkey) performed a Go/No-Go visuomotor task. Results were reported for two monkeys. The task required the subject hold down a lever during an interval having a randomly determined length while a stimulus appeared. On Go trials, a reward was given if the monkey released the lever within 500 ms. The purpose of the study was to look for coordinated rhythmic activity across the recording sites during a task that required focused attention. Of particular interest was the range of frequencies identified as *beta oscillations*, which the authors took to be 14–30 Hz. The specific question was whether local field potentials in sensory and motor regions exhibit co-ordinated patterns within the beta range of frequencies. □

The theoretical framework of such efforts begins, again, with stationarity. A joint process $\{(X_t, Y_t), t \in \mathcal{Z}\}$ is said to be *strictly stationary* if the joint distribution of $\{(X_t, Y_t), \ldots, (X_{t+h}, Y_{t+h})\}$ is the same as that of $\{(X_s, Y_s), \ldots, (X_{s+h}, Y_{s+h})\}$ for all integers s, t, h. The process is *weakly stationary* if each of X_t and Y_t is weakly stationary with means and covariance functions μ_X, $\gamma_X(h)$ and μ_Y, $\gamma_Y(h)$, and, in addition, the cross-covariance function

$$\gamma_{XY}(s,t) = E((X_s - \mu_X)(Y_t - \mu_Y))$$

depends on s and t only through their difference $h = t - s$, in which case we write it in the form

$$\gamma_{XY}(h) = E((X_{t-h} - \mu_X)(Y_t - \mu_Y)).$$

Note that $\gamma_{XY}(h) = \gamma_{YX}(-h)$. The *cross-correlation* function of $\{(X_t, Y_t)\}$ is

$$\rho_{XY}(h) = \frac{\gamma_{XY}(h)}{\sigma_X \sigma_Y}$$

where $\sigma_X = \sqrt{\gamma_X(0)}$ and similarly for Y_t. The cross-correlation $\rho_{XY}(h)$ is the ordinary correlation between the random variable X_{t-h} and Y_t. Just as the ordinary correlation ρ may be interpreted as a measure of linear association between two random variables, the cross-correlation $\rho(h)$ may be interpreted as a measure of linear association between two stationary processes at lag h. The cross-covariance and cross-correlation functions are estimated by their sample counterparts:

$$\hat{\gamma}_{XY}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (x_t - \bar{x})(y_{t+h} - \bar{y})$$

with $\hat{\gamma}_{XY}(-h) = \hat{\gamma}_{YX}(h)$, and

$$\hat{\rho}(h) = \frac{\hat{\gamma}_{XY}(h)}{\hat{\sigma}_X \hat{\sigma}_Y}.$$

The univariate Eqs. (18.29)–(18.31) have immediate extensions to the bivariate case: if

$$\sum_{h=-\infty}^{\infty} |\gamma_{XY}(h)| < \infty$$

then there is a *cross-spectral density function* $f_{XY}(\omega)$ for which

$$\gamma_{XY}(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} f_{XY}(\omega) d\omega \tag{18.51}$$

and

$$f_{XY}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{XY}(h) e^{-2\pi i \omega h}.$$

The cross-spectral density is, in general, complex valued. Because $\gamma_{YX}(h) = \gamma_{XY}(-h)$ we have

$$f_{YX}(\omega) = \overline{f_{XY}(\omega)} \tag{18.52}$$

i.e., $f_{YX}(\omega)$ is the complex conjugate of $f_{XY}(\omega)$. In Section 18.3.1 we said that a smoothed periodogram could be considered an estimator of the theoretical spectral

density, and we based that interpretation on a finite-sample expression (18.33), which gave the periodogram as a scaled DFT of the sample covariance function. Similarly, an estimate $\hat{f}_{XY}(\omega)$ of $f_{XY}(\omega)$ may be obtained by smoothing a scaled DFT of the sample cross-covariance function $\hat{\gamma}_{XY}(h)$. In Section 18.5.1 we discuss the important concept of *coherence*, which is defined in terms of the cross-spectral density.

18.5.1 The coherence $\rho_{XY}(\omega)$ between two series X and Y may be considered the correlation of their ω-frequency components.

There is a very nice way to decompose into frequencies the linear dependence between a pair of stationary time series. This frequency-based measure of linear dependence forms an analogy with ordinary correlation which, as we noted in Section 4.2.1, may be interpreted as a measure of linear association. To substantiate this interpretation for the ordinary correlation ρ between two random variables Y and X we provided on p. 81 a theorem concerning the linear prediction of Y from $\alpha + \beta X$, giving the formula for α and β that minimized the mean squared error of prediction, $E\left((Y - \alpha - \beta X)^2\right)$ and showing that when these optimal values of α and β are plugged in, the minimum mean squared error became

$$E\left((Y - \alpha - \beta X)^2\right) = \sigma_Y^2(1 - \rho^2), \tag{18.53}$$

which was Eq. (4.11).

In Eq. (18.53) we considered the linear prediction of Y based on X, meaning the prediction of Y based on a linear function of X. The analogous problem for $\{(X_t, Y_t), t \in \mathcal{Z}\}$ is to assume

$$Y_t = \sum_{h=-\infty}^{\infty} \beta_h X_{t-h} + W_t, \tag{18.54}$$

where W_t is a stationary process independent of $\{X_t\}$, with $E(W_t) = 0$ and $V(W_t) = \sigma_W^2$, and to minimize the mean squared error

$$MSE = E\left(Y_t - \sum_{h=-\infty}^{\infty} \beta_h X_{t-h}\right)^2. \tag{18.55}$$

Some manipulations show that the solution satisfies

$$\min MSE = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_Y(\omega)(1 - \rho_{XY}(\omega)^2)d\omega \tag{18.56}$$

where

$$\rho_{XY}(\omega)^2 = \frac{|f_{XY}(\omega)|^2}{f_X(\omega) f_Y(\omega)} \tag{18.57}$$

is the *squared coherence*. Thus, in analogy with (18.53), $f_Y(\omega)(1 - \rho_{XY}(\omega)^2)$ is the ω-component of the minimum-*MSE* fit of (18.54). In (18.56) we have $MSE \geq 0$ and $f_Y(\omega) \geq 0$, which together imply that $0 \leq \rho_{XY}(\omega)^2 \leq 1$ for all ω, and when

$$Y_t = \sum_{h=-\infty}^{\infty} \beta_h X_{t-h}$$

we have $\rho_{XY}(\omega)^2 = 1$ for all ω. These facts, together with (18.56), give the interpretation that the squared coherence is a frequency-based analogue to squared correlation between two theoretical time series.

Additional details: The interpretation of coherence in terms of correlation may be pushed further, but is somewhat subtle. In defining the cross-spectral spectral density we mentioned that it is complex valued. Let $\theta(f_{XY}(\omega))$ be the phase of $f_{XY}(\omega)$, which we may write in terms of the real and imaginary parts of $f_{XY}(\omega)$,

$$\theta(f_{XY}(\omega)) = \arctan \frac{Im(f_{XY}(\omega))}{Re(f_{XY}(\omega))}$$

so that

$$f_{XY}(\omega) = |f_{XY}| \exp\left(i\theta(f_{XY}(\omega))\right).$$

The function $\theta(f_{XY}(\omega))$ is often called the *phase coherence*. The *coherence* is then the complex-valued function defined by

$$\rho_{XY}(\omega) = \frac{f_{XY}(\omega)}{\sqrt{f_X(\omega) f_Y(\omega)}}.$$

This complex-valued coherence contains phase information, which is necessary when considering the tendency of two signal components at frequency ω to vary together. The magnitude of the coherence is often considered to be a measure of phase-locking of the two signals, but it also depends on the relationship of their amplitudes.
A more complete explanation of coherence is beyond the scope of our presentation here.[14] □

From a pair of observed time series the squared coherence may be estimated by

[14] One helpful fact is that an average coherence across a given frequency band may be shown to be equal to the complex-valued correlation between band-pass filtered versions of the two series; see Ombao and Vanbellegem (2008).

$$\hat{\rho}^2_{XY}(\omega) = \frac{|\hat{f}_{XY}(\omega)|^2}{\hat{f}_X(\omega)\hat{f}_Y(\omega)} \tag{18.58}$$

where, again, $\hat{f}_{XY}(\omega)$ is a smoothed version of the DFT of $\hat{\gamma}_{XY}(h)$. However, the smoothing in this estimation process is crucial. The raw cross-periodogram $I_{XY}(\omega)$ satisfies the relationship

$$|I_{XY}(\omega)|^2 = I_X(\omega)I_Y(\omega)$$

so that plugging the raw periodograms into (18.58) will always yield the value 1. Thus, again, it is imperative to smooth periodograms before interpreting them.

Example 18.4 (continued from p. 553) Brovelli et al. collected approximately 900 successful Go trials, using data from 90 ms prior to stimulus onset to 500 ms after onset. They subtracted out the trial-averaged signals to produce approximately stationary multiple time series. To look for the presence of beta oscillations in sensorimotor cortex they recorded from six sites in one animal and four in another. The sites are shown in Fig. 18.17. The sites shown in part A of the figure appear to be in (1) the arm area of primary motor cortex (M1), (2) the arm area of sensory cortex (S1), (3) anterior intraparietal cortex (AIP, object and hand shape representation), (4) lateral intraparietal cortex (used in guiding saccades and identifying visual locations), (5) ventral premotor cortex, (6) dorsal premotor cortex. In part B of the figure the sites appear to be in (1) the wrist area of M1 or ventral premotor cortex, (2) the wrist area of S1, (3) AIP, (4) medial intraparietal cortex (related to goals or targets of intended reach).

The authors computed squared coherence for each pair of sites, as in (18.57), with ω in the beta range, then found the maximum squared coherence across all values of ω, and performed a permutation significance test (see Section 11.2.1) to see whether that maximum was sufficiently large to form clear evidence of underlying coherence in LFP across brain regions. Their results are depicted on the left side of Fig. 18.17. The authors found that primary motor cortex (M1, site 1 in both monkeys), primary sensory cortex (S1, site 2), and anterior intraparietal cortex (AIP, site 3) were all engaged in coherent oscillatory activity during the task. □

18.5.2 In examining cross-correlation or coherence of two time series it is advisable first to pre-whiten the series.

In Section 12.2.3 we highlighted the importance of the assumption of independent errors in linear regression: we showed that the squared correlation between two *independent* $AR(1)$ time series is likely to be statistically significant, erroneously indicating association. A similar phenomenon occurs for the cross-correlation, and for coherence. To avoid it, the serial dependence should be removed from the two series before the cross-correlation or coherence is computed. For example, if we have two series $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ we could fit appropriate AR models to each

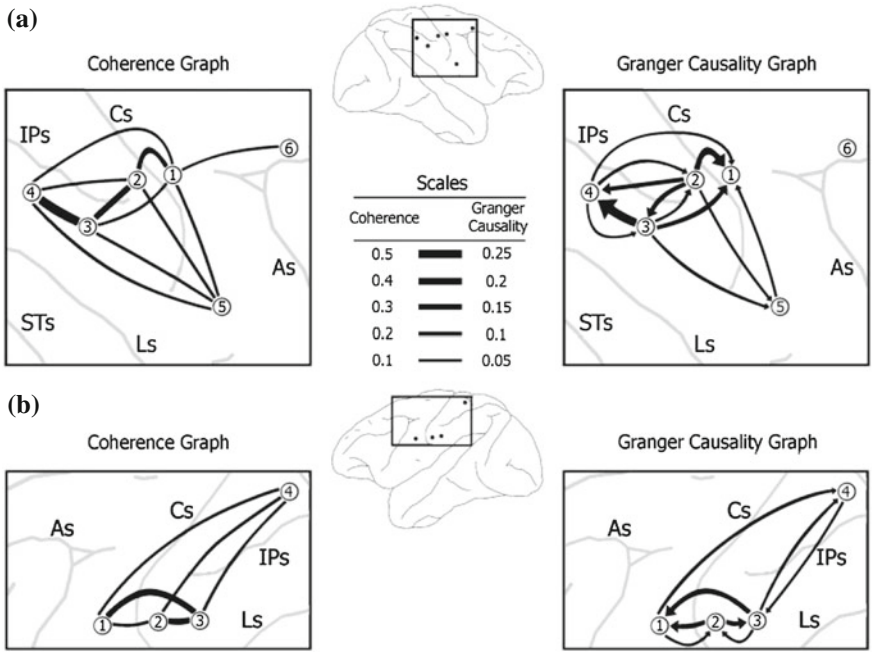

Fig. 18.17 Figure adapted from Brovelli et al. showing coherence and Granger causality among six recording sites in one monkey (part A) and four in another (part B). On the *left* are lines representing statistically significant coherence between a pair of sites ($p < .005$ based on a permutation test with a correction for multiple comparisons), with thickness indicating the magnitude of coherence as shown on the scale graphic in the middle of the figure. On the *right* are lines, some of which have *arrows*, representing statistically significant Granger causality, with magnitudes again indicated by line thickness as shown on the scale graphic in the middle of the figure. Recording sites are shown above and below the scale graphic.

series and then work instead with the residuals obtained from subtracting the *AR* fits. An alternative procedure involves fitting an *AR* (or *ARMA*) model then applying a suitable filter that removes the serial dependence. See Box et al. (2008) for discussion of this approach.

Example 18.2 (continued from p. 518) In their study Logothetis et al. (2001) reported the distribution of R^2 values between[15] LFP and BOLD signals across trials, which were generally substantial, with a mean of .52. Before computing these correlations, however, they pre-whitened the series using $AR(10)$ models. □

[15] Actually, they reported R^2 between stimulus-based impulse response functions (see p. 544) found from the LFP and BOLD signals.

18.5.3 *Granger causality measures the linear predictability of one time series by another.*

The squared coherence provides a frequency-based measure of linear association between two time series. Just as the correlation $Cor(X, Y)$ is symmetrical in its arguments X and Y, so too is the squared coherence. In contrast, regression is directional. We now develop a simple directional assessment of linear predictability of one time series from another.

The idea is very simple. In ordinary regression we assess the influence of a variable (or set of variables) X_2 on Y in the presence of another variable (or set of variables) X_1 by examining the reduction in variance when we compare the regression of Y on (X_1, X_2) with the regression of Y on X_1 alone. If the variance is reduced sufficiently, then we conclude that X_2 helps explain (predict) Y. Here, we replace Y with Y_t, replace X_1 with $\{Y_s, s < t\}$ and X_2 with $\{X_s, s < t\}$. In other words, we examine the additional contribution to predicting Y_t made by the past observations of X_s after accounting for the autocorrelation in $\{Y_t\}$. The "causality" part comes when the past of X_s helps predict Y_t but the past of Y_s does *not* help predict X_t.

Let us begin by defining what it means for $\{(X_t, Y_t), t \in \mathcal{Z}\}$ to follow a joint $AR(p)$ process. Working by analogy with the definition (18.27), we write

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \sum_{i=1}^{p} \begin{pmatrix} \phi_i^{XX} & \phi_i^{XY} \\ \phi_i^{YX} & \phi_i^{YY} \end{pmatrix} \begin{pmatrix} X_{t-i} \\ Y_{t-i} \end{pmatrix} + \begin{pmatrix} W_t^{X|XY} \\ W_t^{Y|XY} \end{pmatrix} \tag{18.59}$$

where $W_t^{X|XY}$ and $W_t^{Y|XY}$ are independently $N(0, \sigma^2_{X|XY})$ and $N(0, \sigma^2_{Y|XY})$. The notational superscripts and subscripts $X|XY$ and $Y|XY$ are used to indicate variables or variances for the joint $AR(p)$ model (18.59), in which both $X_1, \ldots, X_{t-p}$ and $Y_1, \ldots, Y_{t-p}$ appear on the right-hand side. This is in contrast to the usual univariate $AR(p)$ models for $\{Y_t, t \in \mathcal{Z}\}$,

$$Y_t = \sum_{i=1}^{p} \phi_i^Y Y_{t-i} + W_t^Y, \tag{18.60}$$

where W_t^Y are independently[16] $N(0, \sigma^2_{Y|Y})$, and for $\{X_t, t \in \mathcal{Z}\}$,

$$X_t = \sum_{i=1}^{p} \phi_i^X X_{t-i} + W_t^X, \tag{18.61}$$

where W_t^X are independently $N(0, \sigma^2_{X|X})$. We may now say that $\{X_t, t \in \mathcal{Z}\}$ is predictive of $\{Y_t, t \in \mathcal{Z}\}$ if $\sigma_{Y|XY} < \sigma_{Y|Y}$. In this situation, $\{X_t, t \in \mathcal{Z}\}$ is also said to be

[16] Here $\sigma^2_{Y|Y}$ is a constant; the notation is intended only to indicate that it is the error variance when Y appears on both the left-hand side and the right-hand side of the model.

Granger causal of $\{Y_t, t \in \mathcal{Z}\}$. Similarly, we say $\{Y_t, t \in \mathcal{Z}\}$ is predictive (Granger causal) of $\{X_t, t \in \mathcal{Z}\}$ if $\sigma_{X|XY} < \sigma_{X|X}$. This kind of predictability is often quantified by the *Granger causality measure*

$$F_{X\to Y} = 2\log\frac{\sigma_{Y|Y}}{\sigma_{Y|XY}}.$$

Theoretical analysis of this approach was given by Geweke (1982), based on earlier work by Granger (1969).[17]

In applications, to evaluate whether a time series $x_t, t = 1, \ldots, n$ is predictive of $y_t, t = 1, \ldots, n$, the basic procedure is to (1) fit a bivariate $AR(p)$ model, then (2) test the hypothesis $H_0 : \phi_i^{YX} = 0$ for all i, which is equivalent to testing $H_0 : F_{X\to Y} = 0$.

Illustration As an illustration, we simulated a bivariate time series of length 1,000 using the model

$$\begin{aligned} X_t &= .5X_{t-1} + U_t \\ Y_t &= .2Y_{t-1} + .5X_{t-1} + V_t \end{aligned}$$

where $U_t \sim N(0, (.2)^2)$ and $V_t \sim N(0, (.2)^2)$, independently. We then fit a linear regression model of the form

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{t-1} + \epsilon_t$$

and, similarly, fit another model of the same form but with the roles of X and Y reversed. The results for the two regressions are shown in the following table.

Variable	Coefficient	Std. Err.	t-ratio	p-value
Intercept	−.001	.006	−.211	.83
x_{t-1}	.496	.012	42.7	$<10^{-15}$
y_{t-1}	.192	.018	10.7	$<10^{-15}$
Intercept	.008	.016	.536	.59
x_{t-1}	.508	.029	17.1	$< 10^{-15}$
y_{t-1}	−.055	.045	−1.3	.228

As expected, the first fit indicates that X_{t-1} provides additional information beyond Y_{t-1} in predicting Y_t, while the second fit shows that Y_{t-1} does *not* provide additional information beyond X_{t-1} in predicting X_t. This is sometimes summarized by saying

[17] In addition, Geweke (1982) defined a spectral measure $f_{X\to Y}(\omega)$ representing the ω-component of Granger causality in the sense that

$$F_{X\to Y} = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_{X\to Y}(\omega)d\omega.$$

X_t is *causally* related to Y_t, but we must keep in mind that "causal" is used in a predictive, time-directed sense. □

This illustration sweeps under the rug the selection of auto-regressive order p in part of the problem, in step (1) above. In applications this is non-trivial, and care should be taken to make sure interpretations do not depend on choices of p that involve substantial uncertainty.

Example 18.4 (continued from p. 556) Results of Brovelli et al. based on coherence analysis were discussed on p. 556 and were displayed on the left-hand side of Fig. 18.17. Those authors went on to fit an $AR(10)$ model to the data from both monkeys, noting that $AR(5)$ and $AR(15)$ gave consistent results, and that AIC (see Section 11.1.6) would select $AR(15)$ (they considered $AR(p)$ models up through order $p = 15$). They then applied Granger causality[18] analysis, which allowed them to produce the additional directional interpretations shown on the right-hand side of Fig. 18.17. In particular, beta rhythms in primary sensory cortex (site 2 in both monkeys) were predictive of the rhythms in other locations, while primary motor cortex (site 1) tended to be predicted by both sensory and AIP signals and was itself only weakly predictive of signals at other sites. □

[18] They used the spectral decomposition mentioned in the footnote on p. 559 to plot the frequency representation of Granger causality, found its peak, and performed a permutation test analogously to what they had done in analyzing coherence.