Investigating Economic Trends and Cycles

D.S.G. Pollock *

Methods are described for extracting the trend from an economic data sequence and for isolating the cycles that surround it. The latter often consist of a business cycle of variable duration and a perennial seasonal cycle.

There is no evident point in the frequency spectrum where the trend ends and the business cycle begins. Therefore, unless it can be represented by a simple analytic function, such as an exponential growth path, there is bound to be a degree of arbitrariness in the definition of the trend.

The business cycle, however defined, is liable to have an upper limit to its frequency range that falls short of the Nyquist frequency, which is the maximum observable frequency in sampled data. This must be taken into account in fitting an ARMA model to the detrended data.

* *Address:* Department of Economics, University of Leicester,
Leicester LE1 7RH, United Kingdom,

Email: stephen_pollock@sigmapi.u-net.com

1. Introduction

It has been traditional in economics to decompose time series—more accurately described as temporal sequences—into a variety of components, some or all of which may be present in a particular instance. The essential decomposition is a multiplicative one of the form

$$Y(t) = T(t) \times C(t) \times S(t) \times E(t), \tag{1}$$

where

$T(t)$ is the global trend,

$C(t)$ is a secular cycle, or business cycle,

$S(t)$ is the seasonal variation and

$E(t)$ is an irregular component.

Occasionally, other cycles of relatively long durations are included. Amongst these are the mysterious Kondratieff cycle, reflecting the ebb and flow of human fortunes over half a century, the Shumpeterian cycle, reflecting currents and tides of technological innovation, and the demographic cycle, reflecting the fluctuations in the procreative urges of human beings.

The factors $C(t)$, $S(t)$ and $E(t)$ in equation (1) serve to modulate the trend $T(t)$ by inducing fluctuations in its trajectory. They take the generic form of $X(t) = 1 + \xi(t)$, where $\xi(t)$ is a process that fluctuates about a mean of zero.

Typically, $Y(t)$ and $T(t)$ are strictly positive and, therefore, the modulating factors, which are usually deemed to act independently of each other, must also be bounded away from zero. This condition will be satisfied whenever the generic factor can be expressed in an exponential form:

$$X(t) = 1 + \xi(t) = 1 + \sum_{j=1}^{\infty} \frac{\{x(t)\}^j}{j!} = \exp\{x(t)\}. \tag{2}$$

In that case, it is appropriate to take logarithms of the expression (1) and to work with an alternative additive decomposition instead of the multiplicative one. This is

$$y(t) = \tau(t) + c(t) + s(t) + \varepsilon(t), \tag{3}$$

where $y(t) = \ln Y(t)$, $\tau(t) = \ln T(t)$, $c(t) = \ln C(t)$, $s(t) = \ln S(t)$ and $\varepsilon(t) = \ln E(t)$. An additional assumption, which might be plausible, is that the components $c(t)$, $s(t)$, and $\varepsilon(t)$ have amplitudes that remain roughly constant over time.

In the absence of extraneous information that correlates them with other variables, it is impossible to distinguish the components of (3) perfectly, one from another, unless they occupy separate frequency bands. If their bands do overlap, then any separation of the components will be tentative and doubtful. Thus, a sequence that is deemed to represent one of the components will comprise, to some extent, elements that rightfully belong to the other components.

However, as we shall see, the components of an econometric data sequence often reside within bands of frequencies that are separated by wide dead spaces

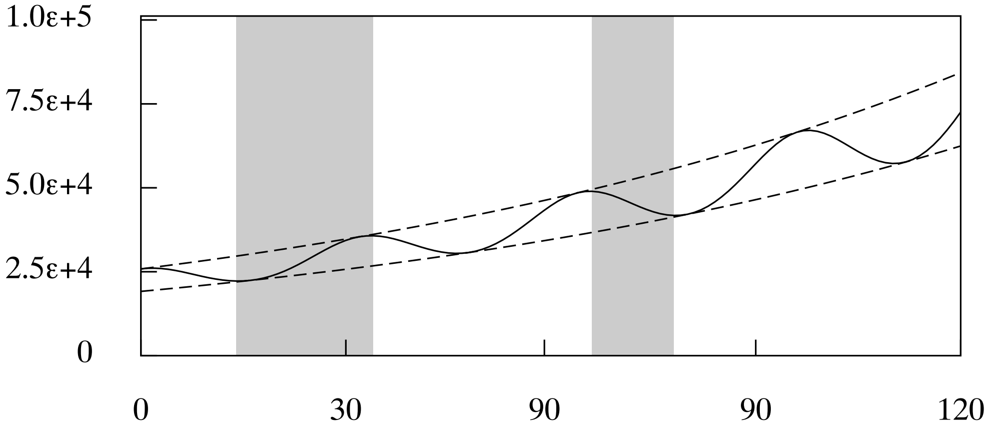

Figure 1. The function $Y(t) = \beta \exp\{rt + \gamma \cos(\omega t)\}$ as a model of the business cycle. Observe that, when $r > 0$, the duration of an expansion exceeds the duration of a contraction.

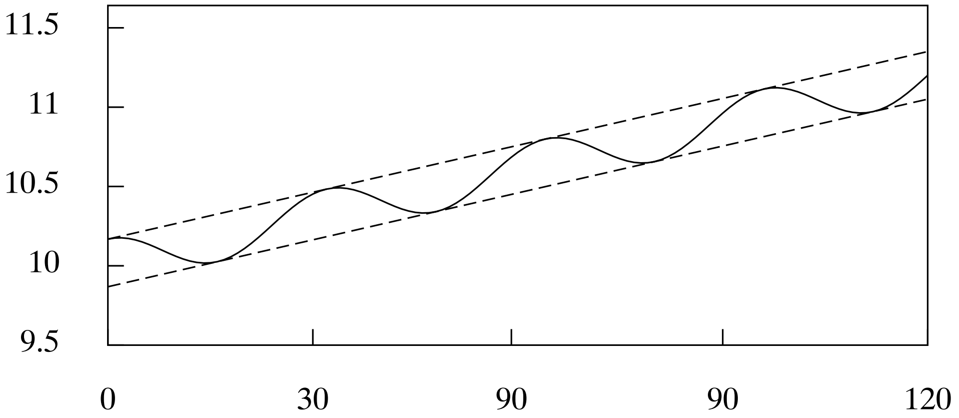

Figure 2. The function $\ln\{Y(t)\} = \ln\{\beta\} + rt + \gamma \cos(\omega t)$ representing the logarithmic business cycle data. The durations of the expansions and the contractions are not affected by the transformation.

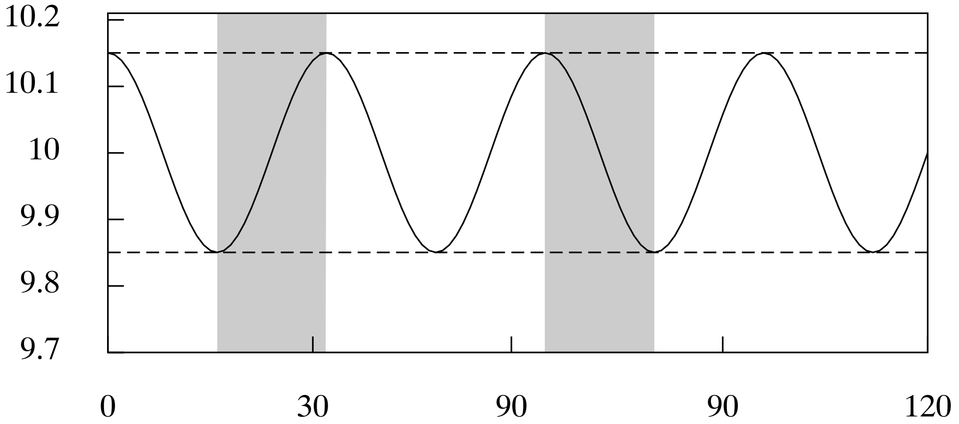

Figure 3. The function $\mu + \gamma \cos(\omega t)$ representing the detrended business cycle. The durations of the expansions and the contractions are equal.

where there are no spectral elements of any significance. The possibility of definitely separating the components is greater than analysts are likely to perceive unless they work in the frequency domain.

The exception concerns the separation of the business cycle from the trend. These components are liable to be merged within a single spectral structure; and there is no uniquely appropriate way of separating them. Their separation depends upon adopting whatever convention best suits the purposes of the analysis. No such difficulties will affect the simple schematic model of the business cycle that we shall consider in the next section.

2. A schematic model of the business cycle

In order to extract the modulating components from the data, it is also necessary to remove the trend component from $Y(t)$. To understand what is at issue in detrending the data, it is helpful to look at a simple schematic model comprising an exponential growth trajectory $T(t) = \beta\exp\{rt\}$, with $r > 0$, that is modulated by a exponentiated cosine function $C(t) = \exp\{\gamma\cos(\omega t)\}$ to create a model for the trajectory of aggregate income:

$$Y(t) = \beta\exp\{rt + \gamma\cos(\omega t)\}. \tag{4}$$

The resulting business cycles, which are depicted in Figure 1, have an asymmetric appearance. Their contractions are of lesser duration than their expansions; and they become shorter as the growth rate r increases.

Eventually, when the rate exceeds a certain value, the periods of contraction will disappear and, in place of the local minima, there will be only points of inflection. In fact, the condition for the existence of local minima is that $\omega\gamma > r$, which is to say the product of the amplitude of the cycles and their angular velocity must exceed the growth rate of the trend.

Next, we take logarithms of the data to obtain a model, represented in Figure 2, that has additive trend and cyclical components. This gives

$$\ln\{Y(t)\} = y(t) = \mu + rt + \gamma\cos(\omega t), \tag{5}$$

where $\mu = \ln\{\beta\}$. Since logs effect a monotonic transformation, there is no displacement of the local maxima and minima. However, the amplitude of the fluctuations around the trend, which has become linear in the logs, is now constant.

The final step is to create a stationary function by eliminating the trend. There are two equivalent ways of doing this in the context of the schematic model. On the one hand, the linear trend $\xi(t) = \mu + rt$ can be subtracted from $y(t)$ to create the pure business cycle $\gamma\cos(\omega t)$. Alternatively, the function $y(t)$ can be differentiated to give $dy(t)/dt = r - \gamma\omega\sin(\omega t)$. When the latter is adjusted by subtracting the growth rate r, by dividing by ω and by displacing its phase by $-\pi/2$ radians—which entails replacing the argument t by $t - \pi/2$—we obtain the function $\gamma\cos(\omega t)$ again. Through the process of detrending, the phases of expansion and contraction acquire equal durations, and the asymmetry of the business cycle vanishes.

There is an enduring division of opinion, in the literature of economics, on whether we should be looking at the turning points and phase durations of the original data or at those of the detrended data. The task of finding the turning points is often a concern of analysts who wish to make international comparisons of the timing of the business cycle. There is a belief, which bears investigating, that these cycles are becoming increasingly synchronised amongst member countries of the European Union.

However, since the business cycle is a low-frequency component of the data, it is difficult to find the turning points with great accuracy. In fact, the pinnacles and pits that are declared to be the turning points often seem to be the products of whatever high-frequency components happen to remain in the data after it has been subjected to a process of seasonal adjustment.

If the objective is to compare the turning points of the cycles, then the trends should be eliminated from the data. The countries that are to be compared are liable to be growing at differing rates. From the trended data, it will appear that those with higher rates of growth have shorter recessions with delayed onsets, and this can be misleading.

The various indices of an expanding economy will also grow at diverse rates. Unless they are reduced to a common basis by eliminating their trends, their fluctuations cannot be compared easily. Amongst such indices will be the percentage rate of unemployment, which constitutes a trend-stationary sequence. It would be difficult to collate the turning points in this index with those within a rapidly growing series of aggregate income, which might not exhibit any absolute reductions in its level. A trenchant opinion to the contrary, which opposes the practice of detrending the data for the purposes of describing the business cycle, has been offered by Harding and Pagan (2002).

3. The methods of Fourier analysis

A means of extracting the cyclical components from a data sequence is to regress it on a set of trigonometrical functions. The relevant procedures have been described within the context of the statistical analysis of time series by numerous authors, including Bloomfield (1975), Fuller (1976) and Priestley (1989).

In the Fourier decomposition of a finite sequence $\{x_t; t = 0, 1, \ldots, T-1\}$, the T data points are expressed as a weighted sum of an equal number of trigonometrical functions of frequencies that are equally spaced in the interval $[0, \pi]$.

We define $[T/2]$ to be the integer part to $T/2$, which will be $n = T/2$, if T is even, or $(T-1)/2$, if T is odd. Then

$$\begin{aligned} x_t &= \sum_{j=0}^{[T/2]} \Big\{ \alpha_j \cos(\omega_j t) + \beta_j \sin(\omega_j t) \Big\} \\ &= \sum_{j=0}^{[T/2]} \rho_j \cos(\omega_j t - \theta_j). \end{aligned} \tag{6}$$

Here, $\rho_j^2 = \alpha_j^2 + \beta_j^2$ and $\theta_j = \tan^{-1}(\beta_j/\alpha_j)$, whilst $\alpha_j = \rho_j \cos(\theta_j)$ and $\beta_j = \rho_j \sin(\theta_j)$. The equality of (6) follows in view of the trigonometrical identity

$$\cos(A - B) = \cos(A)\cos(B) + \sin(A)\sin(B). \tag{7}$$

The frequency $\omega_j = 2\pi j/T$ is a multiple of the fundamental frequency $\omega_1 = 2\pi/T$. The latter belongs to a sine and a cosine function that complete a single cycle in the time spanned by the data. The zero frequency ω_0 is associated with the constant function $\cos(\omega_0 t) = \cos(0) = 1$, whereas $\sin(\omega_0 t) = \sin(0) = 0$.

If $T = 2n$ is an even number, then the highest frequency is $\omega_n = \pi$; and, within (6), there are $\cos(\omega_n t) = \cos(\pi t) = (-1)^t$ and $\sin(\omega_n t) = \sin(\pi t) = 0$. If T is an odd number, then the highest frequency is $\pi(T-1)/T$, and there are both a sine and a cosine function at this frequency. Counting the number of nonzero functions in both cases shows that they are equal in number to the data points. Therefore, there is a one-to-one correspondence between the data points and the coefficients of the nonzero functions in the Fourier expression of (6).

In equation (6), the temporal index $t \in \{0, 1, \ldots, T-1\}$ assumes integer values. However, by allowing $t \in [0, T)$ to vary continuously, one can generate a continuous function that interpolates the T data points. This method of generating the continuous function from sampled values may be described as Fourier interpolation. It is notable that the interpolated function is analytic in the sense that it possesses derivatives of all orders.

Although the process generating the data may contain components of frequencies higher than the Nyquist frequency, these will not be detected when it is sampled regularly at unit intervals of time. In fact, the effects on the process of components of frequencies in excess of the Nyquist value will be confounded with those of frequencies that fall below it.

To demonstrate this, consider the case where the process contains a component that is a pure cosine wave of unit amplitude and zero phase and of a frequency ω that lies in the interval $\pi < \omega < 2\pi$. Let $\omega^* = 2\pi - \omega$. Then,

$$\begin{aligned}\cos(\omega t) &= \cos\big\{(2\pi - \omega^*)t\big\}\\ &= \cos(2\pi)\cos(\omega^* t) + \sin(2\pi)\sin(\omega^* t)\\ &= \cos(\omega^* t),\end{aligned} \tag{8}$$

which indicates that ω and ω^* are observationally indistinguishable. Here, $\omega^* < \pi$ is described as the alias of $\omega > \pi$.

Since the trigonometrical functions are mutually orthogonal, the Fourier coefficients can be obtained via a set of T simple inner-product formulae, which are in the form of ordinary univariate least-squares regressions, with the values of the sine and cosine functions at the points $t = 0, 1, \ldots, T-1$ as the regressors.

Let $c_j = [c_{0,j}, \ldots, c_{T-1,j}]'$ and $s_j = [s_{0,j}, \ldots, s_{T-1,j}]'$ represent vectors of T values of the generic functions $\cos(\omega_j t)$ and $\sin(\omega_j t)$ respectively, and let $x = [x_0, \ldots, x_{T-1}]'$ be the vector of the sample data and $\iota = [1, \ldots, 1]'$ a vector of units. The 'regression' formulae for the Fourier coefficients are

$$\alpha_0 = (\iota'\iota)^{-1}\iota' x = \frac{1}{T}\sum_t x_t = \bar{x}, \tag{9}$$

$$\alpha_j = (c_j' c_j)^{-1} c_j' x = \frac{2}{T}\sum_t x_t \cos(\omega_j t), \tag{10}$$

$$\beta_j = (s_j's_j)^{-1}s_j'x = \frac{2}{T}\sum_t x_t \sin(\omega_j t), \tag{11}$$

$$\alpha_n = (c_n'c_n)^{-1}c_n'x = \frac{1}{T}\sum_t (-1)^t x_t. \tag{12}$$

However, in calculating the coefficients, it is more efficient to use the family of specialised algorithms known as fast Fourier transforms, which deliver complex-valued spectral ordinates from which the Fourier coefficients are obtained directly. (See, for example, Pollock 1999.)

The power of a sequence is the time average of its energy. It is synonymous with the mean-square deviation which, in statistical terms, is its variance. The power of the sequence $x_j(t) = \rho_j \cos(\omega_j t)$ is $\rho_j^2/2$. This result can be obtained in view of the identity $\cos^2(\omega_j t) = \{1 + \cos(2\omega_j t)\}/2$, for the average of $\cos(2\omega_j t)$ over an integral number of cycles is zero. The assemblage of values $\rho_j^2/2; j = 1, 2, \ldots, [T/2]$ constitutes the power spectrum of $x(t)$, which becomes the periodogram when scaled by a factor T. Their sum equals the variance of the sequence. If $T = 2n$ is even, then

$$\frac{1}{T}\sum_{t=0}^{T-1}(x_t - \bar{x})^2 = \frac{1}{2}\sum_{j=1}^{n-1}\rho_j^2 + \alpha_n^2. \tag{13}$$

Otherwise, if T is odd, then the summation runs up to $(T-1)/2$, and the term α_n^2 is missing.

The indefinite sequence $x(t) = \{x_t; t = 0, \pm 1, \pm 2, \ldots\}$, expressed in the manner of (6), is periodic with a period T equal to the length of the sample. It is described as the periodic extension of the sample, and it may be obtained be replicating sample elements over all preceding and succeeding intervals of T points. An alternative way of forming the periodic sequence is by wrapping the sample around a circle of circumference T. Then, the periodic sequence is generated by travelling perpetually around the circle.

3.1 Approximations, resampling and Fourier interpolation.

By letting $t = 0, \ldots, T-1$ in equation (6), the data sequence $\{x_t; t = 0, \ldots, T-1\}$ is generated exactly. An approximation to the sequence may be generated by taking a partial sum comprising the terms of (6) that are associated with the Fourier frequencies $\omega_0, \ldots, \omega_d$, where $d < [T/2]$. It is straightforward to demonstrate that this is the best approximation, in the least-squares sense, amongst all of the so-called trigonometrical polynomials of degree d that comprise the sinusoidal functions in question.

The result concerning the best approximation extends to the continuous functions that are derived by allowing t to vary continuously in the interval $[0, T)$. That is to say, the continuous function derived from the partial Fourier sum comprising frequencies no higher than $\omega_d = 2\pi d/T$ is the minimum-mean-square approximation to the continuous function derived from (6) by letting t vary continuously.

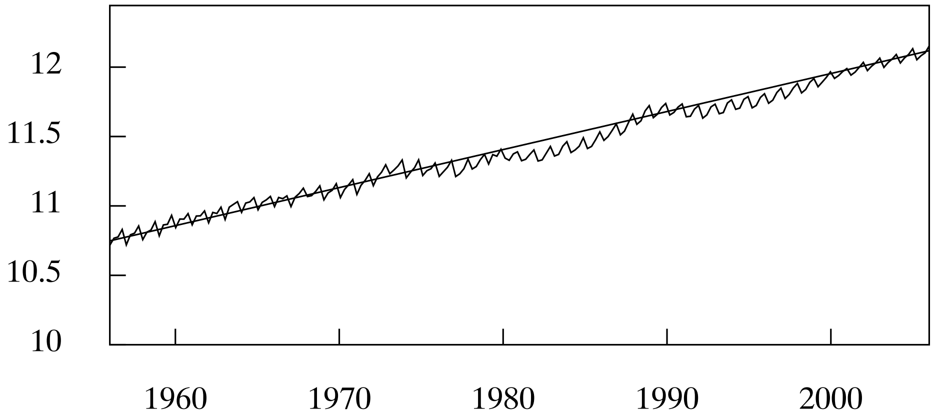

Figure 4. The quarterly sequence of the logarithms of household expenditure in the U.K. for the years 1956 to 2005, together with an interpolated linear trend.

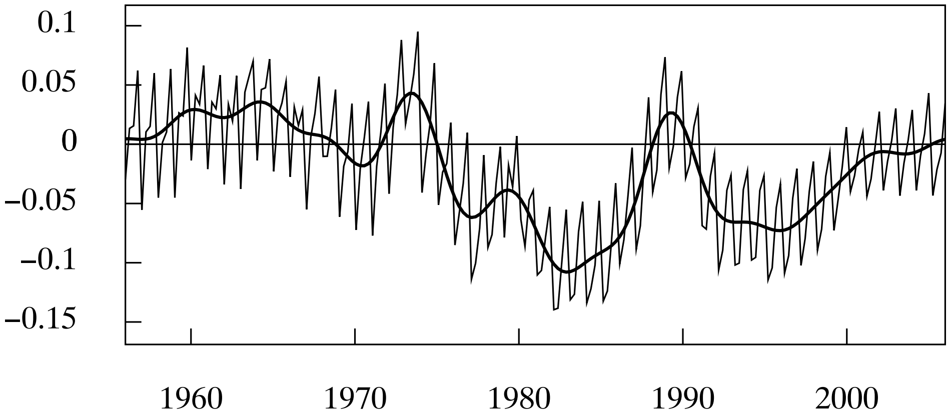

Figure 5. The residual deviations of the logarithmic expenditure data from the linear trend of Figure 4. The interpolated line, which represents the business cycle, has been synthesised from the Fourier ordinates in the frequency interval $[0, \pi/8]$.

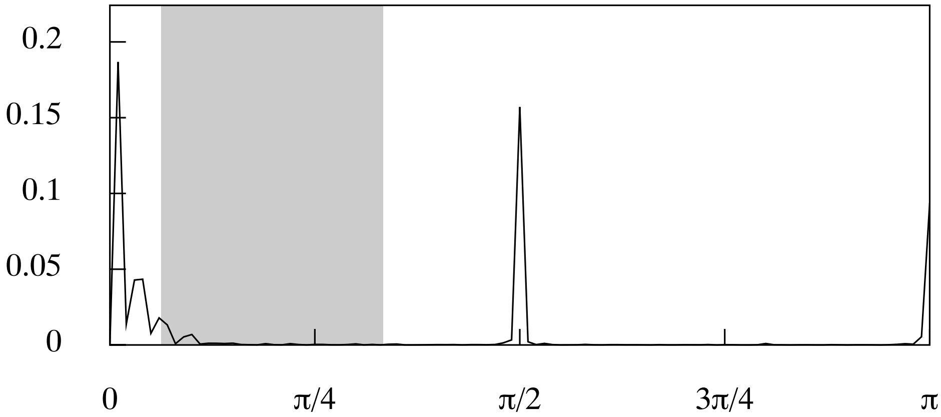

Figure 6. The periodogram of the residual sequence of Figure 5. A band, with a lower bound of $\pi/16$ radians and an upper bound of $\pi/3$ radians, is masking the periodogram.

We may exclude the sine function of frequency ω_d from the Fourier sum. Then, the continuous approximation is given by

$$\begin{aligned} z(t) &= \sum_{j=0}^{d} \left\{ \alpha_j \cos\left(\frac{2\pi jt}{T}\right) \right\} + \sum_{j=1}^{d-1} \left\{ \beta_j \sin\left(\frac{2\pi jt}{T}\right) \right\} \\ &= \sum_{j=0}^{d} \left\{ \alpha_j \cos\left(\frac{2\pi j\tau}{N}\right) \right\} + \sum_{j=1}^{d-1} \left\{ \beta_j \sin\left(\frac{2\pi j\tau}{N}\right) \right\}, \end{aligned} \tag{14}$$

where $\tau = tN/T$ with $N = 2d$, which is the total number of the Fourier coefficients. Here, τ varies continuously in $[0, N)$, whereas t varies continuously in $[0, T)$. On the RHS, there is a new set of Fourier frequencies $\{2\pi j/N; j = 0, 1, \ldots, d\}$.

The N coefficients $\{\alpha_0, \alpha_1, \beta_1, \ldots, \alpha_{d-1}, \beta_{d-1}, \alpha_d\}$ bear a one-to-one correspondence with the set of N ordinates $\{z_\tau = z(\tau T/N); \tau = 0, \ldots, N-1\}$ sampled at intervals of $\pi/\omega_d = T/N$ from $z(t)$. The consequence is that $z(t)$ is fully represented by the resampled data $z_\tau; \tau = 0, \ldots, N-1$, from which it may be derived by Fourier interpolation.

The result concerning the optimality of the approximation is a weak one; for it is possible that the preponderance of the variance of the data will be explained by sinusoids at frequencies that lie outside the range $[\omega_0, \ldots, \omega_d]$. The matter can be judged with reference to the periodogram of the data sequence, which constitutes a frequency-specific analysis of variance.

Example. Figure 4 represents the logarithms of the data on quarterly real household expenditure in the U.K. for the period 1956–2005, through which a linear function had been interpolated so as to pass through the midst of the data points of the first and the final years.

This interpolation is designed to minimise any disjunction that might otherwise occur where the ends data sequence meet, when it is mapped onto the circumference of a circle. A trend line fitted by ordinary least-squares regression would have a lesser gradient, which would raise the final years above the line. This would be a reflection of the relatively prosperity of the times.

The residual deviations of the expenditure data from the trend line of Figure 4 are represented in Figure 5, and their periodogram is in Figure 6. Within this periodogram, the spectral structure extending from zero frequency up to $\pi/8$ belongs to the business cycle. The prominent spikes located at the frequency $\pi/2$ and at the limiting Nyquist frequency of π are the property of the seasonal fluctuations. Elsewhere in the periodogram, there are wide dead spaces, which are punctuated by the spectral traces of minor elements of noise.

The slowly varying continuous function interpolated through the deviations of Figure 5 has been created by combining a set of sine and cosine functions of increasing frequencies in the manner of (14), with the frequencies extending no further than $\omega_d = \pi/8$, and by letting t vary continuously in the interval $[0, T)$. This is a representation of the business cycle as it affects household expenditure. Observe that, since it is analytic, the turning points of this function can be determined via its first derivative.

3.2 Complex exponentials

In dealing with the mathematics of the Fourier transform, it is common to use complex exponential functions in place of sines and cosines. This makes the expressions more concise. According to Euler's equations, there are

$$\cos(\omega_j t) = \frac{1}{2}(e^{\mathrm{i}\omega_j t} + e^{-\mathrm{i}\omega_j t}) \quad \text{and} \quad \sin(\omega_j t) = \frac{-\mathrm{i}}{2}(e^{\mathrm{i}\omega_j t} - e^{-\mathrm{i}\omega_j t}), \tag{15}$$

where $\mathrm{i} = \sqrt{-1}$. Therefore, equation (6) can be expressed as

$$x_t = \alpha_0 + \sum_{j=1}^{[T/2]} \frac{\alpha_j + \mathrm{i}\beta_j}{2} e^{-\mathrm{i}\omega_j t} + \sum_{j=1}^{[T/2]} \frac{\alpha_j - \mathrm{i}\beta_j}{2} e^{\mathrm{i}\omega_j t}, \tag{16}$$

which can be written concisely as

$$x_t = \sum_{j=0}^{T-1} \xi_j e^{\mathrm{i}\omega_j t}, \tag{17}$$

where

$$\xi_0 = \alpha_0, \quad \xi_j = \frac{\alpha_j - \mathrm{i}\beta_j}{2} \quad \text{and} \quad \xi_{T-j} = \xi_j^* = \frac{\alpha_j + \mathrm{i}\beta_j}{2}. \tag{18}$$

Equation (17) may be described as the inverse Fourier transform. The direct transform is the mapping from the data sequence within the time domain to the sequence of Fourier ordinates in the frequency domain. The relationship between the discrete periodic function and its Fourier transform can be summarised by writing

$$x_t = \sum_{j=0}^{T-1} \xi_j e^{\mathrm{i}\omega_j t} \quad \longleftrightarrow \quad \xi_j = \frac{1}{T}\sum_{t=0}^{T-1} x_t e^{-\mathrm{i}\omega_j t} dt. \tag{19}$$

For matrix representations of these transforms, one may define

$$\begin{aligned} U &= T^{-1/2}[\exp\{-\mathrm{i}2\pi tj/T\}; t, j = 0, \ldots, T-1], \\ \bar{U} &= T^{-1/2}[\exp\{\mathrm{i}2\pi tj/T\}; t, j = 0, \ldots, T-1], \end{aligned} \tag{20}$$

which are unitary complex matrices such that $U\bar{U} = \bar{U}U = I_T$. Then,

$$x = T^{1/2}\bar{U}\xi \quad \longleftrightarrow \quad \xi = T^{-1/2}Ux, \tag{21}$$

where $x = [x_0, x_1, \ldots x_{T-1}]'$ and $\xi = [\xi_0, \xi_1, \ldots \xi_{T-1}]'$ are the vectors of the data and of their Fourier ordinates respectively.

4. Spectral representations of a stationary process

The various equations of the Fourier analysis of a finite data sequence can also be used to describe the processes that generate the data. Thus, within the equation

$$\begin{aligned} y_t &= \sum_{j=0}^{n} \{\alpha_j \cos(\omega_j t) + \beta_j \sin(\omega_j t)\} \\ &= \zeta_0 + \sum_{j=1}^{n} \left\{\zeta_j e^{\mathrm{i}\omega_j t} + \zeta_j^* e^{-\mathrm{i}\omega_j t}\right\}, \end{aligned} \tag{22}$$

the quantities α_j, β_j can be taken to represent independent real-valued random variables, and the quantities

$$\zeta_j = \frac{\alpha_j - \mathrm{i}\beta_j}{2} \qquad \text{and} \qquad \zeta_j^* = \frac{\alpha_j + \mathrm{i}\beta_j}{2} \tag{23}$$

can be regarded as complex-valued random variables.

The autocovariance of the elements y_t and y_s is given by

$$\begin{aligned} E(y_t y_s) = \sum_{j=0}^{n}\sum_{k=0}^{n} E\Big[&\zeta_j\zeta_k e^{\mathrm{i}(\omega_j t+\omega_k s)} + \zeta_j\zeta_k^* e^{\mathrm{i}(\omega_j t-\omega_k s)} \\ &+\zeta_j^*\zeta_k e^{\mathrm{i}(\omega_k s-\omega_j t)} + \zeta_j^*\zeta_k^* e^{-\mathrm{i}(\omega_j t+\omega_k s)}\Big]. \end{aligned} \tag{24}$$

The condition of stationarity requires that the covariance should be a function only of the temporal separation $|t-s|$ of y_t and y_s. For this, it is necessary that

$$E(\zeta_j\zeta_k) = E(\zeta_j^*\zeta_k^*) = E(\zeta_j^*\zeta_k) = E(\zeta_j\zeta_k^*) = 0, \tag{25}$$

whenever $j \neq k$. Also, the conditions

$$E(\zeta_j^2) = 0 \quad \text{and} \quad E(\zeta_j^{*2}) = 0 \tag{26}$$

must hold for all j. For (25) and (26) to hold, it is sufficient that

$$E(\alpha_j\beta_k) = 0 \quad \text{for all} \quad j, k \tag{27}$$

and that

$$E(\alpha_j\alpha_k) = E(\beta_j\beta_k) = \begin{cases} 0, & \text{if } j \neq k; \\ \sigma_j^2, & \text{if } j = k. \end{cases} \tag{28}$$

An implication of the equality of the variances of α_j and β_j is that the phase angle θ_j is uniformly distributed in the interval $[-\pi, \pi]$.

Under these conditions, the autocovariance of the process at lag $\tau = t - s$ will be given by

$$\gamma_\tau = \sum_{j=0}^{n} \sigma_j^2 \cos\omega_j\tau. \tag{29}$$

The variance of the process is just

$$\gamma_0 = \sum_{j=0}^{n} \sigma_j^2, \tag{30}$$

which is the sum of the variances of the n individual periodic components. This is analogous to equation (13)

The stochastic model of equation (22) may be extended to encompass processes defined over the entire set of positive and negative integers as well as processes that are continuous in time. First, we may consider extending the length T of the sample indefinitely. As T and n increase, the Fourier coefficients become more numerous and more densely packed in the interval $[0, \pi]$. Also, given that the variance of the process is bounded, the variance of the individual coefficients must decrease.

To accommodate these changes, we may write $\alpha_j = dA(\omega_j)$ and $\beta_j = dB(\omega_j)$, where $A(\omega)$, $B(\omega)$ are cumulative step functions with discontinuities at the points $\{\omega_j; j = 0, \ldots, n\}$. In the limit, the summation in (22) is replaced by an integral, and the expression becomes

$$\begin{aligned} y(t) &= \int_0^{\pi} \{\cos(\omega t) dA(\omega) + \sin(\omega t) dB)\omega)\} \\ &= \int_{-\pi}^{\pi} e^{\mathrm{i}\omega t} dZ(\omega), \end{aligned} \tag{31}$$

where

$$\begin{aligned} dZ(\omega) &= \frac{1}{2}\{dA(\omega) - \mathrm{i}dB(\omega)\} \quad \text{and} \\ dZ(-\omega) = dZ^*(\omega) &= \frac{1}{2}\{dA(\omega) + \mathrm{i}dB(\omega)\}. \end{aligned} \tag{32}$$

Also, $y(t) = \{y_t; t = 0, \pm 1, \pm 2, \ldots\}$ stands for a doubly-infinite data sequence.

The assumptions regarding $dA(\omega)$ and $dB(\omega)$ are analogous to those regarding the random variables α_j and β_j, which are their prototypes. It is assumed that $A(\omega)$ and $B(\omega)$ represent a pair of stochastic processes of zero mean, which are indexed on the continuous parameter ω. Thus

$$E\big\{dA(\omega)\big\} = E\big\{dB(\omega)\big\} = 0. \tag{33}$$

It is also assumed that the two processes are mutually uncorrelated and that non overlapping increments within each process are uncorrelated. Thus

$$\begin{aligned} E\big\{dA(\omega)dB(\lambda)\big\} &= 0 \quad \text{for all} \quad \omega, \lambda, \\ E\big\{dA(\omega)dA(\lambda)\big\} &= 0 \quad \text{if} \quad \omega \neq \lambda, \\ E\big\{dB(\omega)dB(\lambda)\big\} &= 0 \quad \text{if} \quad \omega \neq \lambda. \end{aligned} \tag{34}$$

The variance of the increments is given by

$$V\big\{dA(\omega)\big\} = V\big\{dB(\omega)\big\} = 2dF(\omega). \tag{35}$$

The function $F(\omega)$, which is defined provisionally over the interval $[0, \pi]$, is described as the spectral distribution function. The properties of variances imply that it is a non decreasing function of ω. In the case where the process $y(t)$ is purely random, $F(\omega)$ is a continuous differentiable function. Its derivative $f(\omega)$, which is nonnegative, is described as the spectral density function.

The domain of the functions $A(\omega)$, $B(\omega)$ may be extended from $[0, \pi]$ to $[-\pi, \pi]$ by regarding $A(\omega)$ as an even function such that $A(-\omega) = A(\omega)$ and by regarding $B(\omega)$ as an odd function such that $B(-\omega) = -B(\omega)$. Then, $dZ^*(\omega) = dZ(-\omega)$, in accordance with (32). From the conditions of (34), it follows that

$$\begin{aligned} E\big\{dZ(\omega)dZ^*(\lambda)\big\} &= E\big\{dZ(\omega)dZ(-\lambda)\big\} = 0 \quad \text{if} \quad \omega \neq \lambda, \\ E\big\{dZ(\omega)dZ^*(\omega)\big\} &= E\big\{dZ(\omega)dZ(-\omega)\big\} = dF(\omega), \end{aligned} \tag{36}$$

where the domain of $F(\omega)$ is now the interval $[-\pi, \pi]$

The sequence of the autocovariances of the process $y(t)$ may be expressed in terms of the spectrum of the process. From (36), it follows that the autocovariance of $y(t)$ at lag $\tau = t - s$ is given by

$$\begin{aligned} \gamma_\tau = C(y_t, y_s) &= E\bigg\{ \int_\omega e^{\mathrm{i}\omega t} dZ(\omega) \int_\lambda e^{\mathrm{i}\lambda s} dZ(\lambda) \bigg\} \\ &= \int_\omega \int_\lambda e^{\mathrm{i}\omega t} e^{\mathrm{i}\lambda s} E\big\{dZ(\omega)dZ(\lambda)\big\} \\ &= \int_\omega e^{\mathrm{i}\omega\tau} E\big\{dZ(\omega)dZ^*(\omega)\big\} \\ &= \int_{-\pi}^{\pi} e^{\mathrm{i}\omega\tau} dF(\omega). \end{aligned} \tag{37}$$

In the case of a continuous spectral distribution function, we may write $dF(\omega) = f(\omega)d\omega$ in the final expression, where $f(\omega)$ is the spectral density function. If $f(\omega) = \sigma^2/2\pi$, then there is $\gamma_0 = \sigma^2$ and $\gamma_\tau = 0$ for all $\tau \neq 0$, which are the characteristics of a white-noise process comprising a sequence of independently and identically distributed random variables. Thus, a white-noise process has a uniform spectral density function.

The second way of extending the model is to allow the rate of sampling to increase indefinitely. In the limit, the sampled sequence becomes a continuum. Equation (31) will serve to represent a continuous process on the understanding that t is now a continuous variable. However, if the discrete-time process has been subject to aliasing, then the range of the frequency integral will increase as the rate of sampling increases.

Under any circumstances, it seems reasonable to postulate an upper limit to the range of the frequencies comprised by a stochastic process. However, within the conventional theory of continuous stochastic processes, it is common to consider an unbounded range of frequencies. In that case, we obtain a spectral representation of a stochastic process of the form

$$y(t) = \int_{-\infty}^{\infty} e^{\mathrm{i}\omega t} dZ(\omega). \tag{38}$$

This representation is capable, nevertheless, of subsuming a process that is limited in frequency. If the bandwidth of $Z(\omega)$ is indeed unbounded, then (38) becomes the spectral representation of a process comprising a continuous succession of infinitesimal impacts, which generates a trajectory that is everywhere continuous but nowhere differentiable.

Example. Figure 7 shows the spectral density function of an autoregressive moving-average ARMA(2, 2) process $y(t)$, described by the equation $\alpha(z)y(z) = \mu(z)\varepsilon(z)$, where $\alpha(z)$ and $\mu(z)$ are quadratic polynomials and $y(z)$ and $\varepsilon(z)$ are, respectively, the z-transforms of the data sequence $y(t) = \{y_t; t = 0, \pm 1, \pm 2, \ldots\}$ and of a white-noise sequence $\varepsilon(t) = \{\varepsilon_t; t = 0, \pm 1, \pm 2, \ldots\}$ of independently and identically distributed random variables.

The ARMA(2, 2) process has been formed by the additive combination a second-order autoregressive AR(2) process and an independent white-noise process. The autoregressive polynomial is $\alpha(z) = 1 + 2\rho\cos(\theta)z + \rho^2z^2$, which has conjugate complex roots of which the polar forms are $\rho\exp\{\pm \mathrm{i}\theta\}$. In the example, the modulus of the roots is $\rho = 0.9$ and their argument is $\theta = \pi/4$ radians.

The spectral density function attains a non-zero minimum at $\omega = \pi$. However, it is possible to decompose the ARMA(2, 2) process into an ARMA(2, 1) process and a white-noise component that has the maximum variance compatible with such a decomposition. This is a so-called canonical decomposition of the ARMA process. The moving-average polynomial of the resulting ARMA(2, 1) process is $1 + z$, which has a zero at $\omega = \pi$. By maximising the variance of the white-noise component, an ARMA component is derived that is as smooth and as regular as possible.

Canonical decompositions are entailed in a method for extracting unobserved components from data sequences described by ARIMA models, which will be discussed in section 6.3.

Figure 7 also shows a periodogram that has been calculated from a sample of 256 points generated by the ARMA(2, 2) process. Its volatility contrasts markedly with the smoothness of the spectrum. The periodogram has half as many ordinates as the data sequence and it inherits this volatility directly from the data. A non-parametric estimate of the spectrum may be obtained by smoothing the ordinates of the periodogram with an appropriately chosen moving average, or by subjecting the empirical autocovariances to an equivalent weighting operation before transforming them to the frequency domain.

4.1 The frequency-domain analysis of filtering

It is a straightforward matter to derive the spectrum of a process $y(t)$ formed by mapping the process $x(t)$ through a linear filter. If

$$x(t) = \int_{\omega} e^{\mathrm{i}\omega t} dZ_x(\omega), \tag{39}$$

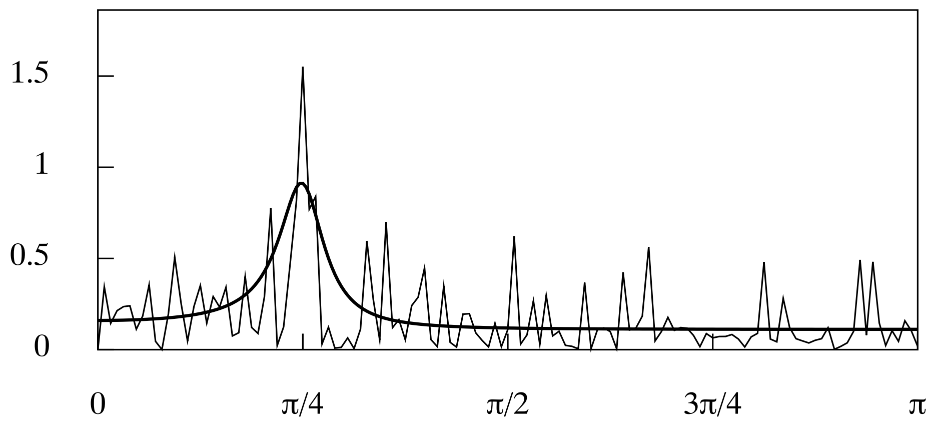

Figure 7. The periodogram of 256 points of a pseudo-random ARMA(2, 2) process overlaid by the spectral density function of the process.

then the filtered process is

$$\begin{aligned} y(t) &= \sum_j \psi_j x(t-j) \\ &= \sum_j \psi_j \left\{ \int_\omega e^{\mathrm{i}\omega(t-j)} dZ_x(\omega) \right\} \\ &= \int_\omega e^{\mathrm{i}\omega t} \left(\sum_j \psi_j e^{-\mathrm{i}\omega j} \right) dZ_x(\omega). \end{aligned} \tag{40}$$

On writing $\sum \psi_j e^{-\mathrm{i}\omega j} = \psi(\omega)$, which is the frequency response function of the filter, this becomes

$$\begin{aligned} y(t) &= \int_\omega e^{\mathrm{i}\omega t} \psi(\omega) dZ_x(\omega) \\ &= \int_\omega e^{\mathrm{i}\omega t} dZ_y(\omega). \end{aligned} \tag{41}$$

If the process $x(t)$ has a spectral density function $f_x(\omega)$, which will allow one to write $dF(\omega) = f(\omega)d\omega$ in equation (36), then the spectral density function $f_y(\omega)$ of the filtered process $y(t)$ will be given by

$$\begin{aligned} f_y(\omega)d\omega &= E\big\{dZ_y(\omega)dZ_y^*(\omega)\big\} \\ &= \psi(\omega)\psi^*(\omega)E\big\{dZ_x(\omega)dZ_x^*(\omega)\big\} \\ &= |\psi(\omega)|^2 f_x(\omega)d\omega. \end{aligned} \tag{42}$$

The complex-valued frequency-response function $\psi(\omega)$, which characterises the linear filter, can be written in polar form as

$$\psi(\omega) = |\psi(\omega)| e^{-\mathrm{i}\theta(\omega)}, \tag{43}$$

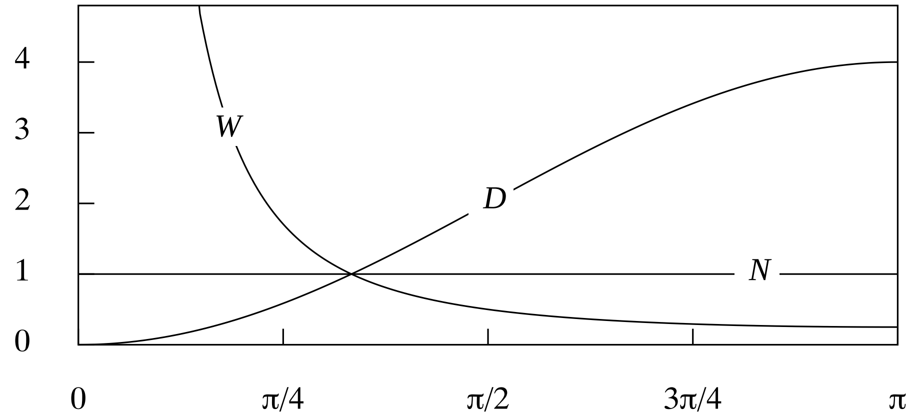

Figure 8. The squared gain of the difference operator, labelled D, and that of the summation operator, labelled W.

The function $|\psi(\omega)|$, which is described as the gain of the filter, indicates the extent to which the amplitude of the cyclical components of which $x(t)$ is composed are altered in the process of filtering.

When $x(t) = \varepsilon(t)$ is a white-noise sequence of independently and identically distributed random variables of variance σ^2, equation (42) gives rise to the expression $f_y(\omega) = \sigma^2|\psi(\omega)|^2 = \sigma^2\psi(\omega)\psi^*(\omega)$, which is the spectral density function of $y(t)$. Then, it is helpful to use the notation of the z-transform whereby $\psi(\omega)$ is written as $\psi(z) = \sum_j \psi_j z^j; z = e^{-\mathrm{i}\omega}$. If we allow z to be an arbitrary complex number, then we can define the autocovariance generating function $\gamma(z) = \sum_\tau \gamma_\tau z^\tau$ wherein $\gamma_\tau = (y_t y_{t-\tau})$. This takes the form of

$$\gamma(z) = \sigma^2\psi(z)\psi(z^{-1}). \tag{44}$$

Example. Figure 8 depicts the squared gain of the difference operator $\nabla(z) = 1 - z$, which is the curve labelled D. The squared gain of $\nabla(z)$ is obtained by setting $z = \exp\{-\mathrm{i}\omega\}$ within $|\nabla(z)|^2 = (1 - z)(1 - z^{-1})$ to give $D(\omega) = 2 - 2\cos(\omega)$, whence $W(\omega) = D^{-1}(\omega)$ can be obtained, which is the squared gain of the summation operator. The product of $D(\omega)$ and $W(\omega)$ is the constant function $N(\omega) = 1$, which also represents the spectral density function or power spectrum of a white-noise process with a variance of $\sigma^2 = 2\pi$. Likewise, $W(\omega)$ represents the pseudo-spectrum of a first-order random walk.

This is not a well-defined spectral density function, since the random walk does not constitute a stationary process of a sort that can be defined over a doubly-infinite set of time indices. The unbounded nature of $W(\omega)$ as $\omega \to 0$ is a testimony to the fact that the variance of the random walk process is proportional to time that has elapsed since its start-up. The variance will be unbounded if the start-up is in the indefinite past.

5. Stochastic accumulation

In the schematic model of the economy, we have envisaged business cycle fluctuations that are purely sinusoidal; and we have considered a trend that follows an exponential growth path. In a realistic depiction of an economy, both of these functions are liable to be more flexible and more variable through time.

Whereas, in some eras, a linear function, interpolated by least-squares regression through the logarithms of the data, will serve as a benchmark about which to measure the cyclical economic activities, the latter usually require to be modelled by a stochastic process. It is arguable that the trend should also be modelled by a stochastic function.

A further feature of the schematic model, which is at odds with the available data, is the continuous nature of its functions. Whereas the processes that generate the data can be thought of as operating in continuous time, the sampled data are sequences of values that are indexed by dates at equal intervals. These data are liable to be modelled via discrete-time stochastic processes. Therefore, some attention needs be paid to the relationship between the discrete data and the underlying continuous process.

The theory of continuous-time stochastic models has been summarised by Bergstrom (1984, 1988), who researched the subject over a 40-year period, beginning in the mid 1960's. His posthumous contributions are to be found in Bergstrom and Nowman (2007), where the contributions of other authors are also referenced.

A linear stochastic process must have a *primum mobile* or forcing function, which is liable to be a stationary process. For the usual discrete-time processes, this is a white-noise sequence of independently and identically distributed random variables. In the theory of continuous stochastic processes, the forcing function consists, almost invariably, of the increments of a Wiener process, which is a process that has an infinite bandwidth in the frequency domain. Already, in section 3, we have encountered a process with a limited bandwidth. Later, in section 9, we shall consider some further implications of a limited bandwidth.

The Wiener process $Z(t)$ is defined by the following conditions:

(a) $Z(0)$ is finite,

(b) $E\{Z(t)\} = 0,$ for all t,

(c) $Z(t)$ is normally distributed,

(d) $dZ(s), dZ(t)$ for all $t \neq s$ are independent stationary increments,

(e) $V\{Z(t+h) - Z(t)\} = \sigma^2 h$ for $h > 0$.

The increments $dZ(s), dZ(t)$ are impulses that have a uniform power spectrum distributed over an infinite range of frequencies corresponding to the entire real line. Sampling $Z(t)$ at regular intervals to form a discrete-time white-noise process $\varepsilon(t) = Z(t+1) - Z(t)$ entails a process of aliasing whereby the spectral power of the cumulated increments gives rise to a uniform spectrum of finite power over the frequency interval $[-\pi, \pi]$.

In general,

$$Z(t) = Z(a) + \int_a^t dZ(\tau), \tag{45}$$

where $Z(a)$ is a finite starting value at time a. However, if $Z(t)$ were differentiable, as some forcing functions may be, then we should have $dZ(t) = \{dZ(t)/dt\}dt$.

The simplest of stochastic differential equations is the first-order equation, which takes the form

$$\frac{dx(t)}{dt} - \lambda x(t) = dZ(t) \qquad \text{or} \qquad (D - \lambda)x(t) = dZ(t). \tag{46}$$

Multiplying throughout by the factor $\exp\{-\lambda t\}$ gives

$$e^{-\lambda t}Dx(t) - \lambda e^{-\lambda t}x(t) = D\{x(t)e^{-\lambda t}\} = e^{-\lambda t}dZ(t), \tag{47}$$

where the first equality follows from the product rule of differentiation. Integrating $D\{x(t)e^{-\lambda t}\} = e^{-\lambda t}dZ(t)$ gives

$$x(t)e^{-\lambda t} = \int_{-\infty}^{t} e^{-\lambda\tau}dZ(\tau) \tag{48}$$

or

$$x(t) = e^{\lambda t}\int_{-\infty}^{t} e^{-\lambda\tau}dZ(\tau) = \int_{-\infty}^{t} e^{\lambda(t-\tau)}dZ(\tau). \tag{49}$$

If we write $x(t) = (D - \lambda)^{-1}dZ(t)$, then we get the result that

$$x(t) = \frac{1}{D - \lambda}dZ(t) = \int_{-\infty}^{t} e^{\lambda(t-\tau)}dZ(\tau), \tag{50}$$

from which it is manifest that the necessary and sufficient condition for stability is that $\lambda < 0$. That is to say, the root of the equation $D - \lambda = 0$, which indicates the rate of decay of the increments, must be less than zero.

The general solution of a differential equation should normally comprise a particular solution, which represents the effects of the initial conditions. However, given that their effects decay as time elapses and given that, in this case, the integral has no lower limit, no account needs to be taken of initial conditions.

When the process is observed at the integer time points $\{t = 0, \pm 1, \pm 2, \ldots\}$, it is appropriate to express it as

$$\begin{aligned} x(t) &= e^{\lambda}\int_{-\infty}^{t-1} e^{\lambda(t-1-\tau)}dZ(\tau) + \int_{t-1}^{t} e^{\lambda(t-\tau)}dZ(\tau) \\ &= e^{\lambda}x(t-1) + \int_{t-1}^{t} e^{\lambda(t-\tau)}dZ(\tau). \end{aligned} \tag{51}$$

This gives rise to a discrete-time equation of the form

$$x(t) = \phi x(t-1) + \varepsilon(t), \quad \text{or} \quad (1 - \phi L)x(t) = \varepsilon(t), \tag{52}$$

where

$$\phi = e^{\lambda} \qquad \text{and} \qquad \varepsilon(t) = \int_{t-1}^{t} e^{\lambda(t-\tau)}dZ(\tau), \tag{53}$$

and where L is the lag operator, which has the effect that $Lx(t) = x(t-1)$.

The second-order equation may be expressed as follows:

$$(D^2 + \varphi_1 D + \varphi_2)x(t) = (D - \lambda_1)(D - \lambda_2)x(t) = dZ(t). \tag{54}$$

Using a partial-fraction expansion, this can be cast in the form of

$$\begin{aligned} x(t) &= \frac{1}{\lambda_1 - \lambda_2}\left\{\frac{1}{D - \lambda_1} - \frac{1}{D - \lambda_2}\right\} dZ(t) \\ &= \int_{-\infty}^{t} \left\{\frac{e^{\lambda_1(t-\tau)} - e^{\lambda_2(t-\tau)}}{\lambda_1 - \lambda_2}\right\} dZ(\tau). \end{aligned} \tag{55}$$

Here, the final equality depends upon the result under (50). If the roots λ_1, λ_2 have real values, then the condition of stability is that $\lambda_1, \lambda_2 < 0$. If the roots are conjugate complex numbers, then the condition for stability is that they must lie in the left half of the complex plane. In that case, the trajectory of $x(t)$ will have a damped quasi-sinusoidal motion of a sort that is characteristic of the business cycle.

Equation (55) gives rise to a second-order difference equation. In the manner that equation (50) leads to equation (52), we get

$$\begin{aligned} x(t) &= \frac{1}{\lambda_1 - \lambda_2}\left\{\frac{\varepsilon_1(t)}{1 - \kappa_1 L} + \frac{\varepsilon_2(t)}{1 - \kappa_2 L}\right\} \\ &= \frac{\theta_0 + \theta_1 L}{1 + \phi_1 L + \phi_2 L}\varepsilon(t). \end{aligned} \tag{56}$$

Here, $(\lambda_1 - \lambda_2)(1 - \kappa_1 L)(1 - \kappa_2 L) = 1 + \phi_1 L + \phi_2 L$, and we have defined $(\theta_0 + \theta_1 L)\varepsilon(t) = (1 - \phi_2 L)\varepsilon_1(t) + (1 - \phi_1 L)\varepsilon_2(t)$, which is a first-order moving-average process. Equation (56) depicts an ARMA$(2, 1)$ process in discrete time. The correspondence between the second-order differential equation and the ARMA$(2, 1)$ process has been discussed by Phadke and Wu (1974) and by Pandit and Wu (1975).

Autoregressive models of other orders may be derived in the same manner as the second-order model by putting polynomial functions of D of the appropriate degrees in place of the quadratic function. The models can also be elaborated by applying a moving-average operator or weighting function $\rho(r)$ to the stochastic forcing function $dZ(t)$. This gives a forcing function in the form of

$$\eta(t) = \int_0^q \rho(\tau) dZ(t - \tau) = \int_{t-q}^{t} \rho(t - \tau) dZ(\tau). \tag{57}$$

The consequence of this elaboration for the corresponding discrete-time ARMA model is that its moving-average parameters are no longer constrained to be functions of the autoregressive parameters alone.

In modelling a stochastic trend, it is common to adopt a first or second-order process in which the roots are set to zeros. In that case, the stochastic increments are accumulated without decay. Therefore, it is crucial to specify the initial conditions of the process. We shall denote the process that is the m-fold integral of the incremental process $dZ(t)$ by $Z^{(m)}(t)$. Then, $Z^{(1)}(t)$ can stand for the Wiener process $Z(t)$, defined previously.

If the process has begun in the indefinite past, then there will be zero probability that its current value will be found within a finite distance from the origin. Therefore, we must impose the condition that, at any time that is at a finite distance both from the origin and from the current time, the process $Z^{(1)}(t)$ assumes a finite value. This allows us to write

$$Z^{(1)}(t) = Z^{(1)}(t-h) + \int_{t-h}^{t} dZ^{(1)}(\tau), \tag{58}$$

where h is an arbitrary finite step in time and $a = t - h$ is a fixed point in time.

On this basis, the value of the integrated process at time t is

$$\begin{aligned} Z^{(2)}(t) &= Z^{(2)}(t-h) + \int_{t-h}^{t} Z^{(1)}(\tau)d\tau \\ &= Z^{(2)}(t-h) + Z^{(1)}(t-h)h + \int_{t-h}^{t} (t-\tau)dZ^{(1)}(\tau). \end{aligned} \tag{59}$$

By proceeding through successive stages, we find that the mth integral is

$$Z^{(m)}(t) = \sum_{k=0}^{m-1} Z^{(m-k)}(t-h)\frac{h^k}{k!} + \int_{t-h}^{t} \frac{(t-\tau)^{m-1}}{(m-1)!}dZ^{(1)}(\tau). \tag{60}$$

Here, the first term on the RHS is a polynomial in h, which is the distance in time from the fixed point a, whereas the second term is the m-fold integral of mean-zero stochastic increments, which constitutes a non-stationary process.

The covariance of the changes $Z^{(j)}(t) - Z^{(j)}(t-h)$ and $Z^{(k)}(t) - Z^{(k)}(t-h)$ of the jth and the kth integrated processes derived from $Z(t)$ is given by

$$\begin{aligned} C\{z^{(j)}(t), z^{(k)}(t)\} &= \int_{s=t-h}^{t} \int_{r=t-h}^{t} \frac{(t-r)^{j-1}(t-s)^{k-1}}{j!k!} E\{dZ(r)dZ(s)\} \\ &= \sigma^2 \int_{t-h}^{t} \frac{(t-r)^{j+k-2}}{j!k!}dr = \sigma^2 \frac{h^{j+k-1}}{(j+k-1)j!k!}. \end{aligned} \tag{61}$$

A straightforward elaboration of the model of a stochastic trend arises when it is assumed that the expected value of the incremental process that is the forcing function has a nonzero mean. Then, $Z(t)$ is replaced by $\mu dt + dZ(t)$. This is the case of stochastic drift. If μ is relatively large, then it will make a significant contribution to the polynomial component, with the effect that the latter may become the dominant component.

5.1 Discrete-time representation of an integrated Wiener process

To derive the discretely sampled version of the integrated Wiener process, it may be assumed that values are sampled at regular intervals of h time units.

Then, using the alternative notation of $\beta(t) = Z^{(1)}(t)$, equation (58) can be written as

$$\beta(t) = \beta(t-h) + \varepsilon(t), \tag{62}$$

where $\varepsilon(t)$ is a white-noise process. With $\tau(t) = Z^{(2)}(t)$, equation (59) can be written as

$$\tau(t) = \tau(t-h) + h\beta(t-h) + \nu(t), \tag{63}$$

where $\nu(t)$ is another white-noise process. Together, the equations (62) and (63) constitute a so-called local linear model in which $\tau(t)$ represents the level and $\beta(t)$ represents the slope parameter. On taking the step length to be $h = 1$, the transition equation for this model is

$$\begin{bmatrix} \tau(t) \\ \beta(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \tau(t-1) \\ \beta(t-1) \end{bmatrix} + \begin{bmatrix} \nu(t) \\ \varepsilon(t) \end{bmatrix}. \tag{64}$$

Using the difference operator $\nabla = 1-L$, the discrete-time processes entailed in this equation can be written as

$$\begin{aligned} \nabla\tau(t) &= \tau(t) - \tau(t-1) = \beta(t-1) + \nu(t), \\ \nabla\beta(t) &= \beta(t) - \beta(t-1) = \varepsilon(t). \end{aligned} \tag{65}$$

Applying the difference operator a second time to the first of these and substituting for $\nabla\beta(t) = \varepsilon(t)$ gives

$$\begin{aligned} \nabla^2\tau(t) &= \nabla\beta(t-1) + \nabla\nu(t) \\ &= \varepsilon(t-1) + \nu(t) - \nu(t-1). \end{aligned} \tag{66}$$

On the RHS of this equation is a sum of stationary stochastic processes, which can be expressed as an ordinary first-order moving-average process. Thus

$$\varepsilon(t-1) + \nu(t) - \nu(t-1) = \eta(t) + \theta\eta(t-1), \tag{67}$$

where $\eta(t)$ is a white-noise process with $V\{\eta(t)\} = \sigma^2_\eta$. Therefore, the sampled version of the integrated Wiener process is an doubly-integrated IMA(2, 1) moving-average model.

The essential task is to find the values of the moving-average parameter θ. Thus is achieved by reference to equation (61), which provides the variances and covariances of the terms on the LHS of (67), from which the autocovariances of the MA process can be found. It can be shown that the variance and the autocovariance at lag 1 of this composite process are given by

$$\gamma_0 = \frac{2\sigma^2_\varepsilon}{3} = \sigma^2_\eta(1+\theta^2) \qquad \text{and} \qquad \gamma_1 = \frac{\sigma^2_\varepsilon}{6} = \sigma^2_\eta\theta. \tag{68}$$

The equations must be solved for θ and σ^2_η. There are two solutions for θ, and we should take the one which fulfils the condition of invertibility: $\theta = 2 - \sqrt{3}$. (See Pollock 1999.)

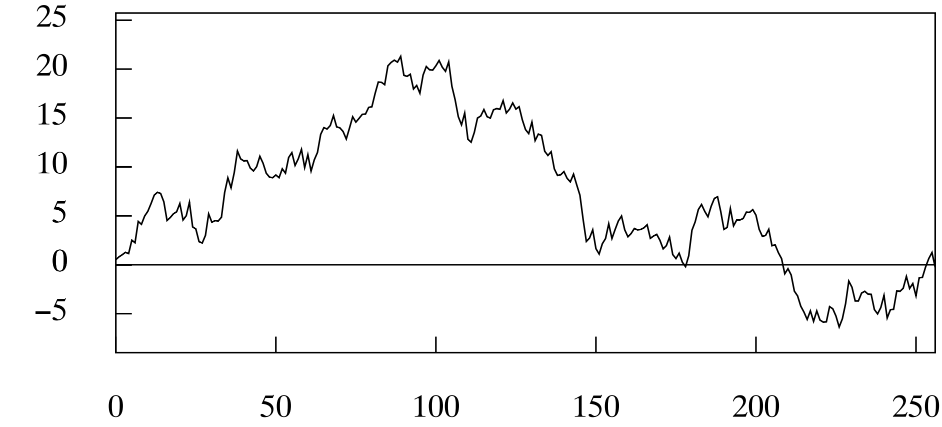

Figure 9. The graph of 256 observations on a simulated series generated by a random walk.

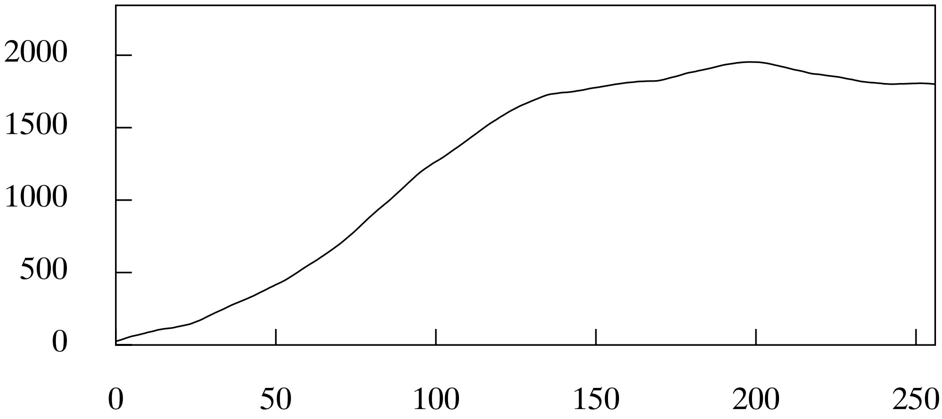

Figure 10. The graph of 256 observations on a simulated series generated by an IMA(2, 1) process that correspond to the sampled version of an integrated Wiener process.

When white-noise errors of observation are superimposed upon values sampled from an integrated Wiener process at regular intervals, the resulting sequence can be described by a doubly-integrated second-order moving-average process in discrete time, which is an IMA(2, 2) process. Such a model provides the basis for the cubic smoothing spline of Reinsch (1976), which can be used to extract an estimate of the trajectory of the underlying integrated Wiener process from the noisy data. The statistical interpretation of the smoothing spline is due to Wahba (1978).

The smoothing spline interpolates cubic polynomial segments between nodes that are derived by smoothing a sequence of sampled data points. The segments are joined in such a way as to ensure that the second derivative of the spline function is continuous at the nodes. An account of the algorithm of the smoothing spline and of its derivation from the statistical model has been provided by Pollock (1999). It is shown that the means by which the nodes are obtained from the data amount to a so-called discrete-time Wiener–Kolmogorov (W–K) filter.

The Wiener–Kolmogorov principle can also be used to derive the so-called Hodrick–Prescott (H–P) filter, which is widely employed in macroeconomic analysis—See Hodrick and Prescott (1980, 1997). The filter, which is presented in section 6.2, is derived from the assumption that the process that generates the trend is a doubly-integrated discrete-time white noise. When white-noise errors are added to the sampled values of the process, the observations are once more described by an IMA(2, 2) model, and the nodes that are generated by the W–K trend-extraction filter are analogous to those of the smoothing spline.

The trend that is generated by the smoothing spline is an aesthetically pleasing curve, of which the smoothness belies the disjunct nature of the stochastic forcing function. That nature is more clearly revealed in the case of a model that postulates a trend that is generated by an ordinary Wiener process, as opposed to an integrated process. The discrete-time observations, which are affected by white-noise errors, are modelled by an IMA(1, 1) process, which also corresponds to the local level model that has been advocated by Harvey (1985, 1989) amongst others. The function that provides statistical estimates of the trend at the nodes and at the points between them has jointed linear segments.

It should be recognised that, if the forcing function were assumed to be bounded in frequency, then the interpolating function would be a smooth one, generated by a Fourier interpolatison, that would have no discontinuities at the nodes.

In section 9, we shall return to the question of how best to specify the continuous-time forcing function. In the next section, we shall deal exclusively with discrete-time models, and we shall examine various ways of decomposing into its component parts a model of an aggregate process that combines the trend and the cycles.

Example. A Wiener process, which is everywhere continuous but nowhere differentiable, can be represented graphically only via its sampled ordinates. If the sampling is sufficiently rapid to give a separation between adjacent points that is below the limits of visual acuity, then the sampled process, which constitutes a discrete-time random walk, will give the same visual impression as the underlying Wiener process. This is the intended effect of Figure 9.

Figure 10 depicts the trajectory of the IMA(2, 1) process that represents the sampled version of an integrated Wiener process. This is a much smoother trajectory than that of the random walk. The extra smoothness can be attributed to the effect of the summation operator, of which the squared gain has been depicted in Figure 8. The operator amplifies the sinusoidal elements in the lower part of the frequency range and it attenuates those in the upper part.

6. Decomposition of discrete-time ARIMA processes

An autoregressive moving-average (ARMA) model can be represented by the equation

$$\sum_{i=0}^{p} \phi_i y_{t-i} = \sum_{i=0}^{q} \theta_i \varepsilon_{t-i} \quad \text{with} \quad \phi_0 = \theta_0 = 1, \tag{69}$$

where t has whatever range is appropriate to the analysis. To exploit the algebra of polynomial operators, the equation can be embedded within the system

$$\phi(z)y(z) = \theta(z)\varepsilon(z), \tag{70}$$

where $\varepsilon(z) = z^t\{\varepsilon_t + \varepsilon_{t-1}z^{-1} + \cdots\}$ is a z-transform of the infinite white-noise forcing function or disturbance sequence $\{\varepsilon_{t-i}; i = 0, 1, \ldots\}$ and where $y(z)$ is the z-transform of the corresponding data sequence. The embedded equation will be associated with z^t.

The polynomials $\theta(z)$ and $\phi(z)$ must have all their roots outside the unit circle to make their inverses $\theta^{-1}(z)$ and $\phi(z)^{-1}$ amenable to power series expansions when $|z| \geq 1$. Then, it is possible to represent the system of (70) by the equation $y(z) = \phi^{-1}(z)\theta(z)\varepsilon(z)$.

An autoregressive integrated moving-average (ARIMA) process represents the accumulation of the output of an ARMA process. On defining the (backwards) difference operator $\nabla(z) = 1 - z$, the dth-order model can be represented by

$$\nabla^d(z)\alpha(z)y(z) = \theta(z)\varepsilon(z). \tag{71}$$

The inverse of the difference operator is the summation operator $\nabla^{-1}(z) = \{1 + z + z^2 + \cdots\}$, and this might be used in representing the system of (71), alternatively, by the equation $y(z) = \nabla^{-d}(z)\alpha^{-1}(z)\theta(z)\varepsilon(z)$.

The difficulty here is that, if it is formed from an infinite number of independently and identically distributed random variables, the disturbance sequence cannot have a finite sum. For this reason, it appears that the algebra of polynomial operators cannot be applied to the analysis of nonstationary processes.

The usual recourse in the face of this problem is scrupulously to avoid the use of the cumulation operator $\nabla^{-1}(z)$ and to represent the integrated system only in the form of (71). This is not a wholly adequate solution to the problem since, to exploit the algebra of the operators, it is necessary to define the inverses of all of the polynomial operators. An alternative solution is to constrain the disturbance sequence to be absolutely summable, which appears to negate the assumption that it is generated by a stationary stochastic process.

The proper recourse is to replace the process of indefinite summation by a definite summation that depends upon supplying the system with initial conditions at some adjacent points in time. To show what this entails, we may consider the system of equations that is derived from (69) by setting $t = 0, 1, \ldots, T-1$. The set of T equations can be arrayed in a matrix format as follows:

$$\begin{bmatrix} y_0 & y_{-1} & \cdots & y_{-p} \\ y_1 & y_0 & \cdots & y_{1-p} \\ \vdots & \vdots & \ddots & \vdots \\ y_p & y_{p-1} & \cdots & y_0 \\ \vdots & \vdots & & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_{T-p-1} \end{bmatrix} \begin{bmatrix} 1 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \varepsilon_0 & \varepsilon_{-1} & \cdots & \varepsilon_{-q} \\ \varepsilon_1 & \varepsilon_0 & \cdots & \varepsilon_{1-q} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_q & \varepsilon_{q-1} & \cdots & \varepsilon_0 \\ \vdots & \vdots & & \vdots \\ \varepsilon_{T-1} & \varepsilon_{T-2} & \cdots & \varepsilon_{T-q-1} \end{bmatrix} \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_q \end{bmatrix}. \tag{72}$$

Apart from the elements $y_0, y_1, \ldots, y_{T-1}$ and $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{T-1}$, which fall within the indicated period, these equations comprise the values $y_{-p}, \ldots, y_{-1}$ and

$\varepsilon_{-q},\ldots,\varepsilon_{-1}$, which are to be found in the top-right corners of the matrices, and which constitute the initial conditions at the start-up time of $t = 0$.

Each of the elements within this display can be associated with the power of z that is indicated by the value of its subscripted index. In that case, the system can be represented by equation (70) with the constituent polynomials defined as follows:

$$\begin{aligned} y(z) &= y_{-p}z^{-p} + \cdots + y_0 + y_1 z + \cdots + y_{T-1}z^{T-1}, \\ \varepsilon(z) &= \varepsilon_{-q}z^{-q} + \cdots + \varepsilon_0 + \varepsilon_1 z + \cdots + \varepsilon_{T-1}z^{T-1}, \\ \phi(z) &= 1 + \phi_1 z + \cdots + \phi_p z^p \quad \text{and} \\ \theta(z) &= 1 + \theta_1 z + \cdots + \theta_q z^q. \end{aligned} \tag{73}$$

This scheme applies regardless of the values of the roots of the polynomial operators $\phi(z)$ and $\theta(z)$. Therefore, it can accommodate the case where $\phi(z) = \nabla^d(z)\alpha(z)$, which is that of equation (71). One of the virtues of this notation is that it is not burdened by an explicit representation of the initial conditions. At a later stage, in section 7, we shall need to represent the initial conditions explicitly.

A trend has only a tenuous existence within the context of a univariate ARIMA model of the sort represented by equation (71). In such a model, it amounts to nothing more that the accumulation of the fluctuations that are created by applying a filter $\theta(z)/\alpha(z)$ to a white-noise sequence $\varepsilon(t)$ of independently and identically distributed random variables.

If the trend and the transitory motions that accompany it are due to the same motive force, which is the white-noise process, then it is difficult to draw a distinction between them. However, a distinction can be made by attributing the trend to the unit roots within $\nabla^d(z) = (1-z)^d$ and by attributing the transitory motions to the stable roots of the autoregressive operator $\alpha(z)$. This is what the decomposition of Beveridge and Nelson (1981) achieves.

Faced with the insistence that the trend and the fluctuations are due to separate sources, an obvious recourse is to attribute separate and independent ARIMA models to each of them. In that case, the aggregate data are also described by a univariate ARIMA model. Provided that their models have distinct parameters, Wiener–Kolmogorov (W–K) filters may be used tentatively to extract the independent components from the data.

The assumption that the components originate from transformations of white-noise sequences implies that their spectra extend over the entire frequency range of $[0, \pi]$. This means that they are bound to overlap substantially. In practice, the spectral structures of the components are often confined to frequency bands that are separated by wide spectral dead spaces. In that case, the separation of one component from another can be achieved in a more decisive manner than the W–K filters will usually allow.

6.1 The Beveridge–Nelson decomposition

The Beveridge–Nelson decomposition relates to an ARIMA model with a first-order integration and with stochastic drift. This can be represented in z-

transform notation by

$$y(z) = \frac{\mu(z)}{\nabla(z)} + \frac{\theta(z)}{\alpha(z)\nabla(z)}\varepsilon(z). \tag{74}$$

If the system has a start-up at $t = 0$, then $\mu(z)$, which represents the drift, is the z-transform of a sequence that is constant over the integers $0, 1, \ldots, t$ and zero-valued for $t < 0$. The operator associated with $\varepsilon(z)$ has the following partial-fraction decomposition:

$$\frac{\theta(z)}{\alpha(z)\nabla(z)} = \frac{\rho(z)}{\alpha(z)} + \frac{\delta}{\nabla(z)}. \tag{75}$$

Multiplying both sides by $\nabla(z) = 1 - z$ and setting $z = 1$ gives $\delta = \theta(1)/\alpha(1)$, where the numerator and the denominator are just the sums of the polynomial coefficients. Substituting the result into equation (74) creates an additive decomposition of the form $y(z) = \tau(z) + \zeta(z)$, wherein

$$\tau(z) = \frac{1}{\nabla(z)}\left\{\mu(z) + \delta\varepsilon(z)\right\}, \tag{76}$$

$$\zeta(z) = \frac{\rho(z)}{\alpha(z)}\varepsilon(z) \tag{77}$$

are respectively the trend component and the transitory component. This is the so-called Beveridge–Nelson decomposition.

The trend component of the Beveridge–Nelson decomposition is a first-order random walk with drift, whereas the transient component is an ARMA process. The distinguishing feature of the decomposition is that both components have the same forcing function. It is easy to see that

$$\tau(z) = \frac{\theta(1)}{\alpha(1)}\frac{\alpha(z)}{\theta(z)}y(z), \tag{78}$$

which is to say that the estimate of the trend is derived by applying an ordinary linear filter to the data sequence. The effect of the filter is to eliminate the ARMA factor from the data so as to deliver a pure random walk.

A common objection to the Beveridge–Nelson decomposition is that the resulting trend is liable to be too rough. This is a consequence of the fact that a random walk that is an accumulation of independently and identically distributed random variables comprises elements at all frequencies up to the limiting Nyquist frequency of π radians per sample period. Also, the decomposition makes no provision for the presence of seasonal fluctuations in the data. A more elaborate model can be proposed with the aim of overcoming these objections

Consider the multiplicative seasonal ARIMA model of Box and Jenkins (1976), which can be represented by the equation

$$\nabla^d(z)\nabla_s^D(z)y(z) = \mu(z) + \frac{\theta(z)\Theta(z^s)}{\alpha(z)A(z^s)}\varepsilon(z). \tag{79}$$

Here, $\alpha(z)$ and $\theta(z)$ are the autoregressive and moving-average polynomials that have appeared in equation (74), whereas $A(z)$ and $\Theta(z)$ are seasonal operators. Whereas $\nabla(z)$ continues to represent the ordinary difference operator, there is now a seasonal difference operator $\nabla_s(z) = 1 - z^s = (1-z)S(z)$, which forms the differences between the data from the same season (or month) of two successive years. The factors of this operator are the ordinary difference operator and a seasonal summation operator $S(z) = 1 + z + z^2 + \cdots + z^{s-1}$. A decomposition can now be found of the form $y(z) = \tau(z) + \sigma(z) + \zeta(z)$, where

$$\tau(z) = \frac{1}{\nabla^{d+D}}\{\mu(z) + \alpha(z)\varepsilon(z)\}, \tag{80}$$

$$\sigma(z) = \frac{\beta(z)}{S^D(z)}\varepsilon(z), \tag{81}$$

$$\zeta(z) = \frac{\gamma(z)}{\alpha(z)A(z^s)}\varepsilon(z), \tag{82}$$

are, respectively, the trend, the seasonal component and the transient component. If the degree $d+D$ of the (ordinary) difference operator exceeds unity, then the trend is liable to be smoother than one generated by a first-order random walk. Also, the effect of $\alpha(z)$ might be further to attenuate the high-frequency elements of the forcing function, thereby enhancing the smoothness of the trend.

To enhance the smoothness of the trend and of the seasonal component yet further, an irregular component could be incorporated in the decomposition. The irregular elements could be extracted from the trend and the seasonal component and assigned to this additional term, which could be regarded as statistically independent of the primary forcing function $\varepsilon(t)$. However, from this point of view, it is natural to consider a model in which each of the components is driven by a statistically independent forcing function. Such a model is the basis of the Wiener–Kolmogorov methodology for signal extraction.

6.2 Wiener–Kolmogorov filtering

The modern theory of statistical signal extraction was formulated independently by Wiener (1941) and Kolmogorov (1941), who arrived at the same results in different ways. Whereas Kolmogorov took a time-domain approach to the problem, Wiener worked primarily in the frequency domain. However, the unification of the two approaches was soon achieved, and a modern account of the theory, which encompasses both, has been provided by Whittle (1983).

The purpose of a Wiener–Kolmogorov (W–K) filter is to extract an estimate of a signal sequence $\xi(t)$ from an observable data sequence

$$y(t) = \xi(t) + \eta(t), \tag{83}$$

which is afflicted by the noise $\eta(t)$. According to the classical assumptions, which we shall later amend, the signal and the noise are generated by zero-mean stationary stochastic processes that are mutually independent. Also, the assumption is made that the data constitute a doubly-infinite sequence. It follows that the autocovariance generating function of the data is the sum of the autocovariance generating functions of its two components. Thus

$$\gamma^{yy}(z) = \gamma^{\xi\xi}(z) + \gamma^{\eta\eta}(z) \quad \text{and} \quad \gamma^{\xi\xi}(z) = \gamma^{y\xi}(z). \tag{84}$$

These functions are amenable to the so-called Cramér–Wold factorisation, and they may be written as

$$\gamma^{yy}(z) = \phi(z^{-1})\phi(z), \quad \gamma^{\xi\xi}(z) = \theta(z^{-1})\theta(z), \quad \gamma^{\eta\eta}(z) = \theta_\eta(z^{-1})\theta_\eta(z). \tag{85}$$

The estimate x_t of the signal element ξ_t is a linear combination of the elements of the data sequence:

$$x_t = \sum_j \beta_j y_{t-j}. \tag{86}$$

The principle of minimum-mean-square-error estimation indicates that the estimation errors must be statistically uncorrelated with the elements of the information set. Thus, the following condition applies for all k:

$$\begin{aligned} 0 &= E\Big\{y_{t-k}(\xi_t - x_t)\Big\} \\ &= E(y_{t-k}\xi_t) - \sum_j \beta_j E(y_{t-k}y_{t-j}) \\ &= \gamma_k^{y\xi} - \sum_j \beta_j \gamma_{k-j}^{yy}. \end{aligned} \tag{87}$$

The equation may be expressed, in terms of the z-transforms, as

$$\gamma^{y\xi}(z) = \beta(z)\gamma^{yy}(z), \tag{88}$$

It follows that

$$\begin{aligned} \beta(z) &= \frac{\gamma^{y\xi}(z)}{\gamma^{yy}(z)} \\ &= \frac{\gamma^{\xi\xi}(z)}{\gamma^{\xi\xi}(z) + \gamma^{\eta\eta}(z)} = \frac{\theta(z^{-1})\theta(z)}{\rho(z^{-1})\rho(z)}. \end{aligned} \tag{89}$$

Now, by setting $z = \exp\{i\omega\}$, one can derive the frequency-response function of the filter that is used in estimating the signal $\xi(t)$. The effect of the filter is to multiply each of the frequency elements of $y(t)$ by the fraction of its variance that is attributable to the signal. The same principle applies to the estimation of the residual component. This is obtained using the complementary filter

$$\beta^c(z) = 1 - \beta(z) = \frac{\gamma^{\eta\eta}(z)}{\gamma^{\xi\xi}(z) + \gamma^{\eta\eta}(z)}. \tag{90}$$

The estimated signal component may be obtained by filtering the data in two passes according to the following equations:

$$\phi(z)q(z) = \theta(z)y(z), \qquad \phi(z^{-1})x(z^{-1}) = \theta(z^{-1})q(z^{-1}). \tag{91}$$

The first equation relates to a process that runs forwards in time to generate the elements of an intermediate sequence, represented by the coefficients of $q(z)$. The second equation represents a process that runs backwards to deliver the estimates of the signal, represented by the coefficients of $x(z)$.

The Wiener–Kolmogorov methodology can be applied to non stationary data with minor adaptations. A model of the processes underlying the data can be adopted that has the form of

$$\begin{aligned}\nabla^d(z)y(z) &= \nabla^d(z)\{\xi(z)+\eta(z)\} = \delta(z)+\kappa(z)\\ &= (1+z)^n\zeta(z)+(1-z)^m\varepsilon(z),\end{aligned} \tag{92}$$

where $\zeta(z)$ and $\varepsilon(z)$ are the z-transforms of two independent white-noise sequences $\zeta(t)$ and $\varepsilon(t)$. The condition $m \geq d$ is necessary to ensure the stationarity of $\eta(t)$, which is obtained from $\varepsilon(t)$ by differencing $m-d$ times. Then, the filter that is applied to $y(t)$ to estimate $\xi(t)$, which is the d-fold integral of $\delta(t)$, takes the form of

$$\beta(z) = \frac{\sigma_\zeta^2(1+z^{-1})^n(1+z)^n}{\sigma_\zeta^2(1+z^{-1})^n(1+z)^n + \sigma_\varepsilon^2(1-z^{-1})^m(1-z)^m}, \tag{93}$$

regardless of the degree d of differencing that would be necessary to reduce $y(t)$ to stationarity.

Two special cases are of interest. By setting $d=m=2$ and $n=0$ in (92), a model is obtained of a second-order random walk $\xi(t)$ affected by white-noise errors of observation $\eta(t)=\varepsilon(t)$. The resulting lowpass W–K filter, in the form of

$$\beta(z) = \frac{1}{1+\lambda(1-z^{-1})^2(1-z)^2} \quad \text{with} \quad \lambda = \frac{\sigma_\eta^2}{\sigma_\delta^2}, \tag{94}$$

is the Hodrick–Prescott (H–P) filter. The complementary highpass filter, which generates the residue, is

$$\beta^c(z) = \frac{(1-z^{-1})^2(1-z)^2}{\lambda^{-1}+(1-z^{-1})^2(1-z)^2}. \tag{95}$$

Here, λ, which is described as the smoothing parameter, is the single adjustable parameter of the filter.

By setting $m=n$, a filter for estimating $\xi(t)$ is obtained that takes the form of

$$\begin{aligned}\beta(z) &= \frac{\sigma_\zeta^2(1+z^{-1})^n(1+z)^n}{\sigma_\zeta^2(1+z^{-1})^n(1+z)^n + \sigma_\varepsilon^2(1-z^{-1})^n(1-z)^n}\\ &= \frac{1}{1+\lambda\left(\mathrm{i}\dfrac{1-z}{1+z}\right)^{2n}} \quad \text{with} \quad \lambda = \frac{\sigma_\varepsilon^2}{\sigma_\zeta^2}.\end{aligned} \tag{96}$$

This is the formula for the Butterworth lowpass digital filter. The filter has two adjustable parameters, and, therefore, it is a more flexible device than the H–P filter. First, there is the parameter λ. This can be expressed as

$$\lambda = \{1/\tan(\omega_d)\}^{2n}, \tag{97}$$

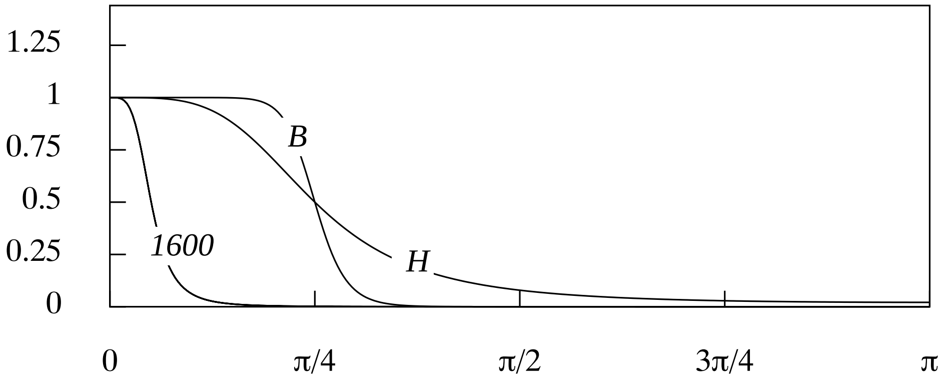

Figure 11. The gain of the Hodrick–Prescott H and of the Butterworth filter B with nominal cut-off points at $\pi/4$ radians, together with the gain of a Hodrick–Prescott filter with a smoothing parameter of 1600.

where ω_d is the nominal cut-off point of the filter, which is the mid point in the transition of the filter's frequency response from its pass band to its stop band. The second of the adjustable parameters is n, which denotes the order of the filter. As n increases, the transition between the pass band and the stop band becomes more abrupt.

These filters can be applied to the nonstationary data sequence $y(t)$ in the manner indicated by equation (91), provided that the appropriate initial conditions are supplied with which to start the recursions. However, by concentrating on the estimation of the residual sequence $\eta(t)$, which corresponds to a stationary process, it is possible to avoid the need for nonzero initial conditions. Then, the estimate of $\eta(t)$ can be subtracted from $y(t)$ to obtain the estimate of $\xi(t)$.

The Hodrick–Prescott filter has many antecedents. Its invention cannot reasonably be attributed to Hodrick and Prescott (1980, 1997), who cited Whittaker (1923) as one of their sources. Leser (1961) also provided a complete derivation of the filter at an earlier date. The Butterworth filter is a commonplace of electrical engineering. The digital version of the filter has been described in an econometric context by Pollock (2000) and by Gómez (2001). It has been applied to climatological data by Harvey and Mills (2003).

Example. Figure 11 shows the gain functions of the three filters overlaid on the same diagram. The lowpass Hodrick–Prescott filter with a smoothing parameter of $\lambda = 1600$ is commonly recommended for estimating the trend in quarterly economic data. The corresponding gain function is marked in the diagram by the number 1600.

An alternative to specifying the smoothing parameter directly is to specify the frequency value ω_d for which the gain is $\beta(\omega_d) = 0.5$. For the H–P filter, the correspondence between ω_d and λ is as follows:

$$\lambda = \frac{1}{4\{1 - \cos(\omega_d)\}^2} \qquad \text{and} \qquad \omega_d = \cos^{-1}(1 - 1\sqrt{4\lambda}). \tag{98}$$

The function labelled B is the gain of the filter for which $\omega_d = \pi/4$.

The frequency ω_d corresponds to the mid-point in the transition between the pass band and the stop band of the filter. This might be described as the nominal cut-off frequency, but, in the case of the H–P filter, this is a misnomer, on account of the very gradual transition of the gain. The Butterworth filter is capable of a much more rapid transition. The curve labelled B corresponds to the gain of a Butterworth filter with $n = 6$ and $\omega_d = \pi/4$.

6.3 Structural ARIMA models

The Hodrick–Prescott filter and the Butterworth filter are appropriate to the task of extracting the trend or the trend/cycle component from a data sequence without regard to the structure of the residual component. More elaborate filters are available that also take account of a seasonal component.

Consider, therefore, a seasonal autoregressive moving-average model of the form

$$y(z) = \frac{\theta(z)}{\phi(z)}\varepsilon(z) = \frac{\theta(z)}{\phi_S(z)\phi_T(z)}\varepsilon(z), \tag{99}$$

where $\phi_S(z)$ contains the seasonal autoregressive factors and $\phi_T(z)$ contains the non-seasonal factors.

The denominator contains both an ordinary differencing operator $\nabla^d(z)$ and a seasonal differencing operator $\nabla_s^D(z) = \nabla^D(z)S^D(z)$. The operator $\nabla_s(z) = 1 - z^s = (1 - z)S(z)$ forms the differences between the data from the same season (or month) of two successive years. Its factors are the ordinary difference operator and a seasonal summation operator $S(z) = 1 + z + z^2 + \cdots + z^{s-1}$. The factorisation of the seasonal operator implies that the overall degree of differencing within the ARIMA model is $d + D$. The factor $\nabla^{d+D}(z)$ is assigned to $\phi_T(z)$, whereas $S^D(z)$ belongs to $\phi_S(z)$.

On the assumption that the degree of the moving-average polynomial $\theta(z)$ is at least equal to that of the denominator polynomial $\phi(z)$, there is a partial-fraction decomposition of the autocovariance generating function of the model into three components, which correspond to the trend effect, the seasonal effect and an irregular influence. Thus

$$\frac{\theta(z^{-1})\theta(z)}{\phi_S(z^{-1})\phi_T(z^{-1})\phi_T(z)\phi_S(z)} = \frac{Q_T(z)}{\phi_T(z^{-1})\phi_T(z)} + \frac{Q_S(z)}{\phi_S(z^{-1})\phi_S(z)} + R(z). \tag{100}$$

Here, the first two components on the RHS represent proper rational fractions, whereas the final component is an ordinary polynomial. If the degree of the moving-average polynomial is less than that of the denominator polynomial, then the irregular component is missing from the decomposition in the first instance.

To obtain the spectral density function of $y(t)$, we set $z = e^{-\mathrm{i}\omega}$, where $\omega \in [0, \pi]$. (This function is more properly described as a pseudo-spectrum in view of the singularities occasioned by the unit roots in the denominators of the first two components.) The spectral decomposition corresponding to equation (100) can be written as

$$f(\omega) = f(\omega)_T + f(\omega)_S + f(\omega)_R, \tag{101}$$

where $f(\omega) = \theta(e^{i\omega})\theta(e^{-i\omega})/\{\phi(e^{i\omega})\phi(e^{-i\omega})\}$.

Let $\nu_T = \min\{f(\omega)_T\}$ and $\nu_S = \min\{f(\omega)_S\}$. These correspond to the elements of white noise embedded in $f(\omega)_T$ and $f(\omega)_S$. The principle of canonical decomposition is that the white-noise elements should be reassigned to the residual component. On defining

$$\begin{aligned}\gamma_T(z)\gamma_T(z^{-1}) &= Q_T(z) - \nu_T\phi_T(z)\phi_T(z^{-1}),\\ \gamma_S(z)\gamma_S(z^{-1}) &= Q_S(z) - \nu_S\phi_S(z)\phi_S(z^{-1}),\\ \text{and}\quad \rho(z)\rho(z^{-1}) &= R(z) + \nu_T + \nu_S,\end{aligned}\tag{102}$$

the canonical decomposition of the generating function can be represented by

$$\frac{\theta(z)\theta(z^{-1})}{\phi(z)\phi(z^{-1})} = \frac{\gamma_T(z)\gamma_T(z^{-1})}{\phi_T(z)\phi_T(z^{-1})} + \frac{\gamma_S(z)\gamma_S(z^{-1})}{\phi_S(z)\phi_S(z^{-1})} + \rho(z)\rho(z^{-1}).\tag{103}$$

There are now two improper rational functions on the RHS, which have equal degrees in their numerators and denominators.

According to Wiener–Kolmogorov theory, the optimal signal-extraction filter for the trend component is

$$\begin{aligned}\beta_T(z) &= \frac{\gamma_T(z)\gamma_T(z^{-1})}{\phi_T(z)\phi_T z^{-1})} \times \frac{\phi_S(z)\phi_T(z)\phi_T(z^{-1})\phi_S(z^{-1})}{\theta(z)\theta(z^{-1})}\\ &= \frac{\gamma_T(z)\gamma_T(z^{-1})\phi_S(z)\phi_S(z^{-1})}{\theta(z)\theta(z^{-1})} = \frac{C_T(z)}{\theta(z)\theta(z^{-1})}.\end{aligned}\tag{104}$$

This has the form of the ratio of the autocovariance generating function of the trend component to the autocovariance generating function of the process $y(t)$. This formulation presupposes a doubly-infinite data sequence, so it must be translated into a form that can be implemented with finite sequences.

The approach to the estimation of unobserved components that adopts the principle of canonical decompositions has been advocated by Hillmer and Tiao (1982) and by Maravall and Pierce (1987). It has been implemented in the TRAMO–SEATS program of Gómes and Maravall (1996) and of Caporello and Maravall (2004), which builds upon the work of Burman (1980). A comparative analysis of the STAMP and TRAMO–SEATS programs has been provided by Pollock (2002b)

6.4 The state space form of the structural model

In the foregoing approach to modelling the components of a structural time series model, an aggregate univariate process is first estimated and then decomposed into its components. An alternative approach is to model the individual components from the start as separate entities, which are described by independent linear stochastic models.

Provision can be made for a cyclical component which is distinct from the trend component, but, if this is omitted, then the disaggregated model commonly takes the form of $y(z) = \tau(z) + \sigma(z) + \eta(z)$, where

$$\tau(z) = \frac{(1 + \alpha z)}{\nabla^2(z)}\zeta(z), \tag{105}$$

$$\sigma(z) = \frac{1}{S(z)}\omega(z). \tag{106}$$

Then, $\tau(t)$ is the trend, $\sigma(t)$ is the seasonal component and $\eta(t)$ is the irregular noise. Here, there are three independent white-noise processes driving the model, which are $\zeta(t)$, $\omega(t)$ and $\eta(t)$. The model has been described by Harvey (1989) as the basic structural model. A reason for omitting the cyclical or business-cycle component from this model is the difficulty in separating it from the trend component.

The trend process is usually depicted as the product of two processes that constitute the so-called local linear model, which has already been described in section 5.1:

$$\tau(t) = \tau(t-1) + \beta(t) + \nu(t), \tag{107}$$

$$\beta(t) = \beta(t-1) + \varepsilon(t). \tag{108}$$

The first of these describes the level of the trend process and the second describes its slope.

A more elaborate seasonal model is available that generates more regular cycles. A moving-average operator $M(z)$ can be included in the numerator of the expression on the RHS of (106) to give $\sigma(z) = \{M(z)/S(z)\}\omega(z)$. The autoregressive operator may be factorised as $S(z) = \prod_{j=1}^{s-1}(1 - e^{2\pi j/s})$, where s is the number of observations per annum. The complementary moving-average operator will have the form of $M(z) = \prod_{j=1}^{s-1}(1 - \rho e^{2\pi j/s})$, where $\rho < 1$ is close to unity. The zeros of the moving-average operator will serve largely to negate the effects of the poles of the autoregressive operator, except at the seasonal frequencies, where prominent spectral spikes will be found.

The basic structural model, without the elaboration of a seasonal moving-average component, can be represented in a state-space form that comprises a transition equation, which describes a first-order vector autoregressive process, and an accompanying measurement equation. For notational convenience, let $s = 4$, which corresponds to the case of quarterly observations on annual data. Then, the transition equation, which gathers together equations (106), (107) and (108), is

$$\begin{bmatrix} \tau(t) \\ \beta(t) \\ \sigma(t) \\ \sigma(t-1) \\ \sigma(t-2) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tau(t-1) \\ \beta(t-1) \\ \sigma(t-1) \\ \sigma(t-2) \\ \sigma(t-3) \end{bmatrix} + \begin{bmatrix} \nu(t) \\ \varepsilon(t) \\ \omega(t) \\ 0 \\ 0 \end{bmatrix}. \tag{109}$$

This incorporates the transition equation of the non-seasonal local linear model that has been given by (64). The observation equation, which combines the

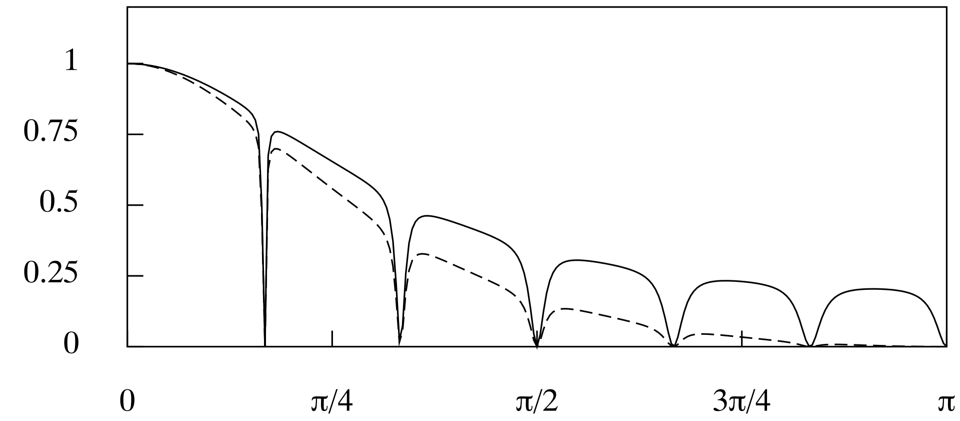

Figure 12. The gain function of the trend-extraction filter obtained from the STAMP program (solid line) together with that of the canonical trend-extraction filter (broken line).

current values of the components, is

$$y(t) = [1 \quad 0 \quad 1 \quad 0 \quad 0] \begin{bmatrix} \tau(t) \\ \beta(t) \\ \sigma(t) \\ \sigma(t-1) \\ \sigma(t-2) \end{bmatrix} + \eta(t). \tag{110}$$

The state-space model is amenable to the Kalman filter and the associated smoothing algorithms, which can be used in estimating the parameters of the model and in extracting estimates of the so-called unobserved components $\tau(t)$, $\sigma(t)$ and $\varepsilon(t)$. These algorithms have been described by Pollock (2003a).

Disaggregated structural time-series models have been treated at length in the book by Harvey (1989). The methodology has been implemented in the STAMP program, which is described by Koopman, Harvey, Doornick and Shephard (2007). A similar approach has been pursued in a program within the Captain MATLAB Toolbox, which has been described by Pedregal, Taylor and Young (2004).

Example. Figure 12 show the gain of the trend extraction filter that is associated with a disaggregated structural model that has been applied to the monthly airline passenger data of Box and Jenkins (1976).

The solid line represents the gain of the ordinary filter and the broken line represents the gain of the filter that is obtained when the principle of canonical decomposition is applied to the components of the model. In that case, the white noise that is contained in the components is removed and reassigned to the residual component.

The indentations in the gain function at the seasonal frequencies $\pi j/6; j = 1,\ldots,6$ are due to the zeros of the filter that are to be found on the circumference of the unit circle and which are effective in removing the seasonal fluctuations from the trend.

Disregarding these indentations, the gain of the filters is reduced only gradually as frequency increases. In particular, the ordinary unadjusted filter is liable to transmit a higher proportion of the high-frequency noise of the data. Howerver, given that such high-frequency noise is largely absent from the airline passenger data, it transpires that the effect upon the estimated trend of adopting the principle of canonical decomposition is a minor one.

7. Finite-sample signal extraction

The classical theory of linear filtering relies heavily upon the simplifications that are afforded by the assumption that the data constitute a doubly infinite sequence. The assumption is an acceptable one in the case of finite impulse response (FIR) filters that can be realised via low-order moving-average operators. When such a filter has only a short span, it matters little which assumptions are made about the length of the data sequence. Only at the ends of the data sequence are there liable to be problems.

The assumption of a double-infinite data sequence, also sustains the theory of time-invariant infinite impulse response (IIR) rational filters, such as the Butterworth and Hodrick–Prescott filters of section 6.2, which correspond to moving averages of infinite order. These are not so easily applied to short sequences. Nevertheless, if the data sequence is sufficiently lengthy to allow the transient effects of the arbitrary start-up values to disappear, then such filters can be implemented successfully via bi-directional feedback procedures which comprise only and handful of recent data values. (In effect the start-up values purport to summarise the history of the infinite data sequence, in so far as it affects the IIR filter.)

In econometric applications, attention is often focussed upon the most recent observations at the upper end of a short data sequence. In such cases, a theory of filtering is called for that fully recognises the finite nature of the data sequence. Also, in cases where the data are trended, it becomes essential to supply appropriate nonzero initial conditions to the filter; and these should be the products of a finite-sample theory.

The theory that we shall expound here depends upon replacing the symbol z within the various polynomial operators by a matrix lag operator. However, it is immediately apparent that this replacement alone is insufficient for the purpose of creating adequate finite-sample filters.

To demonstrate the effects of the replacement, let $L_T = [e_1, e_2, \ldots, e_{T-1}, 0]$ be the matrix version of the lag operator, which is formed from the identity matrix $I_T = [e_0, e_1, e_2, \ldots, e_{T-1}]$ of order T by deleting the leading column and by appending a column of zeros to the end of the array. Then, the matrix of order T that corresponds to the p-th difference operator $\nabla^p(z) = (1 - z)^p$ is

$$\nabla_T^p = (I - L_T)^p. \tag{111}$$

We may partition this matrix so that $\nabla_T^p = [Q_*, Q]'$, where Q_*' has p rows. If y is a vector of T elements, then

$$\nabla_T^p y = \begin{bmatrix} Q_*' \\ Q' \end{bmatrix} y = \begin{bmatrix} g_* \\ g \end{bmatrix}; \tag{112}$$

and g_* is liable to be discarded, whereas g will be regarded as the vector of the p-th differences of the data.

The inverse matrix is partitioned conformably to give $\nabla_T^{-p} = [S_*, S]$. It follows that

$$[\,S_* \quad S\,]\begin{bmatrix} Q_*' \\ Q' \end{bmatrix} = S_*Q_*' + SQ' = I_T, \tag{113}$$

and that

$$\begin{bmatrix} Q_*' \\ Q' \end{bmatrix}[\,S_* \quad S\,] = \begin{bmatrix} Q_*'S_* & Q_*'S \\ Q'S_* & Q'S \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & I_{T-p} \end{bmatrix}. \tag{114}$$

If g_* is available, then y can be recovered from g via

$$y = S_*g_* + Sg. \tag{115}$$

The lower-triangular Toeplitz matrix $\nabla_T^{-p} = [S_*, S]$ is completely characterised by its leading column. The elements of that column are the ordinates of a polynomial of degree $p-1$, of which the argument is the row index $t = 0, 1, \ldots, T-1$. Moreover, the leading p columns of the matrix ∇_T^{-p}, which constitute the submatrix S_*, provide a basis for all polynomials of degree $p-1$ that are defined on the integer points $t = 0, 1, \ldots, T-1$.

It follows that $S_*g_* = S_*Q_*'y$ contains the ordinates of a polynomial of degree $p-1$, which is interpolated through the first p elements of y, indexed by $t = 0, 1, \ldots, p-1$, and which is extrapolated over the remaining integers $t = p, p+1, \ldots, T-1$.

7.1 Polynomial regression and H–P filtering

A polynomial that is designed to fit the data should take account of all of the observations in y. Imagine, therefore, that $y = \phi + \eta$, where ϕ contains the ordinates of a polynomial of degree $p-1$ and η is a disturbance term with $E(\eta) = 0$ and $D(\eta) = \Sigma$. Then, in forming an estimate $f = S_*r_*$ of ϕ, we should minimise the sum of squares $\eta'\Sigma^{-1}\eta$. Since the polynomial is fully determined by the elements of a starting-value vector r_*, this is a matter of minimising

$$(y-\phi)'\Sigma^{-1}(y-\phi) = (y - S_*r_*)'\Sigma^{-1}(y - S_*r_*) \tag{116}$$

with respect to r_*. The resulting values are

$$r_* = (S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1}y \quad \text{and} \quad \phi = S_*(S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1}y. \tag{117}$$

An alternative representation of the estimated polynomial is available, which avoids the inversion of Σ. This is provided by the identity

$$\begin{aligned} P_* &= S_*(S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1} \\ &= I - \Sigma Q(Q'\Sigma Q)^{-1}Q' = I - P_Q, \end{aligned} \tag{118}$$

which gives two representations of the projection matrix P_*. The equality follows from the fact that, if $\text{Rank}[R, S_*] = T$ and if $S_*'\Sigma^{-1}R = 0$, then

$$S_*(S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1} = I - R(R'\Sigma^{-1}R)^{-1}R'\Sigma^{-1}. \tag{119}$$

Setting $R = \Sigma Q$ gives the result. It follows that the polynomial fitted to the data by generalised least-squares regression can be written as

$$\phi = y - \Sigma Q(Q'\Sigma Q)^{-1}Q'y. \tag{120}$$

A more general method of curve fitting, which embeds polynomial regression as a special case, is one that involves the minimisation of a combination of two sums of squares. Let x denote the vector of fitted values, which is a sequence of the ordinates of points, equally spaced in time, through which a continuous curve might be interpolated. The criterion for finding the vector is to minimise

$$L = (y - x)'\Sigma^{-1}(y - x) + x'Q\Omega^{-1}Q'x. \tag{121}$$

The first term penalises departures of the resulting curve from the data, whereas the second term imposes a penalty for a lack of smoothness in the curve.

The second term comprises $d = Q'x$, which is the vector of pth-order differences of x. The matrix Ω^{-1} serves to generalise the overall measure of the curvature of the function that has the elements of x as its sampled ordinates, and it serves to regulate the penalty, which may vary over the sample.

Differentiating L with respect to x and setting the result to zero, in accordance with the first-order conditions for a minimum, gives

$$\begin{aligned}\Sigma^{-1}(y - x) &= Q\Omega^{-1}Q'x\\ &= Q\Omega^{-1}d.\end{aligned} \tag{122}$$

Multiplying the equation by $Q'\Sigma$ gives $Q'(y-x) = Q'y-d = Q'\Sigma Q\Omega^{-1}d$, whence $\Omega^{-1}d = (\Omega+Q'\Sigma Q)^{-1}Q'y$. Putting this into the equation $x = y-\Sigma Q\Omega^{-1}d$ gives

$$x = y - \Sigma Q(\Omega + Q'\Sigma Q)^{-1}Q'y. \tag{123}$$

By setting $\Omega = \lambda^{-1}I$ and $\Sigma = I$ and letting Q' denote the second-order difference operator, the Hodrick–Prescott filter is obtained in the form of

$$x = y - Q(\lambda^{-1}I + Q'Q)^{-1}Q'y. \tag{124}$$

This form is closely related to that of the infinite-sample filter $\beta(z) = 1 - \beta^c(z)$ which invokes equation (95). In the finite-sample version of the filter, the submatrix Q' of $\nabla_T^2 = (I - L_T)^2$ replaces the difference operator $(1 - z)^2$, and Q replaces $(1 - z^{-1})^2$.

If $\Omega = 0$ in (123), and if Q' is the matrix version of the second-difference operator, then the generalised least-squares interpolator of a linear function is derived, which is subsumed under (120).

7.2 Finite-sample Wiener–Kolmogorov filters

To provide a statistical interpretation of the formula of (123), consider a data sequence $y = \xi + \eta$, where $\xi = \phi + \zeta$ is a trend component, which is the sum of a vector ϕ, containing the ordinates of a polynomial of degree p at most, and of a vector ζ from a stochastic process with p unit roots that is driven by a zero-mean forcing function. The term η stands for a vector sampled from a mean-zero stationary stochastic process which is independent of the process driving ξ such that

$$E(\eta) = 0, \qquad D(\eta) = \Sigma \qquad \text{and} \qquad C(\eta, \xi) = 0. \tag{125}$$

If Q' is the p-th difference operator, then $Q'\phi = \mu\iota$, with $\iota = [1, 1, \ldots, 1]'$, will contain a constant sequence of values, which will be zeros if the degree of ϕ is less than p. Also, $Q'\zeta$ will be a vector sampled from a mean-zero stationary process. Therefore, $\delta = Q'\xi$ is from a stationary process with a constant mean. Thus, there is

$$\begin{aligned} Q'y &= Q'\xi + Q'\eta \\ &= \delta + \kappa = g, \end{aligned} \tag{126}$$

where

$$\begin{aligned} E(\delta) &= \mu\iota, \qquad D(\delta) = \Omega, \\ E(\kappa) &= 0, \qquad D(\kappa) = Q'\Sigma Q. \end{aligned} \tag{127}$$

Now consider the conditional expectation of η given $g = Q'y$, which is also its minimum-mean-square-error estimator on the assumption that the various stochastic processes are normally distributed. This is

$$\begin{aligned} E(\eta|g) &= E(\eta) + C(\eta, g)D^{-1}(g)\{g - E(g)\} \\ &= \Sigma Q(\Omega + Q'\Sigma Q)^{-1}\{Q'y - \mu\iota\}. \end{aligned} \tag{128}$$

Here, if the vector $E(g) = \mu\iota$ is nonzero it will, nevertheless, be virtually nullified by the matrix $\Sigma Q(\Omega + Q'\Sigma Q)^{-1}$, which is a matrix version of a highpass filter. Therefore, it may be deleted from the expressions of (128). Next, since $\xi = y - \eta$, the estimate of the trend is $x = E(\xi|g) = y - E(\eta|g)$, which is exactly equation (123).

The Hodrick–Prescott filter may be derived by specialising the statistical assumptions of (125) and (127). It is assumed that

$$D(\eta) = \Sigma = \sigma_\eta^2 I, \qquad D(\delta) = \Omega = \sigma_\delta^2 I \qquad \text{and} \qquad \lambda = \frac{\sigma_\eta^2}{\sigma_\delta^2}. \tag{129}$$

Putting these details into equation (123) gives equation (124).

It is straightforward to derive the dispersion matrices that are found within the formulae for the finite-sample estimators from the corresponding autocovariance generating functions. Let $\gamma(z) = \{\gamma_0 + \gamma_1(z + z^{-1}) + \gamma_2(z^2 + z^{-2}) + \cdots\}$ denote the autocovariance generating function of a stationary stochastic process. Then, the corresponding dispersion matrix for a sample of T consecutive elements drawn from the process is

$$\Gamma = \gamma_0 I_T + \sum_{\tau=1}^{T-1} \gamma_\tau (L_T^\tau + F_T^\tau), \tag{130}$$

where $F_T = L_T'$ is in place of z^{-1}. Since L_T and F_T are nilpotent of degree T, such that $L_T^q, F_T^q = 0$ when $q \geq T$, the index of summation has an upper limit of $T - 1$.

7.3 The polynomial component

The formula (123) tends to conceal the presence of polynomial components within the sequences that are generated by filtering the nonstationary data. An alternative procedure, which we have already adopted in detrending the logarithmic consumption data of the U.K. in the example following (14), is to extract a polynomial trend from the nonstationary data before applying a filter to the residual sequence, which will have the characteristics of a sequence generated by a stationary process, provided that the polynomial is of a sufficient degree.

Another procedure that can be followed requires the data to be reduced to stationarity by a process of differencing, before it is filtered. The filtered output can be re-inflated thereafter to obtain estimates of the components of the nonstationary process. It transpires that, in the context of Wiener–Kolmogorov filtering, such a procedure produces estimates that are identical to those that are delivered by the finite-sample filter of (123).

To demonstrate this result, we shall assume that, within $y = \xi + \eta$, the vector ξ is generated by a stochastic process with p unit roots driven by a mean-zero white-noise process. The vector η is assumed to be from a stationary process. Therefore, the specifications of (125) and (127) remain, but we may choose to set $E(\delta) = 0$, if only to confirm that the polynomial component will arise just as surely in the absence of stochastic drift.

Let the estimates of ξ, η, $\delta = Q'\xi$ and $\kappa = Q'\eta$ be denoted by x, h, d and k respectively. Then, the Wiener–Kolmogorov, minimum-mean-square-error estimates of the differenced components are

$$E(\delta|g) = d = D(\delta)\{D(\delta) + D(\kappa)\}^{-1}g = \Omega(\Omega + Q'\Sigma Q)^{-1}Q'y, \qquad (131)$$

$$E(\kappa|g) = k = D(\kappa)\{D(\delta) + D(\kappa)\}^{-1}g = Q'\Sigma Q(\Omega + Q'\Sigma Q)^{-1}Q'y. \qquad (132)$$

The estimates of ξ and η may be obtained by integrating, or re-inflating, the components of the differenced data to give

$$x = S_*d_* + Sd \qquad \text{and} \qquad h = S_*k_* + Sk, \qquad (133)$$

where S_*d_* and S_*k_* are vectors of the ordinates of polynomials of degree p. For this representation, the polynomial parameters, in the form of the starting values d_* and h_*, are required.

The initial conditions in d_* should be chosen so as to ensure that the estimated trend is aligned as closely as possible with the data. The criterion is

$$\text{Minimise} \quad (y - S_*d_* - Sd)'\Sigma^{-1}(y - S_*d_* - Sd) \quad \text{with respect to} \quad d_*. \qquad (134)$$

The solution for the starting values is

$$d_* = (S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1}(y - Sd). \qquad (135)$$

The equivalent starting values of k_* are obtained by minimising the (generalised) sum of squares of the fluctuations:

$$\text{Minimise} \quad (S_*k_* + Sk)'\Sigma^{-1}(S_*k_* + Sk) \quad \text{with respect to} \quad k_*. \tag{136}$$

The solution is

$$k_* = -(S_*'\Sigma^{-1}S_*)^{-1}S_*'\Sigma^{-1}Sk. \tag{137}$$

The starting values k_* and d_* can be eliminated from the expressions for x and h in (133), which provide the estimates of the components. Thus, using expression $I - P_* = P_Q$ from (118), we get

$$\begin{aligned} h &= Sk + S_*k_* \\ &= (I - P_*)Sk = P_QSk. \end{aligned} \tag{138}$$

Then, by using the expression for k from (132) together with the identity $Q'S = I_T$, we get

$$h = \Sigma Q(\Omega + Q'\Sigma Q)^{-1}Q'y. \tag{139}$$

This agrees with (128) in the case where $\mu = 0$. The condition that $x + h = y$, which is that the sum of the estimated components equals the data vector, indicates that

$$\begin{aligned} x &= y - h \\ &= y - \Sigma Q(\Omega + Q'\Sigma Q)^{-1}Q'y, \end{aligned} \tag{140}$$

which is equation (123) again.

Observe that the filter matrix $Z_\eta = \Sigma Q(\Omega + Q'\Sigma Q)^{-1}$ of (140), which delivers $h = Z_\eta g$, differs from the matrix $Z_\kappa = Q'Z_\eta$ of (132), which delivers $k = Z_\kappa g$, only in respect of the matrix difference operator Q'. The effect of omitting the operator is to remove the need for re-inflating the filtered components and thus to remove the need for the starting values. These matters have been discussed at greater length by Pollock (2006).

8. The Fourier methods of signal extraction

If the data are generated by a stationary stochastic process, then it may be reasonable to regard them as the product of a circular process, of which the Fourier representation is readily available. There are some advantages in exploiting the Fourier representation by performing the essential filtering operations in the frequency domain—for these are usually aimed at suppressing or attenuating some of the cyclical elements of the data. It is also straightforward to provide a time-domain interpretation of the frequency domain operations, and the possibility exists of performing the equivalent operations in either domain.

The dispersion matrix of a circular stochastic process is obtained from the autocovariance generating function $\gamma(z)$ by replacing the argument z by the circulant matrix $K_T = [e_1, \ldots, e_{T-1}, e_0]$, which is formed from the identity matrix

$I_T = [e_0, e_1, \ldots, e_{T-1}]$ by moving the leading column to the back of the array. In this way, the generating function $\gamma(z)$ gives rise to the matrix

$$\begin{aligned}\Gamma^\circ &= \gamma(K_T)\\ &= \gamma_0 I_T + \sum_{\tau=1}^{\infty}\gamma_\tau(K_T^\tau + K_T^{-\tau})\\ &= \gamma_0 I_T + \sum_{\tau=1}^{T-1}\gamma_\tau^\circ(K_T^\tau + K_T^{-\tau}).\end{aligned} \tag{141}$$

It can be seen from this that the circular autocovariances would be obtained by wrapping the sequence of ordinary autocovariances around a circle of circumference T and adding the overlying values. Thus

$$\gamma_\tau^\circ = \sum_{j=0}^{\infty}\gamma_{jT+\tau}, \quad \text{with} \quad \tau = 0, \ldots, T-1. \tag{142}$$

Given that $\lim(\tau \to \infty)\gamma_\tau = 0$, it follows that $\gamma_\tau^\circ \to \gamma_\tau$ as $T \to \infty$, which is to say that the circular autocovariances converge to the ordinary autocovariances as the circle expands.

The circulant autocovariance matrix is amenable to a spectral factorisation of the form

$$\Omega^\circ = \gamma(K_T) = \bar{U}\gamma(D)U, \tag{143}$$

wherein U and $\bar{U}$ are the unitary matrices defined by (20) and

$$D = \text{diag}(\exp\{\mathrm{i}2\pi j/T\}; j = 0, \ldots, T-1) \tag{144}$$

is a diagonal matrix whose elements are the T roots of unity, which are found on the circumference of the unit circle in the complex plane. Then, $\gamma(D)$ is the diagonal matrix formed by replacing the argument z within $\gamma(z)$ by D.

The jth element of the diagonal matrix $\gamma(D)$ is

$$\gamma(\exp\{\mathrm{i}\omega_j\}) = \gamma_0 + 2\sum_{\tau=1}^{\infty}\gamma_\tau\cos(\omega_j\tau). \tag{145}$$

This represents the cosine Fourier transform of the sequence of the ordinary autocovariances; and it corresponds to an ordinate (scaled by 2π) sampled at the point $\omega_j = 2\pi j/T$, which is a Fourier frequency, from the spectral density function of the linear (i.e. non-circular) stationary stochastic process.

The theory of circulant matrices has been described by Gray (2002) and by Pollock (2002a). Both authors provide abundant additional references.

The method of Wiener–Kolmogorov filtering can also be implemented using the circulant dispersion matrices that are given by

$$\begin{aligned}&\Omega_\delta^\circ = \bar{U}\gamma_\delta(D)U, \quad \Omega_\kappa^\circ = \bar{U}\gamma_\kappa(D)U \quad \text{and}\\ &\Omega^\circ = \Omega_\delta^\circ + \Omega_\kappa^\circ = \bar{U}\{\gamma_\delta(D) + \gamma_\kappa(D)\}U,\end{aligned} \tag{146}$$

wherein the diagonal matrices $\gamma_\delta(D)$ and $\gamma_\kappa(D)$ contain the ordinates of the spectral density functions of the component processes. By replacing the dispersion matrices of (131) and (132) by their circulant counterparts, we derive the following formulae:

$$d = \bar{U}\gamma_\delta(D)\{\gamma_\delta(D) + \gamma_\kappa(D)\}^{-1}Ug = P_\delta g, \tag{147}$$

$$k = \bar{U}\gamma_\kappa(D)\{\gamma_\delta(D) + \gamma_\kappa(D)\}^{-1}Ug = P_\kappa g. \tag{148}$$

We may note that P_δ and P_κ are circulant matrices.

The filtering formulae may be implemented in the following way. First, a Fourier transform is applied to the (differenced) data vector g to give Ug, which resides in the frequency domain. Then, the elements of the transformed vector are multiplied by those of the diagonal weighting matrices $J_\delta = \gamma_\delta(D)\{\gamma_\delta(D) + \gamma_\kappa(D)\}^{-1}$ and $J_\kappa = \gamma_\kappa(D)\{\gamma_\delta(D) + \gamma_\kappa(D)\}^{-1}$. Finally, the products are carried back into the time domain by the inverse Fourier transform, which is represented by the matrix $\bar{U}$. (An efficient implementation of a mixed-radix fast Fourier transform, which is designed to cope with samples of arbitrary sizes, has been provided by Pollock (1999). The usual algorithms demand a sample size of $T = 2^n$.)

An advantage of the Fourier method is that it is possible to effect a total suppression of the elements within the stop band of the desired frequency response. Also, the transition between the pass band and the stop band can be confined to the interval between adjacent Fourier frequencies, which means that it can be perfectly abrupt.

Neither of these features are available to the ordinary finite-sample Wiener–Kolmogorov filters. Nevertheless, it is possible to achieve both of these effects by working in the time domain. This fact is manifest in the formulae of (147) and (148) which entail the equations $d = P_\delta g$ and $k = P_\kappa g$ respectively.

In effect, a pair of wrapped filters can be applied to the data in the time domain via processes of circular convolution. If we can imagine the leading rows of the matrices P_δ and P_κ disposed around a circle of circumference T, then each of the succeeding rows is derived from its predecessor via an anticlockwise rotation through an angle of $2\pi/T$ radians.

Example. It is commonly believed that, in the case of samples of a finite length T, it is impossible to design a filter that will preserve completely all elements within a specified range of frequencies and that will remove all elements outside it. A filter that would achieve such an objective is described as an ideal filter. The ideal lowpass filter with a cut-off frequency of $\omega_d = 2\pi d/T$ has the following frequency response over the interval $[-\pi, \pi]$:

$$\phi(\omega) = \begin{cases} 1, & \text{if } \omega \in [-\omega_d, \omega_d], \\ 0, & \text{otherwise.} \end{cases} \tag{149}$$

The coefficients of the filter are given by the discrete-time sinc function, which

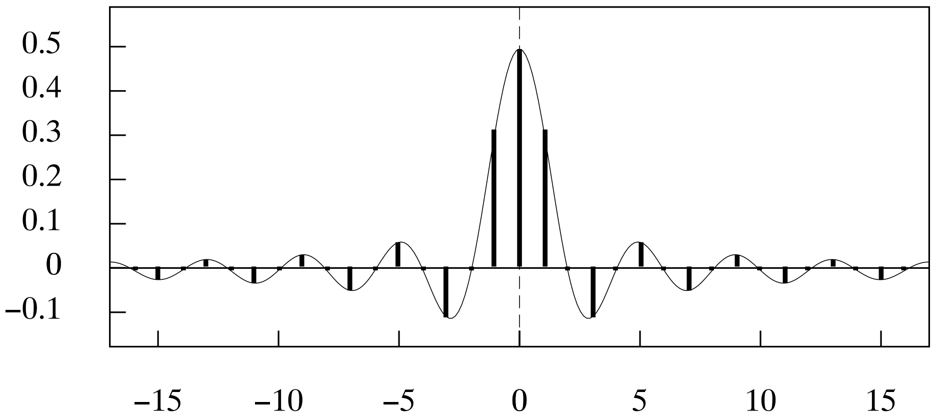

Figure 13. The central coefficients of the Fourier transform of the frequency response of an ideal lowpass filter with a cut-off point at $\omega = \pi/2$. The sequence of coefficients extends indefinitely in both directions. The coefficients are the sampled ordinates of a sinc function.

is the (inverse) Fourier transform of the periodic frequency response function:

$$\beta_k = \frac{1}{2\pi}\int_{-\omega_d}^{\omega_d} e^{i\omega k} d\omega = \begin{cases} \dfrac{\omega_d}{\pi}, & \text{if } k = 0; \\ \dfrac{\sin(k\omega_d)}{\pi k}, & \text{if } k \neq 0\ . \end{cases} \tag{150}$$

Such a frequency response presupposes a doubly-infinite data sequence, in so far as it represents the relative amplification and attenuation of trigonometrical functions that are defined over the entire real line.

The coefficients of (150) form a doubly infinite sequence, of which a central part is illustrated in Figure 13; and, in order to obtain a practical filter, it seems that one must truncate the sequence, retaining only a limited number of its central elements. This truncation gives rise to a filter of which the frequency response has certain undesirable characteristics. (See Figure 19 for an example.)

In particular, there is a ripple effect whereby the gain of the filter fluctuates within the pass band, where it should be constant with a unit value, and within the stop band, where it should be zero-valued. Within the stop band, there is a corresponding problem of leakage whereby the truncated filter transmits elements that ought to be blocked

However, it is clear that an ideal filter can be implemented in the frequency domain by preserving the ordinates of the Fourier transform of the data that are associated with frequencies less than ω_d and by setting all other ordinates to zero. This is a matter of applying the following set of weights to the Fourier ordinates:

$$\lambda_j = \begin{cases} 1, & \text{if } j \in \{-d, \ldots, d\}, \\ 0, & \text{otherwise.} \end{cases} \tag{151}$$

By applying an inverse discrete Fourier transform to these weights, the

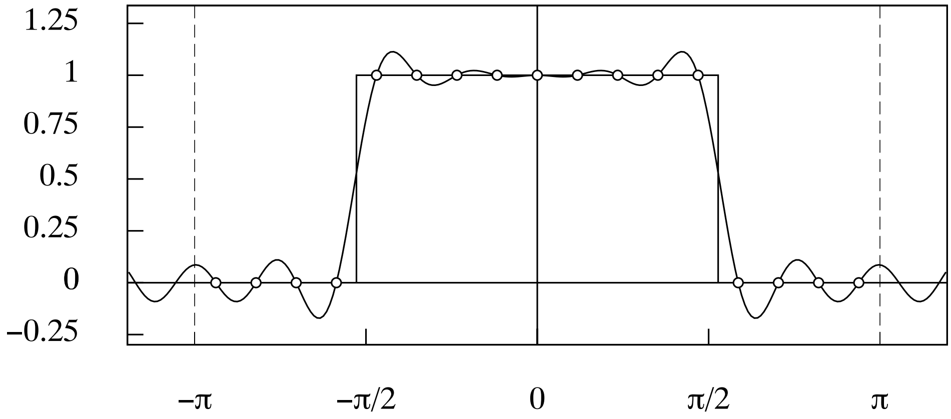

Figure 14. The frequency response of the 17-point wrapped filter defined over the interval $[-\pi, \pi)$. The values at the Fourier frequencies are marked by circles.

coefficients of a circular filter are obtained, of which the values are given by

$$\beta^\circ(k) = \begin{cases} \dfrac{2d+1}{T}, & \text{if } k = 0, \\ \dfrac{\sin([d+1/2]\omega_1 k)}{T\sin(\omega_1 k/2)}, & \text{for } k = 1, \ldots, [T/2], \end{cases} \tag{152}$$

where $\omega_1 = 2\pi/T$. These coefficients would be obtained by wrapping coefficients of (150) around a circle of circumference T and adding the overlying values:

$$\beta_k^\circ = \sum_{j=-\infty}^{\infty} \beta_{jT+k}. \tag{153}$$

Applying the wrapped filter to the finite data sequence via a circular convolution is equivalent to applying the original filter to an infinite periodic extension of the data sequence.

The function of (152) is just an instance of the Dirichlet kernel—see Pollock (1999), for example. Figure 14 depicts the frequency response for this filter at the Fourier frequencies, where $\lambda_j = 0, 1$ in the case where $\omega_d = \pi/2$. It also depicts the continuous frequency response that would be the consequence of applying an ordinary filter with these coefficients to a doubly-infinite data sequence.

8.1 Applying the Fourier method to trended data.

In an ideal application of the Fourier method, it should be possible to wrap the data sequence $y_t; t = 0, \ldots, T-1$ seamlessly around the circle, such that there is no disjunction at the point where the head of the sequence joins the tail. To achieve such an effect, it is common to taper the data so as reduce both ends to zero. To avoid corrupting the sample data, the taper can be applied to some extrapolations of the ends of the sample. However, a data sequence that follows a linear trend is not amenable to tapering, since there is liable to be a radical disjunction at the point where the head joins the tail.

The periodic extension of the linearly trended sequence, which would be generated by travelling around the circle indefinitely, has a saw tooth profile. The corresponding spectrum or periodogram has a *one-over-f* profile that descends, as the frequency increases, in the manner of a rectangular hyperbola, from a high point that is adjacent to the zero frequency to a low point at the limiting frequency. Unless the data are adequately detrended, such a spectrum will serve to conceal all but the most prominent of the harmonic characteristics of the data.

There are two simple ways in which the data may be detrended. The first, which has been described already in section 7.3, is to apply the difference operator to the data as many times as are necessary to reduce them to stationarity. The components that are extracted by filtering the differenced data can be re-inflated, in the manner indicated by equations (133) to (137), to obtain the components of the original data.

We denote the data by y and their differences by $g = Q'y$. The filtered sequence that underlies the trend is denoted by d and the vector of initial conditions by d_*. Then, if we set $\Sigma = I$, the relevant equations for delivering the estimate x of the trend component are

$$x = S_* d_* + Sd \qquad \text{and} \qquad d_* = (S_*' S_*)^{-1} S_*'(y - Sd). \tag{154}$$

The detrended sequence is $h = y - x$. Underlying the detrended sequence is the filtered sequence $k = g - d$, from which the detrended data component may be obtained directly via the equations

$$h = S_* k_* + Sk \qquad \text{and} \qquad k_* = -(S_*' S_*)^{-1} S_*' Sk. \tag{155}$$

Another way of reversing the effects of a differencing operation that has been applied to the data to reduce them to stationarity is to re-inflate the Fourier ordinates of the filtered sequence, using values from the frequency response function of the anti-differencing summation operator. Once the ordinates had been re-inflated within the frequency domain, they can be transformed into the time domain to produce the filtered sequence.

This method is applicable only to components that are bounded away from the zero frequency, since the summation operator has infinite gain at zero. (See Figure 8.) However, if one wishes to apply a lowpass filter to the data, then one has the option of applying the complementary highpass filter and of subtracting the filtered sequence from the original data to generate the lowpass component.

The second way of detrending the data is to extract a polynomial component via an ordinary or a generalised least-squares regression according to the formula of (120). The formula will allow greater weight to be given to the points at both ends of the sample, to ensure that the interpolated curve passes through their midst. This can be achieved by allowing Σ^{-1} to be a diagonal matrix with large values at the ends. In this way, a disjunction in the wrapped version of the residual sequence, or in its periodic extension, can be avoided.

Example. Figure 15 shows the logarithms of the data on aggregate household expenditure in the U.K for the years from 1956 to 2005, throught which a smooth trajectory has been interpolated. This has been obtained by selecting

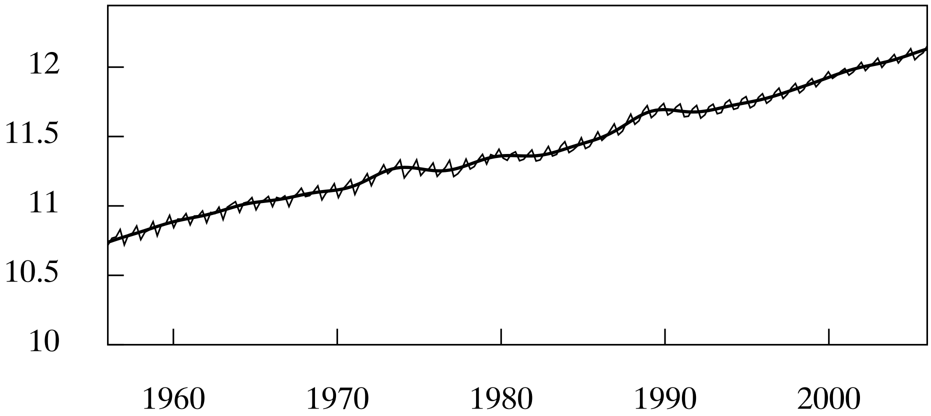

Figure 15. The logarithms of quarterly household expenditure in the U.K., for the years 1956 to 2005, together with an interpolated trend.

the Fourier coefficients of the twice-differenced data that correspond to frequencies in the interval $[0, \pi/8]$. This frequency band has been chosen in the light of the periodogram of Figure 6, which shows that it contains an isolated spectral structure.

The sequence that has been synthesised from these coefficients has been reinflated in the manner indicated by (154) to produce the trajectory. The result of this procedure is a composite of the trend and the business cycle. The same trajectory of aggregate expenditure would have been obtained by adding the business cycle that is depicted in Figure 5 to the linear trend of Figure 4.

9. Band-limited processes

The majority of the methods that we have described for extracting the components of an econometric data sequence presuppose that the data can be described by a univariate ARIMA model. The spectral density function of an ARIMA process is supported on the entire frequency interval $[0, \pi]$, where its ordinates are strictly positive with the possible exception of a few zero-valued ordinates that constitute a set of measure zero. Such zero values will be attributable to the presence of unit roots within the moving-average operator.

It is commonly assumed that the component parts of an aggregate econometric sequence can also be described by ARIMA models. It is on this basis that the Wiener–Kolmogorov filters are derived. However, reference to the periodogram of Figure 6 and to others like it suggests that the components often reside within strictly limited frequency bands which are separated by dead spaces where the spectral ordinates are virtually zeros.

In many circumstances, the disparity between the assumptions underlying the Wiener–Kolmogorov filters and the nature of the data to which they are applied has no adverse effects. A lowpass filter that achieves a gradual transition from a pass band to a stop band within the region of a spectral dead space will be as effective in extracting a low-frequency trend component as is a frequency-domain filter that achieves an abrupt transition between two adjacent Fourier frequencies.

The ordinates at times $t = 0, \ldots, T-1$ of the business cycle that is represented in Figure 5 have been obtained by a Fourier method; but they might have been obtained as well by applying the Butterworth filter of order $n = 6$ and with a nominal cut-off frequency of $\omega_d = \pi/4$ radians, of which the gain is depicted in Figure 11. The principal advantage of the Fourier method, in this context, lies in the ease with which a continuous function can be synthesised from the Fourier coefficients.

Difficulties do arise when an attempt is made to estimate the parameters of an ARMA model from data such as those of Figure 5. A natural objective is to attempt to characterise the business cycle via the parameters of a fitted ARMA model. Such a model is liable to be applied to a seasonally adjusted version of the data, for which the periodogram will lack the spectral spike at the seasonal frequency of $\pi/2$ and at the harmonic frequency of π.

An AR(2) model with complex roots is the simplest of the models that might be appropriate to the purpose. The modulus of its roots should reveal the damping characteristics of the cycles, and their argument should indicate the angular velocity or, equivalently, the length, of the cycles. However, such a model will invariably deliver estimates that imply real-valued roots, which fail adequately to represent the dynamics of the business cycle. (See Pagan, 1997, for example.)

The problem of estimating the business cycle also affects the model-based approaches to econometric signal extraction, which depend upon the prior estimation of an aggregate ARIMA model or upon the estimation of ARIMA components. A business cycle component is usually missing from such models, since the estimation fails to deliver the appropriate complex roots. However, it is straightforward to include a business cycle component with a pre-specified frequency in a disaggregated structural model. (See Harvey, 1985, for example.)

To obtain parametric estimates of the business cycle, it is necessary to remove from the data all but the relevant low-frequency components. This is achieved by selecting the relevant Fourier coefficients from which the business cycle can be constituted via a Fourier synthesis in the manner of (14). Thereafter, it is necessary to sample the continuous function at a rate that will ensure that the Nyquist frequency π corresponds to the highest frequency that is present in the component. A successful ARMA model which represents the complex dynamics of the business cycle can be estimated from the resampled data sequence.

The Shannon–Whittaker sampling theorem indicates that the resampled data contains sufficient information to reconstitute the continuous business-cycle function.

9.1 The Shannon–Whittaker sampling theorem.

Let $x(t)$ be a square integrable continuous signal of which the Fourier transform $\xi(\omega)$ is band limited to the frequency interval $[-\omega_d, \omega_d]$. Then, the signal can be recovered from its sampled ordinates provided that these are separated by a time interval of no more than π/ω_d, which is to say the sinusoidal element of the highest frequency within the signal must take at least two sampling intervals to complete a cycle.

To demonstrate this result, we must consider the Fourier representation of a real–valued square-integrable function $x(t)$ defined over the real line. The following are the corresponding expressions for the function $x(t)$ and its Fourier transform $\xi(\omega)$:

$$x(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{\mathrm{i}\omega t}\xi(\omega)d\omega\longleftrightarrow\xi(\omega)=\int_{-\infty}^{\infty}e^{-\mathrm{i}\omega t}x(t)dt. \tag{156}$$

By sampling $x(t)$ at intervals of π/ω_d, a sequence

$$\{x_\tau=x(\tau[\pi/\omega_d]);\tau=0,\pm1,\pm2,\ldots\}$$

is generated. The elements of the sequence and their Fourier transform $\xi_s(\omega)$ are given by

$$\begin{aligned}x_\tau&=\frac{1}{2\omega_d}\int_{-\omega_d}^{\omega_d}\exp\{\mathrm{i}\omega\tau[\pi/\omega_d]\}\xi_S(\omega)d\omega\\ &\longleftrightarrow\\ \xi_S(\omega)&=\sum_{\tau=-\infty}^{\infty}x_\tau\exp\{-\mathrm{i}\omega\tau[\pi/\omega_d]\}.\end{aligned}\tag{157}$$

Since $\xi(\omega)=\xi_S(\omega)$ is a continuous function defined on the interval $[-\omega_d,\omega_d]$, it may be regarded as a function that is periodic in frequency, with a period of $2\omega_d$. Putting the RHS of (157) into the LHS of (156), and taking the integral over $[-\omega_d,\omega_d]$ in consequence of the band-limited nature of the function $x(t)$, gives

$$\begin{aligned}x(t)&=\frac{1}{2\pi}\int_{-\omega_d}^{\omega_d}\left\{\sum_{\tau=-\infty}^{\infty}x_\tau e^{-\mathrm{i}\omega\tau[\pi/\omega_d]}\right\}e^{\mathrm{i}\omega t}d\omega\\ &=\frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}x_\tau\int_{-\omega_d}^{\omega_d}e^{\mathrm{i}\omega(t-[\tau\pi/\omega_d])}d\omega.\end{aligned}\tag{158}$$

The integral on the RHS is evaluated as

$$\int_{-\omega_d}^{\omega_d}e^{\mathrm{i}\omega(t-[\tau\pi/\omega_d])}d\omega=2\frac{\sin(t\omega_d-\tau\pi)}{t-\tau[\pi/\omega_d]}. \tag{159}$$

Putting this into the RHS of (158) gives

$$x(t)=\sum_{\tau=-\infty}^{\infty}x_\tau\frac{\sin(t\omega_d-\tau\pi)}{\pi(t-\tau[\pi/\omega_d])}=\sum_{k=-\infty}^{\infty}x_\tau\phi_d(\tau-k), \tag{160}$$

where

$$\phi_d(t-\tau)=\frac{\sin(t\omega_d-\tau\pi)}{\pi(t-\tau[\pi/\omega_d])}. \tag{161}$$

When $\tau=0$, this becomes an ordinary sinc function that is a continuous function of t, and which is the Fourier transform of the following frequency function:

$$\phi_d(\omega)=\begin{cases}1, & \text{if } |\omega|\in[0,\omega_d];\\ 0, & \text{otherwise.}\end{cases}\tag{162}$$

When $\tau \neq 0$, it represents a sinc function that has been displaced in time by τ intervals of length π/ω_d. The set of such displaced sinc functions constitutes an orthogonal basis for all continuous functions that are band-limited to the frequency interval $[-\omega_d, \omega_d]$.

In the case of a stationary stochastic process, the sampled sequence is not square summable and, therefore, in a strict sense, this proof of the interpolation via the Nyquist–Shannon Theory does not apply. However, the convergence of the interpolation formula of (160), when $x(\tau) = \{x_\tau; \tau = 0, \pm 1, \pm 2, \ldots\}$ is a stationary sequence, can confirmed by considering a sum with $\tau \in [-N, N]$ for some finite integer N. The variance of the sum of discarded terms can be made arbitrarily small by increasing the value of N.

The reconstruction of a continuous function from its sampled ordinates in the manner suggested by the sampling theorem is not possible in practice, because it requires forming a weighted sum of an infinite number of sinc functions, each of which is supported on the entire real line. Nevertheless, a continuous band-limited periodic function defined on a finite interval—which corresponds to the circumference of a circle—can be reconstituted from a finite number of wrapped or periodic sinc functions, which are Dirichlet kernels by another name. However, the most practical means of reconstituting the function is by a simple Fourier synthesis of the sort described by equation (14).

Example. The analysis of the example following (14) suggests that the business cycle of the detrended logarithmic consumption data fits within the frequency band $[0, \pi/8]$. If this structure can be isolated and thereafter mapped into the frequency interval $[0, \pi]$, then it will be capable of being described by an ordinary linear stochastic model of the ARMA variety. For this purpose, the spectral elements that fall outside the frequency range of the business cycle must first be removed. This operation, which constitutes an anti-alias filtering, may be carried out either in the time domain or in the frequency domain.

Given the availability of the spectral ordinates of the data, it is straightforward to operate in the frequency domain by setting the rejected ordinates to zeros. Then, a continuous low-frequency function can be synthesised from the selected ordinates. An example is provided by the interpolated function in Figure 5. The synthesised function can be resampled at a rate that corresponds to the maximum frequency within the spectral structure of the business cycle.

There is some advantage in fitting a trend function that is more flexible than the straight line of Figure 5. Therefore, a fourth degree polynomial has been fitted to the data by a least-squares regression. The effect is to remove some of the power from the Fourier ordinates adjacent to the zero frequency.

The residual sequence from this polynomial interpolation is show in Figure 16, together with an interpolated function that has been synthesised from the Fourier ordinates that lie in the interval $[0, \pi/8]$. This, function, which purports to represent the business cycle, is devoid of any seasonal fluctuations. Figure 17 displays the periodogram of the residual sequence.

After the removal of all elements of frequencies in excess of $\pi/8$ the data may be resampled at 1/8th of the original rate of observation. This simple fractional rate is a convenient one, since it implies taking one in every eight of

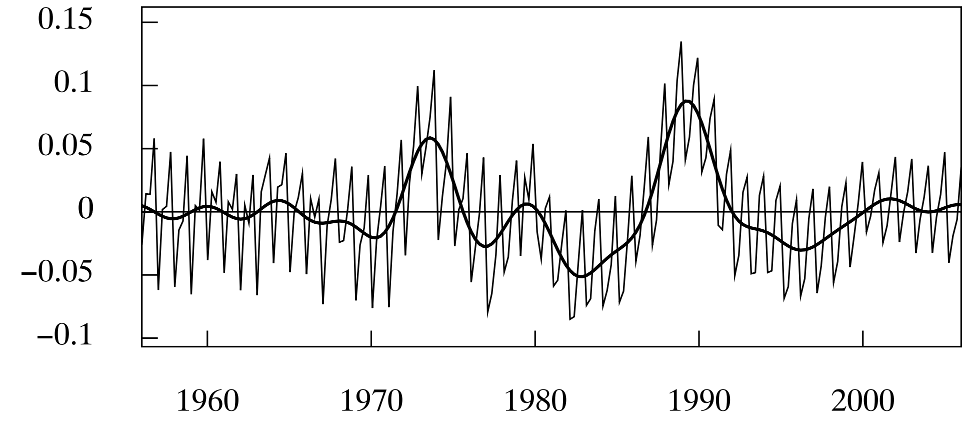

Figure 16. The residuals from fitting a polynomial of degree 4 to the logarithmic expenditure data. The interpolated line, which represents the business cycle, has been synthesised from the Fourier ordinates in the frequency interval $[0, \pi/8]$.

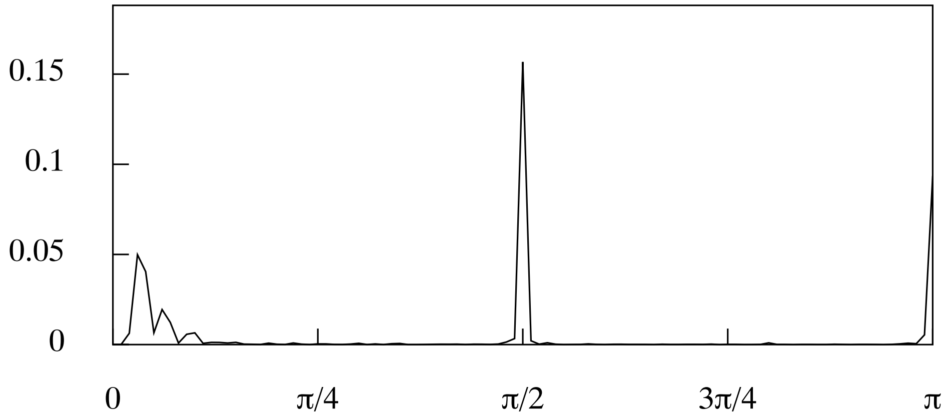

Figure 17. The periodogram of the data sequence of Figure 16.

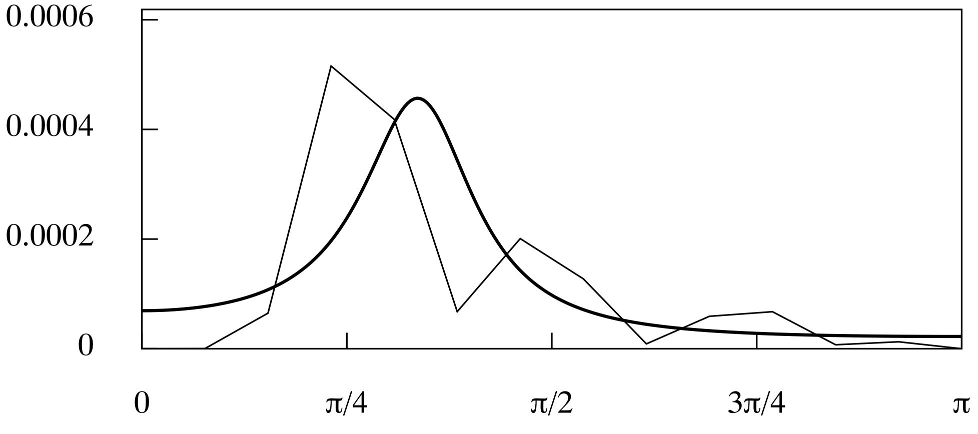

Figure 18. The periodogram of the sub-sampled anti-aliased data with the parametric spectrum of an estimated AR(3) model superimposed.

the anti-aliased data points. In that case, there is no need synthesise a continuous function for the purpose of resampling the data.

The periodogram of the sub sampled anti-aliased data is show in Figure 18 with the parametric spectrum of an estimated AR(3) model superimposed. The periodogram represents a rescaled version of the part of the periodogram of Figure 16 that occupies the frequency range $[0, \pi/8]$; and it appears to be well represented by the parametric spectrum.

The continuous band-limited function of Figure 16 can be recovered from the sub sample by associating to each of its elements an appropriately scaled Dirichlet kernel and, thereafter, by adding these kernels. This demonstrates the one-to-one correspondence that exists between the continuous function and the sub-sampled sequence. This is precisely the one-to-one correspondence that exists between the periodic function $z(t)$, synthesised by equation (14), and its sampled ordinates $\{z_\tau = z(\tau T/N); \tau = 0, 1, \ldots, N-1\}$.

The AR(3) model that underlies the spectral density function of Figure 18 provides a statistical description both of the continuous band-limited function of Figure 16 and of the ordinates sampled from it at the rate of 1 observation in 8 sample periods.

10. Separating the trend and the cycles

The remaining issue to be discussed in this chapter is the matter of separating the trend of an economic data sequence from the cycles that surround it. This is a difficult problem. The trend and the cycles are combined within the same spectral structure and there is rarely any indication, within the periodogram, of where the trend ends and the cycles begin. In the absence of objective criteria for achieving a separation, the definition of the trend is liable to reflect the purposes of the study as well as the circumstances of the economy over the period in question.

A simple prescription that was offered by the pioneering econometrician Tintner (1940, 1953) is that the trend should contain no cyclical motions. This can be interpreted to mean that, if the trend is a differentiable function, then its first derivative should have no more than one local maximum or one local minimum. Such a function can be described as a pure trend. A polynomial function of low degree fitted to the data by least-squares regression is liable to fulfil the requirement; and it can provide an appropriate benchmark for measuring the cyclical variations.

An example of such a trend is the linear function of Figure 5, which has been applied to logarithmic data. When a quadratic function was fitted to the data by least-squares regression, the result was virtually a straight line. The data are from a period that was characterised by uninterrupted economic growth at annual rates that varied little. Therefore, the method of polynomial detrending works well.

In other eras, where there have been marked disruptions, the polynomial method is less appropriate. In order to serve as a benchmark for the ensuing periods of stability, the trend must be made to absorb the disruptions, which implies that it must have a segmented structure. In section 10.2, we will describe a method for achieving this.

A prescription that is to be found in the pioneering work of Burns and Mitchell (1946) is that the business cycle should be defined in terms of a limited band of frequencies. A modern interpretation of this is that the band should comprise the sinusoidal elements of the data that have cyclical durations of no more that eight years and of no less than a year and a half. Such cycles can be extracted from the data via a bandpass filter, as we will discuss below.

The definition seems arbitrary; but it might be justified by proposing that the reactions of economic agents to cycles within the frequency band differ from their reactions to cycles at other frequencies. Thus, it might be argued that their adaptations to cycles of more than eight years duration occur mainly at a subconscious level, whereas cycles of a lesser duration incite conscious reactions.

The growth of an economy may be likened to a process of biological growth, which is affected by events that occur in the course of its evolution. Therefore, a stochastic trend based on the accumulation of random increments has been seen as an appropriate model for an economic trend. This idea has inspired the Beveridge–Nelson decomposition of an ARIMA process, which depicts the trend as an accumulation of disturbances that also give rise to accompanying fluctuations.

In practice, the Beveridge–Nelson decomposition depends upon a linear filter that is applied to the data sequence like any other filter. However, the filtered sequence that represents the trend is liable to include a substantial proportion of the high-frequency elements of the data; and for that reason it may be regarded as unacceptable.

10.1 Bandpass filters

In an attempt to separate a business cycle component from the trend, economists have been resorting increasingly to the use of bandpass filters to implement the definition of Burns and Mitchell (1946). This appears to be in response to the fact that the structural time series methods, which use ARIMA models to represent the unobserved components, fail to isolate the business cycle.

An ideal bandpass filter that transmits all elements within the frequency range $[\alpha, \beta]$ and blocks all others has the following frequency response:

$$\psi(\omega) = \begin{cases} 1, & \text{if } |\omega| \in (\alpha, \beta); \\ 0, & \text{otherwise.} \end{cases} \tag{163}$$

The coefficients of the corresponding time-domain filter are obtained by applying an inverse Fourier transform to this response to give

$$\begin{aligned} \psi(k) = \int_{\alpha}^{\beta} e^{\mathrm{i}k\omega} d\omega &= \frac{1}{\pi k}\{\sin(\beta k) - \sin(\alpha k)\} \\ &= \frac{2}{\pi k}\cos\{(\alpha+\beta)k/2\}\sin\{(\beta-\alpha)k/2\} \\ &= \frac{2}{\pi k}\cos(\gamma k)\sin(\delta k). \end{aligned} \tag{164}$$

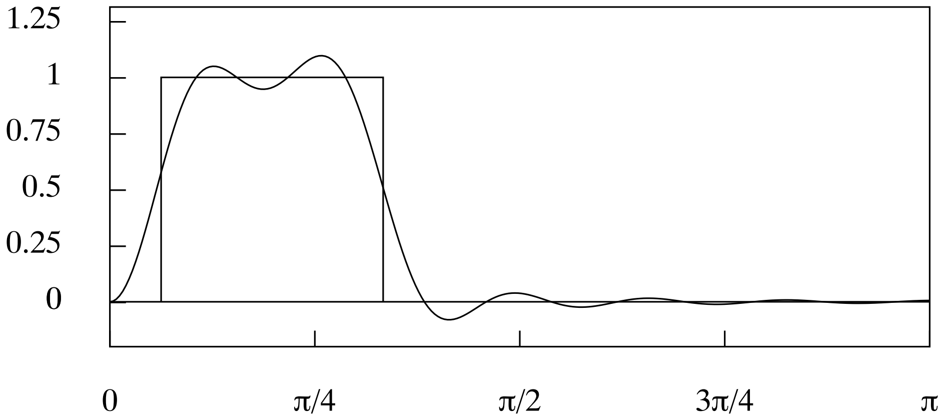

Figure 19. The frequency response of the truncated bandpass filter of 25 coefficients superimposed upon the ideal frequency response. The lower cut-off point is at $\pi/15$ radians (11.25°), corresponding to a period of 6 quarters, and the upper cut-off point is at $\pi/3$ radians (60°), corresponding to a period of the 32 quarters.

Here, $\gamma = (\alpha + \beta)/2$ is the centre of the pass band and $\delta = (\beta - \alpha)/2$ is half its width.

The final equality, which follows from the identity $\sin(A + B) - \sin(A - B) = 2\cos A \sin B$, suggests two interpretations. On the LHS is the difference between the coefficients of two lowpass filters with cut-off frequencies of β and α respectively. On the RHS is the result of shifting a lowpass filter with a cut-off frequency of δ so that its centre is moved from $\omega = 0$ to $\omega = \gamma$.

The process of frequency shifting is best understood by taking account of both positive and negative frequencies when considering the lowpass filter. Then the pass band covers the interval $(-\delta, \delta)$. To convert to the bandpass filter, two copies of the pass band are made that are shifted so that their new centres lie at $-\gamma$ and γ. In the limiting case, the copies are shifted to the centres $-\pi$ and π. There they coincide, and we have $\psi(k) = 2\cos(\pi k)\sin(\delta k)/\pi k$, which constitutes an ideal highpass filter. A bandpass filter can also be expressed as the difference of two such highpass filters

The coefficients of (164) constitute an infinite sequence, which needs to be truncated to produce a practical filter. Alternatively, a wrapped or circular filter may be obtained by sampling the frequency response at a set of equally-spaced points in the frequency range $[-\pi, \pi)$, equal in number to the elements of the data sequence. The wrapped filter is obtained by applying the discrete Fourier transform to the sampled ordinates; and it can be applied to the data sequence by circular convolution.

The z-transform of a set of filter coefficients that are symmetric about the central point and that sum to zero incorporates the factor $(1 - z)(1 - z^{-1}) = -z^{-1}(1-z)^2$. This operator is effective in nullifying a linear trend and in reducing a quadratic trend to a constant. Therefore, such a filter can be applied by linear convolution to a trended data sequence in the expectation that it will produce a stationary filtered sequence.

This is one of the attractions of the truncated bandpass filter that has

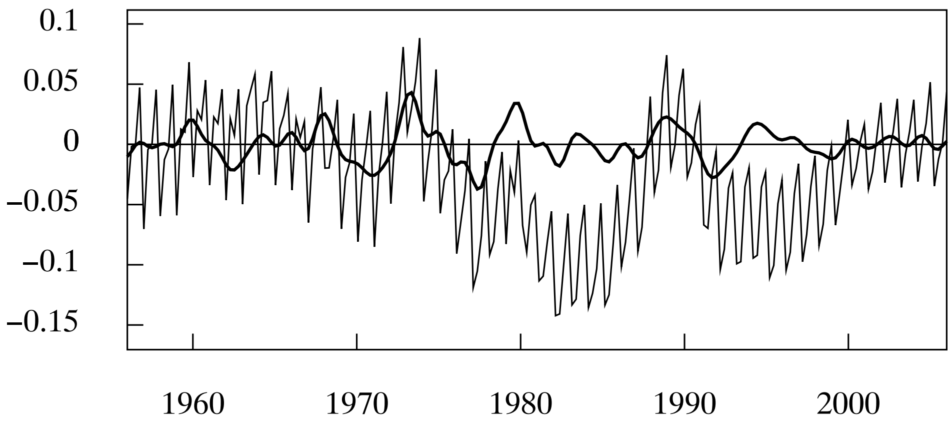

Figure 20. A filtered sequence obtained by applying the bandpass filter of Christiano and Fitzgerald to the logarithms of U.K. household expenditure.

been proposed to economists by Baxter and King (1999). To ensure that the coefficients of the truncated filter do sum to zero, the filter can be expressed as the difference between two truncated versions of the ideal lowpass filter, of which the coefficients have been scaled so as to sum to unity.

The truncated filter has several disadvantages. In the first place, the truncation leads to the phenomenon of leakage that has already been described in section 8. This is illustrated by Figure 19. Also, a finite-order moving-average filter with constant coefficients is incapable of reaching the ends of the sample. This problem occasions a trade-off between the accuracy of the approximation to the ideal filter, which increases with the number of coefficients, and the end-of-sample problem, which is exacerbated by increasing the span of the filter.

There are numerous ways of overcoming the end-of sample problem, including the obvious recourse of extrapolating the sample by forecasting and backcasting it with the help of an ARIMA model that purports to describe the data. Another recourse is to extend the sample by attaching its symmetric reflection to either end. However, if the data are strongly trended this will tend to increase the values at the beginning of the sample and to decrease the values at the end, relative to the values obtained via a linear extrapolation of the sample.

A circular filter should not be applied directly to a trended data sequence. When such a sequence is wrapped around a circle there is liable to be a radical disjunction where the beginning and the end of the sample are joined. The effects of this disjunction are liable to be carried into the filtered sequence in a manner that does not affect the ordinary linear filter. One way of overcoming this difficulty is to apply the circular filter to data that have been reduced to stationarity by differencing. Thereafter, the filtered differenced sequence can be cumulated to obtain an estimate of the business cycle component.

Example. The filter of Baxter and King (1999) is a time-invariant moving average comprising $2q + 1$ of the central coefficients of the ideal infinite-order bandpass filter, which are symmetrically disposed around the central value. These coefficients should re-scaled so that they sum to zero.

The elements of the filtered sequence are given by

$$\begin{aligned}x_t = \phi_q y_{t-q} + \phi_{q-1} y_{t-q+1} + \cdots + \phi_1 y_{t-1} + \phi_0 y_t \\ + \phi_1 y_{t+1} + \cdots + \phi_{q-1} y_{t+q-1} + \phi_q y_{t+q}.\end{aligned} \tag{165}$$

Given a sample $y_0, y_1, \ldots, y_{T-1}$ of T data points, only $T - 2q$ processed values $x_q, x_{q+1}, \ldots, x_{T-q-1}$ are available, since the filter cannot reach the ends of the sample, unless some extrapolations are added to it.

To overcome this difficulty, Chistiano and Fitzgerald (2003) have used a filter that comprises selections of the coefficients of the ideal filter which vary as one moves through the sample. At all times, the central coefficient of the ideal filter is aligned with the current data value. The remainder of the selection consists of the coefficients on either side that fall within the data window. Thus, the filtered values are weighted combinations of all of the sample elements.

In the case of data that might have been generated by a random-walk process, it is proposed to supplement the weighted sum by two additional terms based on the first and the final sample elements, which are the appropriate predictors of the elements of the process that fall outside the data window. In that case, the elements of the filtered sequence will be given by

$$\begin{aligned}x_t = Ay_0 + \phi_t y_0 + \cdots + \phi_1 y_{t-1} + \phi_0 y_t \\ + \phi_1 y_{t+1} + \cdots + \phi_{T-1-t} y_{T-1} + By_{T-1},\end{aligned} \tag{166}$$

where A and B are sums of the coefficients of the ideal filter that lie beyond either end of the data window. Since the filter coefficients must sum to zero, it follows that

$$A = -(\frac{1}{2}\phi_0 + \phi_1 + \cdots + \phi_t) \quad \text{and} \quad B = -(\frac{1}{2}\phi_0 + \phi_1 + \cdots + \phi_{T-t-1}). \tag{168}$$

For data that appear to have been generated by a first-order random walk with a constant drift, it is appropriate to extract a linear trend before filtering the residual sequence. In fact, this has proved to be the usual practice in most circumstances.

It has been proposed to subtract from the data a linear function $f(t) = \alpha + \beta t$ interpolated through the first and the final data points, such that $\alpha = y_0$ and $\beta = (y_{T-1} - y_0)/T$. In that case, there should be $A = B = 0$. This procedure is appropriate to seasonally adjusted data. For data that manifest strong seasonal fluctuations, such as the U.K. expenditure data, a line can be fitted by least squares through the data points of the first and the final years. Figure 20 shows the effect of the application of the filter to the U.K. data adjusted in this manner.

Figure 20 can be compared with Figure 5 and with Figure 16, both of which also purport to show the business cycles that affected the data in question. It is clear that the bandpass filter fails to transmit the appropriate cyclical fluctuations. An explanation for the failure can be found in Figure 6, which shows the periodogram of the linearly detrended data.

The highlighted band in Figure 6 covers the frequency interval $[\pi/16, \pi, 3]$ which, according to Baxter and King, is the frequency range that defines the business cycle. However, this figure indicates that only a small part of the low-frequency component falls within the interval. Therefore, it appears that the definition is at fault. In fact, the leakage that is associated with the filter does allow some of the low-frequency power of the elements that reside in the interval $[0, \pi/16]$ to pass into the filtered sequence.

10.2 Flexible trends and structural breaks

Over a period of a century or so, one can expect to see occasional disturbances that disrupt the steady progress of the economy. To highlight the effects of such breaks, a firm trend function can be fitted to the data to characterise the progress of the economy broadly over the entire period. Such a trend will not be deflected by temporary disruptions, which will be seen in the residual deviations of the data from the trend.

Alternatively, it may be appropriate to absorb the breaks within the trend function. In that case, the trend will not be thrown off course for long by a break; and, therefore, it should serve as a benchmark against which to measure cyclical variations when the economy resumes its normal progress. At best, the residual sequence will serve to indicate how the economy might have behaved in the absence of the break.

Numerous devices have been proposed by economists for accommodating structural breaks, which give rise to segmented trend functions. Mills (2003) has illustrated the effects of some of them by applying them to a common data sequence, which is the annual U.K. output from 1855 to 1999. He has also provided references to an extensive literature in economics concerning structural breaks.

A common theme that unites many of the methods is their use of polynomial segments to represent the trends within subintervals of the data period. There is a problem of how the transition between two adjacent sub-periods should be modelled. This issue has been discussed by Granger and Teräsvirta (1993) and by Teräsvirta (1998). Others have focussed on devising tests to determine the points in time when one statistical regime that describes the data should be replaced by another. Work in this area has been summarised by Perron (2006).

When a smoothing spline is used to interpolate a continuous segmented polynomial function through the data, the smoothness of the function is maintained by imposing the condition that, at the points where they join, the adjacent segments should have equal derivatives, up to some specified order.

The most common smoothing spline is that of Reinsch (1976), which is subject to the condition that the first and second derivatives of adjacent cubic segments should be equal at the joints, which are described as the knots or the nodes. Breaks can be accommodated within such a spline by placing successive nodes in close proximity. Considerable effort has been devoted to developing algorithms that will ensure the optimal placement of the nodes. (See, for example, Luo and Whaba 1997.)

When the abscissae of the nodes correspond to the sample dates, it is possible to increase the flexibility of the spline function by allowing local variations to occur in the smoothing parameter. The same recourse can be used to lend addi-

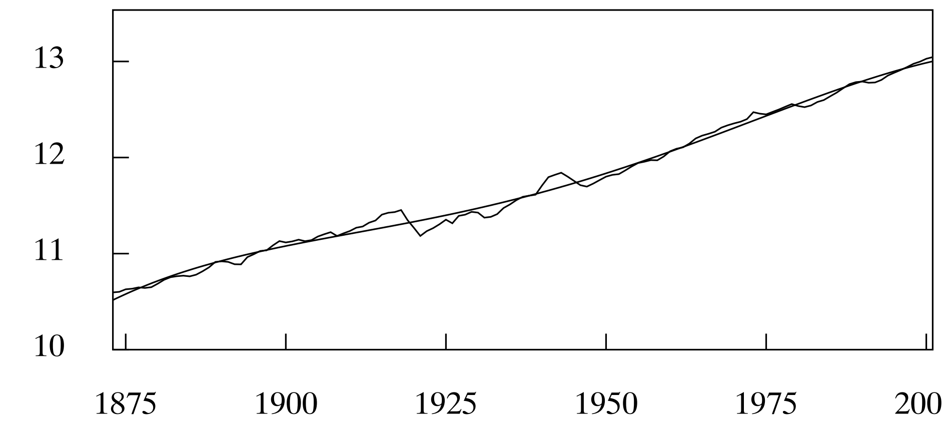

Figure 21. The annual series of the logarithms of real GDP in the U.K., at constant prices, for the years 1873 to 2001. A polynomial of degree 4 has been fitted to the data by least squares regression.

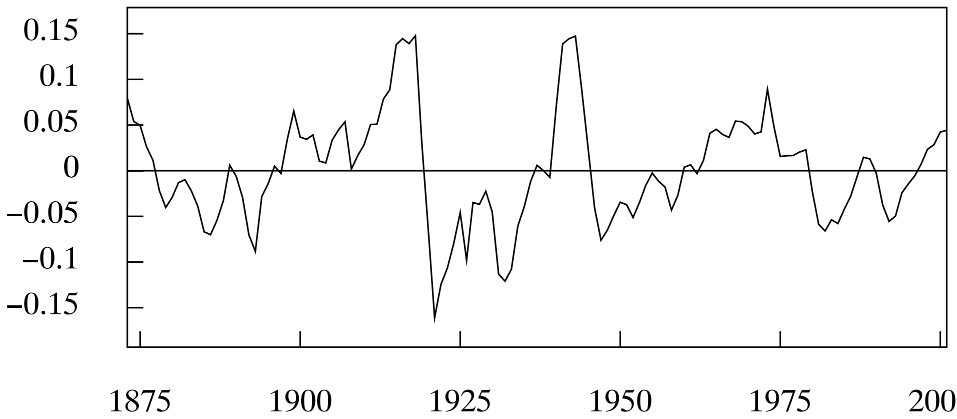

Figure 22. The residual obtained from fitting a polynomial of degree 4 to the logarithmic GDP date of Figure 21.

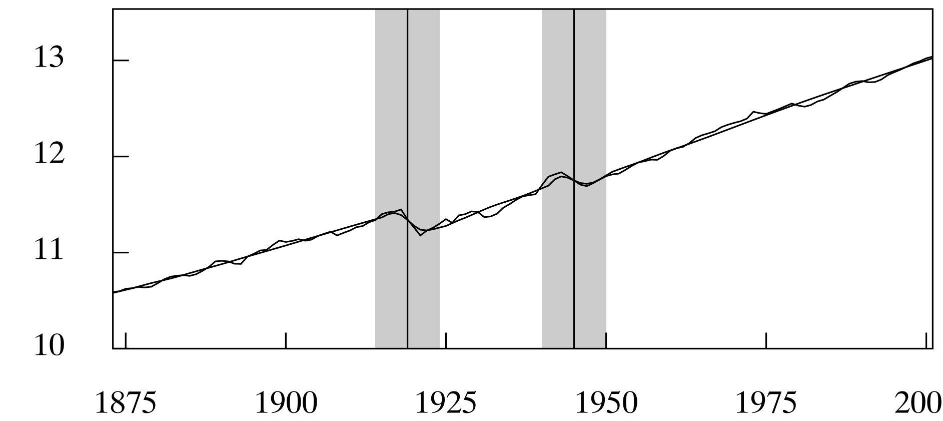

Figure 23. The logarithms of annual U.K. real GDP from 1873 to 2001 with an interpolated trend. The trend is estimated via a filter with a variable smoothing parameter.

tional flexibility to the Hodrick–Prescott filter, which is a device that is appropriate for extracting from noisy data a trend that is generated by a discrete-time process or by a process limited in frequency to the Nyquist value.

The finite-sample version of the Hodrick–Prescott filter is provided by equation (124). Its generalisation is provided by

$$x = y - Q(\Lambda^{-1} + Q'Q)^{-1}Q'y, \tag{167}$$

where $\Lambda = \mathrm{diag}\{\lambda_0, \lambda_1, \ldots, \lambda_{T-3}\}$ is a diagonal matrix of smoothing parameters and Q' is the matrix of the twofold difference operator. In modifying the underlying statistical model of the H–P filter, which is specified by (130), it is the variance σ_δ^2 of the process driving the trend that is allowed to vary, whereas the variance σ_η^2 of the process that is responsible for the errors of observation remains constant.

Setting $\Lambda^{-1} = \lambda^{-1}I$ in (167), which gives the smoothing parameter a globally constant value, produces the Hodrick–Prescott filter. Setting λ_t to a high value where the trend should be stiff and allowing it to take low values where the trend should be flexible will produce a device that can easily absorb structural breaks.

On the assumption that the underlying trend process is limited in frequency by the Nyquist value, it is appropriate to use the method of Fourier interpolation to create a continuous trend based on the elements of the vector x.

Example. An example of a function that fails to accommodate structural breaks is provided by the polynomial of degree 4 that has been interpolated through the logarithms of 129 annual observations of the real GNP of the U.K. This is shown in Figure 21. Figure 22 shows the residual sequence. In both figures, three major events can be recognised. The first is the end of the First World War in 1918, which is followed by a sharp decline in GNP. The second is the recession of 1929 and the third is the end of the Second World War, which is also succeeded by a reduction in income. The recession has less of an impact than one might expect.

Figure 23 shows a trend function that has been fitted using a variable smoothing parameter. In this case, only the end-of-war breaks have been accommodated, leaving the disruptions of the 1929 recession to be expressed in the residual sequence. The effect has been achieved by attributing a greatly reduced value to the smoothing parameter in the vicinity of the post-war breaks. In the areas that are marked by shaded bands, the smoothing parameter has been given a value of 5. Elsewhere, it has been given a high value of 100,000, which results in trend segments that are virtually linear.

11. Summary and conclusions

When confronted by the wide variety of methods that are available for extracting the components of an econometric data sequence, a practitioner is liable to ask for a recommendation of the best method. In the case of business-cycle analysis, there can be no unequivocal answer. The choice of an appropriate method will depend both on the nature of the data and on the purpose of the analysis. It may also depend on the aesthetic preferences of the analyst.

Nevertheless, the choice of a method ought to be made with a view to its effects in the frequency domain. Econometricians working with temporal sequences are, nowadays, paying increasing attention to the frequency aspects of their analyses; and this is where the major emphasis of the present chapter has been placed.

One of the difficulties in analysing business cycles is that there is no unequivocal definition of what constitutes a trend. Often, a clearly defined structure that combines the trend and the cycles can be discerned within the data. An example of the successful extraction of a combination of trend and cycles that has been identified by spectral methods is provided by Figure 15. However, there is hardly ever a case where the data indicates a point within the frequency spectrum of this structure where the trend ends and the cycles begin.

The only unequivocal definition of the trend that might be offered is that it must have a monotonic trajectory that is devoid of cycles, which means, in practice, that it should be modelled by a polynomial of low degree. This was the practice of the generation of pioneering econometricians to which Tintner belonged.

Latterly, this approach has fallen out of favour amongst econometricians. Nowadays, they are liable to describe polynomial trends as deterministic trends, which are contrasted with stochastic trends. The latter are regarded as capable of more realistic representations of economic behaviour. In particular, a stochastic trend can represent a cumulation of random events that effect the development of an economy in the course of time, in the way that the circumstances of their early lives can affect the physical statures of human beings.

Polynomial trends are an essential element within linear models of stochastic accumulation, whether they be represented in continuous time or in discrete time. Therefore, although the conceptual distinction may be a clear one, the practical distinction between a stochastic trend generated by an ARIMA process and a polynomial trend buried in noise is by no means as clear cut as, at first, it might seem to be.

The distinction becomes even more tenuous in the case of an ARIMA model that incorporates stochastic drift. Therefore, notwithstanding the recent efforts of several econometricians, it does not seem to us to be fruitful to employ statistical tests in an attempt to determine which of these alternative statistical structures actually underlies the data.

An opinion to which we adhere in this chapter is that the trend is best regarded as an analytic device, as opposed to an object that subsists within the data that might be uncovered by an appropriate technique. If the trend is to be regarded as an artificial benchmark, then its definition depends largely on what one is intending to measure.

In some cases, when the economy has had an uninterrupted progress, it is straightforward to define an appropriate benchmark. A case in point has been the U.K. economy over the years 1956–2005, of which the aggregate consumption is portrayed in Figures 4–6. For that period, a log-linear trend function provides a datum about which to measure the cyclical variations in consumption.

In other eras and over longer periods, where there have been substantial disruptions to the progress of the economy, the matter becomes more complicated. To highlight the major disruptions, it is appropriate to fit a polynomial of a limited degree over the entire span of the data. An example is provided by Figure 19. There, a fourth degree polynomial, which adheres quite well to the data in the main, also reveals the uncommon circumstances in the periods surrounding the ends of the two world wars.

If the purpose is also to illustrate the normal workings of the economy, then it may be appropriate to fit similar polynomial trends of low degrees to the sub periods that did not experience any disruptions. The overall result will be a segmented curve; and the issue arises of how to join the segments.

The answer that is favoured in this chapter is illustrated in Figure 21, which shows the effect of a filter with a variable smoothing parameter. The resulting curve comprises segments that are virtually straight lines that are interspersed by short segments with rapidly changing slopes.

The disjunctions that occur within the data sequence as consequences of disruptions and breaks give rise to spectra that extend over the entire frequency range. Unless the breaks are absorbed within the trend, the residual sequence will fail to manifest the band-limited structure that we might expect to see in normal periods. Therefore, one of the criteria of a successful elimination of the break is the restoration of a band-limited spectral structure to the trend-cycle component within the residual sequence.

The recognition that, at least for limited periods, the trend-cycle complex is liable to be confined to a limited frequency band gives rise to further opportunities, but it also poses additional problems. The opportunities arise from the possibility of using a Fourier synthesis to create a continuous analytic function to represent the business cycle in isolation or the trend and cycle in combination.

In Figure 5, the business cycle has been synthesised from a limited number of the low-frequency Fourier ordinates of the linearly detrended logarithmic data. The combination of the trend and the cycle can be formed by adding the business-cycle function to the linear trend of Figure 4.

The analytic nature of these functions means that they are amenable to differentiation; and their tuning points are identified as the points where the first derivatives are zero-valued. This method of finding the turning points may be contrasted with the very different procedure of Bry and Boschan (1971) which had been widely adopted by governmental statistical offices, but which often reaches doubtful conclusions.

A problem posed by band-limited processes is that they cannot easily be represented by the ARMA models that are ubiquitous in time-series analysis. Such models are based on the assumption that the spectra of the processes that they represent are supported on a frequency interval that extends as far as the Nyquist frequency, which represents the limit of what is observable in sampled data.

It is often supposed that a discrete-time ARMA process is representing an underlying continuous-time process that has an unbounded frequency range. If that were the case, then the spectral density function defined over the Nyquist interval would be the product of a process of aliasing, whereby the elements of the continuous process that fall outside the Nyquist interval are attributed to frequencies that are inside.

In section 5, we have described a correspondence that would exist between processes that are unbounded in frequency and the discrete time models that would serve to represent them. Nevertheless, we have expressed doubts about the relevance to business-cycle analysis of such unbounded processes.

In section 9, we have argued that processes that are limited in frequency to subintervals of the Nyquist interval, in the way that the business cycle is limited, can be resampled at a reduced rate so as to map their limited supports onto the full Nyquist interval. Thereafter, the ordinary methods of ARMA modelling can be applied to the resampled data. In that case, the Nyquist–Shannon sampling theorem indicates that there is a one-to-one correspondence between the discretely sampled process and an equivalent process in continuous time.

By these means one should be able to find an ARMA model that will capture the dynamics of the business cycle and reveal them in terms of the estimated parameters. In particular, the modulus and the arguments of the roots of the autoregressive operator should reveal the damping characteristics of the cycles and their average periods.

A modern interpretation by Baxter and King (1999) of a prescription of Burns and Mitchell (1946) is that the business cycle should be defined as a band-limited process containing cyclical elements of durations of no less that one and a half years and not exceeding eight years. This appears, at first sight, to be an unequivocal definition. However, there are difficulties in implementing it accurately. Thus, it is commonly believed that the filter that would be required to realise this definition must comprise an infinite number of coefficients; and this is not practical.

In place of the infinite-order filter, a truncated approximation is commonly employed that comprises a limited number of the central coefficients. Such a filter is beset by the phenomenon of leakage, whereby the powerful low-frequency elements that would be blocked by the ideal filter find their way into the estimated business cycle. (In fact, a superior approximation is available in the form of a rational filter. See Pollock (2003b), for example, where a rational function is employed to create a sharp lowpass filter.)

However, it has been show here that the bandpass definition can be fulfilled by selecting the appropriate ordinates of the Fourier transform of the detrended data. The equivalent filter in the time domain is a wrapped or circular filter. Whereas such filters avoid the leakage that besets approximate bandpass filters, they deliver inappropriate estimates of the business cycle when they adhere strictly to the Baxter–King bandpass definition. Moreover, it seems that any success that the approximate bandpass filter may have in representing the business cycle must be due, in some measure, to the leakage.

The conclusion that we have reached ultimately is that, whereas it is sometimes possible to identify a trend-cycle complex within the data, there can be no definitive definition of what constitutes the trend and what, in consequence, must constitute the cyclical component. Therefore, it seems that one must be liberal in allowing any definitions that seem to fulfil their intended purposes. Even when the purpose is mistaken or unfulfilled, we should not automatically reject the resulting definition or the estimates to which it gives rise.

A Computer Program

The computer program that has been used in connection with this chapter is available at the following web address:

`http://www.le.ac.uk/users/dsgp1/`

References

Baxter, M., and R.G. King, (1999), Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series, *Review of Economics and Statistics,* 81, 575–593.

Bergstrom, A.R., (1984) Continuous Time Stochastic Models and Issues of Aggregation Over Time, In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics,* 2, 1146–1212, North-Holland, Amsterdam.

Bergstrom, A.R., (1988) The History of Continuous-Time Econometric Models, *Econometric Theory,* 4, 350–373.

Bergstrom, A.R., and K.B. Nowman (2007), *A Continuous Time Econometric Model of the United Kingdom with Stochastic Trends,* Cambridge University Press, Cambridge.

Beveridge, S., and C.R. Nelson, (1981), A New Approach to the Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the Business Cycle, *Journal of Monetary Economics,* 7, 151–172.

Bloomfield, P., (1976), *Fourier Analysis of Time Series: An Introduction,* John Wiley and Sons, Chichester.

Box, G.E.P., and G.M. Jenkins, (1976), *Time Series Analysis: Forecasting and Control, Revised Edition,* Holden Day, San Francisco.

Bry G., and C. Boschan, (1971), *Cyclical Analysis of Time Series: Selected Procedures and Computer Programs,* National Bureau of Economic Research, New York.

Burman, J.P., (1980), Seasonal Adjustment by Signal Extraction, *Journal of the Royal Statistical Society, Series A,* 143, 321–337.

Burns, A.M., and W.C. Mitchell, (1946), *Measuring Business Cycles,* National Bureau of Economic Research, New York.

Caporello, G., and A. Maravall, (2004), *Program TSW: Revised Reference Manual,* Working Paper 0408, Servicio de Estudios, Banco de España.

Christiano, L.J., and T.J. Fitzgerald, (2003), The Band-pass Filter, *International Economic Review,* 44, 435–465.

Fuller, W.A., (1976), *Introduction to Statistical Time Series,* John Wiley and Sons, New York.

Gómez, V., (2001), The Use of Butterworth Filters for Trend and Cycle Estimation in Economic Time Series, *Journal of Business and Economic Statistics,* 19, 365–373.

Granger C.W.T., and T. Teräsvirta, (1993), *Modelling Nonlinear Economic Relationships,* Oxford University Press.

Gray, R.M., (2002), *Toeplitz and Circulant Matrices: A Review,* Information Systems Laboratory, Department of Electrical Engineering, Stanford University, California, (`http://ee.stanford.edu/gray/~toeplitz.pdf`).

Harding, D., and A. Pagan, (2002), Dissecting the Cycle: A Methodological Investigation, *Journal of Monetary Economics.* 49, 365–381

Harvey, A.C., (1985), Trends and Cycles in Macroeconomic Time Series, *Journal of Business and Economic Statistics,* 3, 216–228,

Harvey, A.C., (1989), *Forecasting, Structural Time Series Models and the Kalman Filter,* Cambridge University Press, Cambridge.

Harvey, D.I., and T. Mills, (2003), Modelling Trends in Central England Temperatures, *Journal of Forecasting,* 22, 35–47.

Hillmer, S.C., and G.C. Tiao, (1982), An ARIMA-Model-Based Approach to Seasonal Adjustment, *Journal of the American Statistical Association,* 77, 63–70.

Hodrick, R.J., and E.C. Prescott, (1980), *Postwar U.S. Business Cycles: An Empirical Investigation,* Working Paper, Carnegie–Mellon University, Pittsburgh, Pennsylvania.

Hodrick R.J., and E.C. Prescott, (1997), Postwar U.S. Business Cycles: An Empirical Investigation, *Journal of Money, Credit and Banking,* 29, 1–16.

Kolmogorov, A.N., (1941), Interpolation and Extrapolation, *Bulletin de l'academie des sciences de U.S.S.R.,* Ser. Math., 5, 3–14.

Koopman, S.J., A.C. Harvey, J.A. Doornick, and N. Shephard, (2007), *STAMP 8.0: Structural Time Series Analyser Modeller and Predictor: The Manual,* Timberlake Consultants Press, London.

Leser, C.E.V., (1961), A Simple Method of Trend Construction, *Journal of the Royal Statistical Society, Series B,* 23, 91–107.

Luo. Z., and Grace Wahba, (1997), Hybrid Adaptive Splines, *Journal of the American Statistical Association,* 92, 107–116.

Maravall, A., and D.A. Pierce, (1987), A Prototypical Seasonal Adjustment Model, *Journal of Time Series Analysis,* 8, 177–193.

Mills, T.C., (2003), *Modelling Trends and Cycles in Economic Time Series,* Palgrave Macmillan, Basingstoke.

Pandit, S.M., and S.M. Wu, (1975), Unique Estimates of the Parameters of a Continuous Stationary Stochastic Process, *Biometrika,* 62, 497–501.

Pagan, A., (1997), Towards an Understanding of Some Business Cycle Characteristics, *The Australian Economic Review,* 30, 1–15.

Pedregal, D.J., C.J. Taylor and P.C. Young (2004), System Identification, Time Series Analysis and Forecasting: The Captain Toolbox. Handbook v1.1.

Perron. P., (2006), Dealing with Structural Breaks, pps. 278–352 in T.C. Mills and K.D. Patterson (eds.), *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory,* Palgrave Macmillan, Basingstoke.

Phadke, M.S., and S.M. Wu, (1974), Modelling of Continuous Stochastic Processes from Discrete Observations with Applications to Sunspot Data, *Journal of the American Statistical Association,* 69, 325–329.

Pierce, D.A., (1979), Signal Extraction in Nonstationary Time Series, *The Annals of Statistics,* 6, 1303–1320.

Pollock, D.S.G., (1999), *A Handbook of Time-Series Analysis, Signal Processing and Dynamics,* Academic Press, London.

Pollock, D.S.G., (2000), Trend Estimation and De-Trending via Rational Square Wave Filters, *Journal of Econometrics,* 99, 317–334.

Pollock, D.S.G., (2002a), Circulant Matrices and Time-Series Analysis, *The International Journal of Mathematical Education in Science and Technology,* 33, 213–230.

Pollock. D.S.G, (2002b), A Review of TSW: the Windows Version of the TRAMO-SEATS Program, *Journal of Applied Econometrics,*17,291–299.

Pollock, D.S.G., (2003a), Sharp Filters for Short Sequences, *Journal of Sttistical Inference and Planning,* 113, 663–683.

Pollock, D.S.G., (2003b), Recursive Estimation in Econometrics, *Journal of Computational Statistics and Data Analysis,* 44, 37–75.

Pollock, D.S.G., (2006), Wiener–Kolmogorov Filtering, Frequency-Selective Filtering and Polynomial Regression, *Econometric Theory,* 23, 71–83.

Priestley, M.B., (1989), *Spectral Analysis and Time Series,* Academic Press, London.

Reinsch, C.H., (1976), Smoothing by Spline Functions, *Numerische Mathematik,* 10, 177–183.

Teräsvirta, T., (1998), Modelling Nonlinear Economic Relationships with Smooth Transitions, pps. 507–552 in A. Ullah and D.E.A. Giles, (eds.) *Handbook of Applied Economic Statistics,* Marcel Dekker, New York.

Tintner, G., (1940), *The Variate Difference Method,* John Wiley and Sons, Bloomington, Indiana.

Tintner, G., (1953), *Econometrics,* John Wiley and Sons, New York.

Wahba, Grace, (1978), Improper Priors, Spline Smoothing and the Problem of Guarding against Model Errors in Regression, *Journal of the Royal Statistical Society, Series B,* 40, 364–372.

Whittaker, E.T., (1923), On a New Method of Graduation, *Proceedings of the Royal Society of Edinburgh,* 44, 77–83.

Whittle, P., (1983), *Prediction and Regulation by Linear Least-Square Methods, Second Revised Edition,* Basil Blackwell, Oxford.

Wiener, N., (1941) *Extrapolation, Interpolation and Smoothing of Stationary Time Series,* Report on the Services Research Project DIC-6037, Published in book form in 1949 by MIT Technology Press and John Wiley and Sons, New York.