IMPROVED FREQUENCY-SELECTIVE FILTERS

by D.S.G. Pollock

Queen Mary, University of London

Email: stephen_pollock@sigmapi.u-net.com

This paper gives an account of some techniques for designing recursive frequency-selective filters which can be applied to data sequences of limited duration which may be nonstationary. The designs are based on the Wiener–Kolmogorov theory of signal extraction which employs a statistical model of the processes generating the data. The statistical model may be regarded as an heuristic device which is designed with a view to ensuring that the resulting signal-extraction filters have certain preconceived properties.

Date: December, 2001

Key words: Signal extraction, Linear filtering, Filter design, Trend estimation, Frequency-domain analysis

JEL Classification: C22

1. Introduction

This paper is concerned with the design and implementation of frequency-selective recursive filters in which the filtered output is incorporated on the input side, at various lags, via a series of feedback terms. Such devices are commonly described as infinite impulse response (IIR) filters. Filters that are devoid of feedback are apt to be described as finite impulse response (FIR) filters.

Relatively little attention has been paid to the design of such IIR digital filters, with the result that there are few clear design principles that can be followed. Part of the reason for this deficiency lies in the success of the FIR filters which are usually preferred when one wishes to avoid inducing non linear phase effects in the output. Also, some of the digital IIR filters that have been derived by translating classical analogue designs have proved eminently successful; and this has reduced the incentive for innovation.

In this paper, our purpose is to design filters with very few coefficients, which show a rapid transition from the pass band to the stop band. Improvements in the rate of transition are usually secured by increasing the number of coefficients within a given family of filters; but we wish to avoid this recourse which is liable to exacerbate the start-up problems that affect the processing of short nonstationary sequences

In our case, the rapid transition is achieved via a filter that operates close to the borders of instability. This can lead to the rapid propagation of numerical rounding error when the filter is realised in low-precision arithmetic—i.e. 16 bits or less. The filters become increasingly unstable as their poles approach the

perimeter of the unit circle; and, therefore, we need to pay particular attention to the location of the poles and to the values of their moduli.

The filter designs presented in this paper are derived with reference to Wiener–Kolmogorov principle of signal extraction which is expounded briefly in section 3 of this paper. (For the original expositions, see Kolmogorov (1941) and Wiener (1950). For an accessible modern treatment of the theory see Whittle (1983).) Filters that fulfil this principle unconditionally are of the bidirectional variety, which must be applied in at least two passes which run forwards and backwards through the data. When there are two passes, the forward pass, which uses a causal or backwards-looking filter, induces a phase lag in the output. The reverse pass, which is applied to the intermediate output, induces an equal and opposite phase lag in reversed time; and the two phase lags cancel each other.

Wiener–Kolmogorov filters are usually derived from a statistical model which depicts the data as a combination of a signal and a noise component, both of which have well-defined statistical properties. One of the advantages of this approach is that the finite-sample versions of the resulting filters can be implemented via the Kalman filter for which computer programs are readily available. (See, for example, Brown and Hwang (1992), Bomhoff (1994), and Koopman, Shephard and Doornick (1999).)

The Kalman filter is a relatively complex device, which is aimed at generating the minimum-mean-square-error estimates at each point in an accumulating information set—see Pollock (1999). The associated smoothing filter, which is applied in a backward pass, effects the retrospective enhancement of the estimates using the information that has accumulated since they were first calculated. If one can take the information set as given, as is the case in all off-line processing, then there are some simpler ways of proceeding. The algorithm presented in section 7 of the paper, which is based on a Cholesky decomposition, is simpler than the Kalman Filter.

In some cases, the quality of the Wiener–Kolmogorov filter depends upon the degree of realism of the underlying statistical model. In this paper, there will be no intention of creating realistic statistical models of the data components; and the models, which are no more than heuristic devices, will be determined solely with a view to ensuring that the resulting signal-extraction filters have certain preconceived properties.

Our object, in the first instance, will be to design a prototype highpass filter with a rapid transition between the pass band and the stop band, which occurs in the vicinity of the frequency value of $\omega_c = \pi/2$ which divides the range of the digital frequencies. Our point of departure will be the classical Butterworth digital filter; and we shall seek to improve on its performance. Sections 4 and 5 of the paper are devoted to this purpose.

Once the prototype filter is available, it can be transformed easily into a filter with an alternative cut-off point. In fact, it can be transformed just as easily into a lowpass filter, a bandpass filter or a bandstop filter. In sixth section of the paper, we shall outline the methods of transforming the filter.

In the final section of the paper, we shall show how to apply a lowpass filter to the task of extracting a trend from a short nonstationary sequence.

2. Filtering in Econometrics

In econometric analysis, linear filters are widely used for removing trends from data series and for removing seasonal fluctuations. They are also used for extracting a range of so-called unobserved components into which an econometric time series can be decomposed. (The techniques of seasonal adjustment are treated in the text of Frances (1996) and in the book edited by Hylleberg (1992). For another excellent but little-know account of the methods of seasonal adjustment, see Stier (1980).)

There are no unique prescriptions for how the filters should be constructed. Their designs are affected by differing views on how the components of a time series have originated and how they are related to each other. However, a way of starting which is common is to model the aggregate time series as an ARIMA process. Such processes are generated by applying a rational filter that has unit roots in its denominator to a white-noise process, which is the forcing function that provides the motive power.

In one perception, a macroeconomic trend represents a long-run growth path, which is affected by the disturbances that impinge upon the economy. In data that have been purged of any seasonal component, the fluctuations that surround the trend will be taken to represent the processes by which the trajectory of the economy converges to the growth path. In this view, the trend and the fluctuations share the same motive power.

If the aggregate non-seasonal series is modelled as an ARIMA process, then the trend and the fluctuations can be separated via a partial-fraction decomposition of the ARIMA operator. The decomposition produces an unstable filter that has the unit roots in its denominator and a stable filter that comprises the remaining roots of the ARIMA denominator. The unstable filter, in conjunction with the white-noise forcing function, accounts for the trend. (A positive constant is liable to be added to the forcing function to create an upwards drift.) The stable filter, in conjunction with the same white-noise forcing function, accounts for the transitory fluctuations. A detailed exposition of this approach is given in the seminal paper of Beveridge and Nelson (1989), and the underlying concepts have been elucidated in a recent paper of Morley *et al.* (1999).

In another view, the trend of a series is an autonomous component that has its own independent motive power, and the fluctuations, which may include seasonal fluctuations, have no effect on its long-term trajectory. The trend is liable to be attributed to an ARIMA process which is independent of the ARMA process generating the transitory fluctuations, and of any seasonal ARIMA process that may be included. Adding these processes creates the ARIMA process that models the aggregate series.

In this case, the trend and the fluctuations can be estimated from the data by using the Wiener–Kolmogorov method of signal extraction that is pursued

in the present paper. The first stage in the process is to find a partial-fraction decomposition of the autocovariance generating function of data, which creates a set of autocovariance structures for its mutually independent unobserved components. The filter for extracting an individual component can be derived by forming the ratio of the autocovariance generating function of the component and the autocovariance generating function of the data. The technique is represented in the present paper by equation (8).

Amongst the filters most commonly used by economists for the purposes of estimating trends is the Hodrick–Prescott filter (1980), which is closely related to the smoothing spline of Reinsch (1976). This filter, which has a single adjustable parameter, is derivable by applying the Wiener–Kolmogorov principle to an abbreviated model in which the trend, which is generated by a second-order random walk, has white-noise disturbances added to it. This filter is the subject of a recent treatise by Kaiser and Maravall (2001).

The model that underlies the Hodrick–Prescott filter often provides an inadequate representation of the processes generating the data. Therefore it is unusual to determine the parameter of the filter by fitting the model to the data. Instead, its value is commonly determined by rule of thumb.

To avoid the arbitrariness of a rule of thumb, it has been proposed that trend estimation and seasonal adjustment in economic time series should be conducted within the framework of a fully-featured model of the data, which attributes separate ARIMA processes to each of the data components. It is argued that, if the model is properly constructed, then the parameters of the various signal-extraction filters will emerge automatically from the process of fitting the model to the data. This approach has been advocated in recent surveys by Maravall (1995) and by Gómez and Maravall (2001), and it is also the basis of the model-based methodology of Harvey (1989).

The implication of depicting the components of an aggregate series as the products of ARIMA processes of low orders is that one expects to find a substantial overlap in the frequency spectra of the individual components. When the spectra do overlap, there is bound to be difficulty in separating the components. The gain profiles of filters aimed at extracting the various components will show very gradual transitions from the pass band to the stop band; and the effect is that the same sinusoidal elements of the data will find their way, with varying degrees of attenuation, into more than one of the estimated components.

Contrary to the presuppositions of the ARIMA-model-based approach, there is evidence that the components of some econometric time series reside in well-defined frequency bands that have little or no overlap. In the most favourable circumstances, the components are separated by spectral dead spaces where there are no elements of any significant power. If the tools are sharp enough, then the components can be extracted without loss or confusion. In such cases, we should use other methods in preference to those of the model-based approach.

The ideal filter for isolating the spectral components of a time series that

fall within a particular range of frequencies is a rectangular window in which the transition from the pass band to the stop band occurs at a point. Baxter and King (1999) have recently investigated the use of moving-average FIR approximations to the ideal filter. However, for a good approximation, such filters must have a wide span involving many coefficients.

The use of a wide-span filter presupposes a lengthy data sequence of which the ends, that the filter cannot reach, can be left unprocessed. In econometric analysis, however, the data sequences are often of a strictly limited duration and they are liable to be strongly trended. In such circumstances, we cannot afford to use a wide-span filter unless the sample can be extrapolated by forecasting and backcasting.

(We should note that, according to one interpretation, some of the more sophisticated finite-sample techniques do entail implicit forecasting techniques. The very successful method of Burman (1980), which implements Wiener–Kolmogorov filters in finite samples, entails explicit extrapolations, but these are, strictly speaking, unnecessary, and the method can be practised within the confines of the sample.)

The Wiener–Kolmogorov filters that are proposed in this paper go much of the way towards meeting the objective of the ideal frequency-selective filter, and they do so at the cost of only a handful of parameters. The manner in which we propose to implement the filters overcomes the problems of processing nonstationary series, and it does not suffer from any impediment in processing the ends of the data sequence.

3. Wiener–Kolmogorov Filters

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

$$y(t) = \xi(t) + \eta(t), \tag{1}$$

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 stationary stochastic processes that are mutually independent. It follows that the autocovariance generating function of the data is the sum of the autocovariance generating functions of its two components. Thus

$$\gamma^{yy}(z) = \gamma^{\xi\xi}(z) + \gamma^{\eta\eta}(z) \quad \text{and} \quad \gamma^{y\xi}(z) = \gamma^{\xi\xi}(z). \tag{2}$$

The signal sequence $\xi(t)$ is estimated via a linear transformation of the data sequence which may be denoted by

$$x(t) = \beta_\xi(L)y(t). \tag{3}$$

Here, $\beta_\xi(L) = \sum_j \beta_j L^j$ is a power-series operator which forms a linear combination of the data elements. The symbol L^j stands for the jth power of the lag operator, the effect of which is that $L^j y(t) = y(t-j)$. The negative powers

of the operator have the opposite effect of reaching forward in time. Thus, $L^{-j}y(t) = y(t+j)$.

Depending on the practical circumstances affecting its implementation, the filter $\beta_\xi(L)$ may be a causal FIR filter (with $j \in [0, p]$), a symmetric two-sided FIR filter (with $j \in [-p, p]$), a causal IIR filter (with $j \in [0, \infty]$), or a bidirectional IIR filter (with no bounds on j).

The principle of minimum-mean-square-error estimation indicates that the sequence formed from the errors of interpolation must be orthogonal to the data sequence, which is to say that the two sequences must be statistically uncorrelated. In the case of a bidirectional IIR filter, the condition of orthogonality is expressed by writing

$$y(t) \perp \{e(t) = \xi(t) - x(t)\} \quad \text{or, equivalently,} \quad \gamma^{ye}(z) = 0, \tag{4}$$

where

$$\begin{aligned} \gamma^{ye}(z) &= \gamma^{y\xi}(z) - \gamma^{yx}(z) \\ &= \gamma^{\xi\xi}(z) - \gamma^{yy}(z)\beta_\xi(z) \end{aligned} \tag{5}$$

is the generating function of the covariances of the errors and the data. Setting this to zero and rearranging gives the normal equations

$$\gamma^{yy}(z)\beta_\xi(z) = \gamma^{\xi\xi}(z). \tag{6}$$

The positive definite autocovariance generating functions within equation (6) are both subject to a so-called Cramér–Wold decomposition which allows them to be written as

$$\gamma^{yy}(z) = \phi(z^{-1})\phi(z) \quad \text{and} \quad \gamma^{\xi\xi}(z) = \delta(z^{-1})\delta(z). \tag{7}$$

For the bidirectional filter, the solution to equation (6) is therefore

$$\beta_\xi(z) = \frac{\gamma^{\xi\xi}(z)}{\gamma^{yy}(z)} = \frac{\delta(z^{-1})\delta(z)}{\phi(z^{-1})\phi(z)}. \tag{8}$$

When z is identified with L and z^{-1} is identified with $L^{-1} = F$, this corresponds the product of a real-time filter $\delta(L)/\gamma(L)$ and a reverse-time filter $\delta(F)/\gamma(F)$.

The principle of minimum-mean-square-error estimation can also be applied in finding the Wiener–Kolmogorov filter for extracting the noise component of the data sequence. This is given by

$$\beta_\eta(z) = \frac{\gamma^{\eta\eta}(z)}{\gamma^{yy}(z)} = 1 - \beta_\xi(z). \tag{9}$$

The condition of complementarity, whereby $\beta_\xi(z) + \beta_\eta(z) = 1$, is amongst the defining characteristics of the Wiener–Kolmogorov filters.

In the case of a causal filter, where $\beta(z)$ contains only nonnegative powers of z, the normal equations take the form of

$$\left[\phi(z^{-1})\phi(z)\beta(z)\right]_+ = \left[\gamma^{\xi\xi}(z)\right]_+, \tag{10}$$

where the subscripted + is to indicate that only the part of the series which contains nonnegative powers of z is to be taken. (This is Whittle's (1983) notation.) The equations imply that

$$\phi(z)\beta(z) = \left[\frac{\gamma^{\xi\xi}(z)}{\phi(z^{-1})}\right]_+, \tag{11}$$

where the subscripted + is missing from the left hand side on account of the fact that it contains only non-negative powers of z. Dividing the equation on both sides by $\phi(z)$ gives a solution for $\beta(z)$ in the form of

$$\beta(z) = \frac{1}{\phi(z)}\left[\frac{\gamma^{\xi\xi}(z)}{\phi(z^{-1})}\right]_+. \tag{12}$$

If the symmetric function $\gamma^{\xi\xi}(z)$ has only a finite number of nonzero coefficients, then the term bearing the + sign will represent the numerator of the rational function denoted by $\beta(z)$.

The minimum-mean-square-error criterion, which the filter of (12) is designed to fulfil, is concerned as much with the avoidance of an excessive phase effect as with the accuracy of the frequency selection. Therefore, for the purposes of frequency-selective filtering, the filter of (12) is liable to be a poor substitute for the bidirectional filter of (8). Moreover, if one is prepared to overlook its phase effect, then the unidirectional filter $\delta(z)/\phi(z)$, which is, so to speak, one half of the bidirectional filter, is liable to provide a superior device for the purposes of frequency selection. However, whereas the phase effect can be overlooked in some engineering applications, such as the analysis of random mechanical vibrations, it is usually of prime importance in economic applications where the correct identification of turning points etc. depends upon phase-neutral filtering.

In the econometric analysis of time series, it is common to model a nonstationary process via an autoregressive operator with roots of unit value, which are on the boundary of instability. Thus, if $\xi(t)$ represents a nonstationary trend component within the equation $y(t) = \xi(t) + \eta(t)$, then, for some value of d, it will transpire that $(I-L)^d\xi(t) = \zeta(t)$ is a stationary process, as is $\eta(t)$. Multiplying the equation throughout by $(I-L)^d$ will give

$$\begin{aligned}(1-L)^d y(t) &= \zeta(t) + (1-L)^d\eta(t)\\ &= \zeta(t) + \kappa(t) = g(t).\end{aligned} \tag{13}$$

The procedure for extracting the estimates $x(t)$ and $h(t)$ of $\xi(t)$ and $\eta(t)$, respectively, begins by filtering $g(t)$ to give $z(t) = \beta_\xi(L)g(t)$ and $k(t) =$

$\beta_\eta(L)g(t)$, which are the estimates of $\zeta(t)$ and $\kappa(t)$, respectively. Thereafter, the sought-after estimates can be obtained by cumulating their differenced versions. However, observe that, since $(1-L)^d h(t) = k(t)$, there is

$$x(t) = y(t) - \frac{k(t)}{(1-L)^d}, \tag{14}$$

and so only one of the estimates, $k(t)$, needs to be cumulated. The other estimate, $x(t)$, can be obtained by subtraction. Moreover, if $\beta_\eta(L)$ contains the factor $(1-L)$ to a degree $n \geq d$—which will prove to be the case—then applying the filter $\beta^*_\eta(L) = (1-L)^{-d}\beta_\eta(L)$ to $g(t)$ will produce $h(t) = \beta^*_\eta(L)g(t)$ directly. Thus one can avoid the need to cumulate the filtered sequence, which means that there will be no need for starting values.

4. Designing the Prototype Filter

An ideal frequency-selective filter is a phase-neutral square-wave filter for which the gain is unity over a certain range of frequencies, described as the passband, and zero over the remaining frequencies, which constitute the stopband. In a lowpass filter β_L, the passband covers a frequency interval $[0, \omega_c)$ ranging from zero to a cut-off point. In the complementary highpass filter β_H, it is the stopband which stands on this interval. Thus

$$|\beta_L(e^{i\omega})| = \begin{cases} 1, & \text{if } \omega < \omega_c \\ 0, & \text{if } \omega > \omega_c \end{cases} \quad \text{and} \quad |\beta_H(e^{i\omega})| = \begin{cases} 0, & \text{if } \omega < \omega_c \\ 1, & \text{if } \omega > \omega_c. \end{cases} \tag{15}$$

The object in constructing a practical frequency-selective filter is to find a rational function, embodying a limited number of coefficients, whose frequency response is a good approximation to the square wave.

In this section, we shall derive a pair of complementary filters that fulfil the specifications of (15) approximately for a cut-off frequency of $\omega_c = \pi/2$. Once we have designed these prototype filters, we shall be able to apply a transformation that shifts the cut-off point from $\omega = \pi/2$ to any other point $\omega_c \in (0, \pi)$.

A preliminary step in designing a pair of complementary filters is to draw up a list of specifications which can be fulfilled in practice. We shall be guided by the following conditions:

$$\begin{aligned}
&\text{(i)} && \beta_L(z^{-1}) = \beta_L(z), \quad \beta_H(z^{-1}) = \beta_H(z), && \textit{Phase-Neutrality} \\
&\text{(ii)} && \beta_L(z) + \beta_H(z) = 1, && \textit{Complementarity} \\
&\text{(iii)} && \beta_L(-z) = \beta_H(z), \quad \beta_H(-z) = \beta_L(z), && \textit{Symmetry} \\
&\text{(iv)} && |\beta_H(1)| = 0, \quad |\beta_H(-1)| = 1, && \textit{Highpass Conditions} \\
&\text{(v)} && |\beta_L(1)| = 1, \quad |\beta_L(-1)| = 0. && \textit{Lowpass Conditions}
\end{aligned} \tag{16}$$

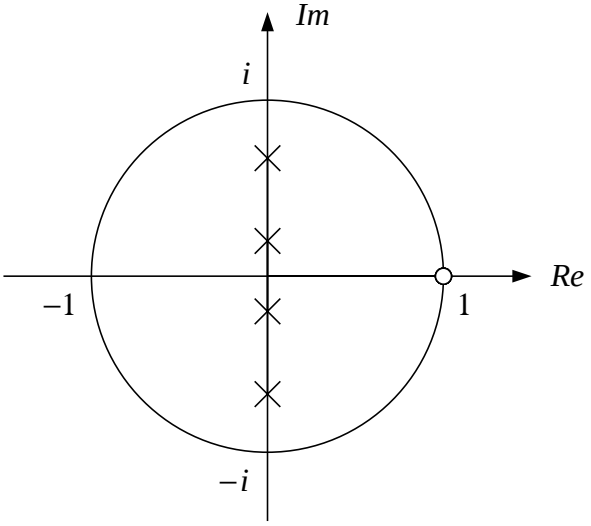

Figure 1. The pole–zero diagram of the fourth-order prototype unidirectional Butterworth filter with a nominal cut-off point of 90 degrees.

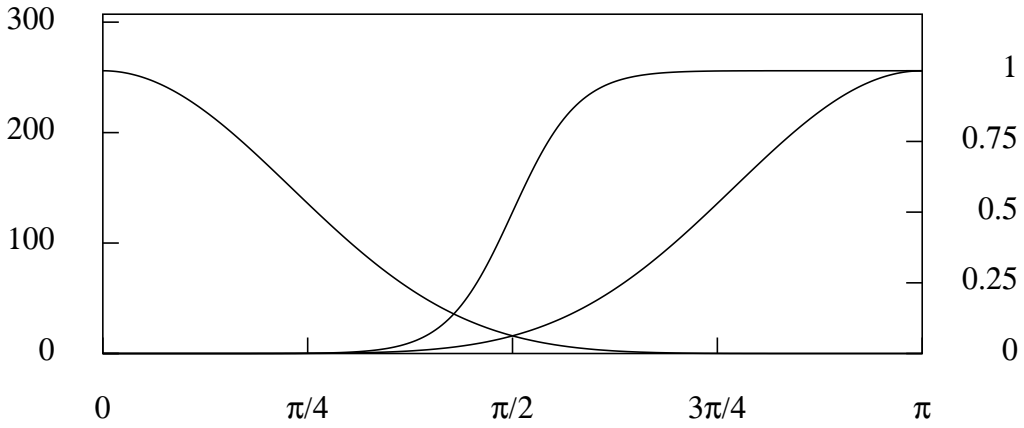

Figure 2. The gain of the bidirectional prototype fourth-order Butterworth filter with a nominal cut-off point of 90 degrees.

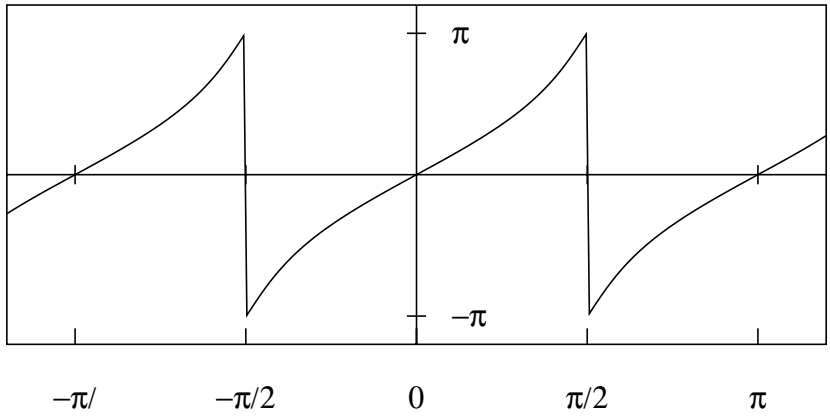

Figure 3. The phase response of the prototype fourth-order unidirectional Butterworth filter with a nominal cut-off point of 90 degrees.

As we have already noted, a bidirectional rational filter in the form of (8) fulfils already the condition (i) of phase neutrality. Given the specification of (8), the condition of complementarity under (ii) implies that the filters must take the form of

$$\beta_L(z) = \frac{\delta_L(z^{-1})\delta_L(z)}{\phi(z^{-1})\phi(z)} \qquad \text{and} \qquad \beta_H(z) = \frac{\delta_H(z^{-1})\delta_H(z)}{\phi(z^{-1})\phi(z)}, \tag{17}$$

where

$$\phi(z^{-1})\phi(z) = \delta_L(z^{-1})\delta_L(z) + \delta_H(z^{-1})\delta_H(z). \tag{18}$$

The roots of this polynomial come in reciprocal pairs and, upon the factorisation of the RHS, the poles that lie outside the unit circle are assigned to $\phi(z)$ whilst those that lie inside are assigned to $\phi(z^{-1})$.

Next, the condition of symmetry under (iii) implies that, when it is reflected about the axis of $\omega = \pi/2$, the frequency response of the lowpass filter becomes the frequency response of the highpass filter. This implies that the cut-off point ω_c must be located at the mid-point frequency of $\pi/2$. The condition requires that $\phi(z) = \phi(-z)$, which implies that every root of $\phi(z) = 0$ must be a purely imaginary number. The condition also requires that

$$\delta_L(z) = \delta_H(-z) \qquad \text{and} \qquad \delta_H(z) = \delta_L(-z). \tag{19}$$

It remains to fulfil the conditions (iv) and (v). Condition (iv) indicates that $\delta_H(z)$ must have a zero at $z = 1$, which is to say that it must incorporate a factor in the form of $(1 - z)^n$. Condition (v) indicates that $\delta_L(z)$ must have a zero at $z = -1$, which is to say that it must incorporate a factor in the form of $(1 + z)^n$.

These conditions (iv) and (v) do not preclude the presence of further factors in $\delta_L(z)$ and $\delta_H(z)$; but, if λ is a root of $\delta_H(z)$, then $-\lambda$ must be a root of $\delta_L(z)$. Thus, in general, the polynomials can take the forms of

$$\begin{aligned} \delta_L(z) &= (1 + z)^n \prod_{i=1}^{m}(1 + \lambda_i z) \qquad \text{and} \\ \delta_H(z) &= (1 - z)^n \prod_{i=1}^{m}(1 - \lambda_i z), \end{aligned} \tag{20}$$

where set of the parameters $\{\pm\lambda_i^{-1}; i = 1, \ldots, m\}$, which are roots of the polynomials, contains conjugate pairs of complex numbers.

The resulting filters are identical to those which would arise from applying the Wiener–Kolmogorov principle of signal extraction to the task of isolating the components $\xi(t)$ and $\eta(t)$ of the model

$$\begin{aligned} y(t) &= \xi(t) + \eta(t) \\ &= (I + L)^n \prod_{i=1}^{m}(I + \lambda_i L)\nu(t) + (1 - L)^n \prod_{i=1}^{m}(I - \lambda_i)\varepsilon(t), \end{aligned} \tag{21}$$

wherein $\nu(t)$ and $\varepsilon(t)$ are independent white-noise processes with a common variance.

Some simple and convenient forms of the filters are obtained by setting

$$\delta_L(z) = (1+z)^n \qquad \text{and} \qquad \delta_H(z) = (1-z)^n. \tag{22}$$

On putting the specification of (22) into (17) and (18), we find that

$$\begin{aligned}\beta_L(z) &= \frac{(1+z^{-1})^n(1+z)^n}{(1+z^{-1})^n(1+z)^n + (1-z^{-1})^n(1-z)^n} \\ &= \frac{1}{1+\left(i\dfrac{1-z}{1+z}\right)^{2n}}\end{aligned} \tag{23}$$

and that

$$\begin{aligned}\beta_H(z) &= \frac{(1-z^{-1})^n(1-z)^n}{(1+z^{-1})^n(1+z)^n + (1-z^{-1})^n(1-z)^n} \\ &= \frac{1}{1+\left(i\dfrac{1+z}{1-z}\right)^{2n}}.\end{aligned} \tag{24}$$

These will be recognised as instances of the digital translation of the Butterworth analogue filter which is familiar in electrical engineering—see, for example, Roberts and Mullis (1987). The translation from the analogue domain to the digital domain is by virtue of the bilinear transformation

$$s(z) = \frac{z-1}{z+1}, \tag{25}$$

which is a mapping from the z-plane, which contains the poles and zeros of the discrete-time digital filter, to the s-plane, which contains the poles of the continuous-time analogue filter. The Butterworth filter represents the simplest way of fulfilling the design features listed under (16)(i)–(v).

We may note that Gómez (1999) has recently discussed the Butterworth digital filter from the point of view of econometric analysis as has Pollock (1997).

5. The Frequency Response of the Filters

We may begin the analysis of the filters by examining the attributes of the Butterworth filter. Figure 1 is the pole–zero diagram for a fourth-order unidirectional digital Butterworth filter. This is for the reverse-time filter, which has its poles inside the unit circle. Figure 2 shows the spectral density functions of the low-frequency signal $\xi(t) = (I+L)^n\nu(t)$ and the high-frequency noise component $\eta(t) = (I-L)^n\varepsilon(t)$ for the case where $V\{\nu(t)\} = V\{\eta(t)\}$, together

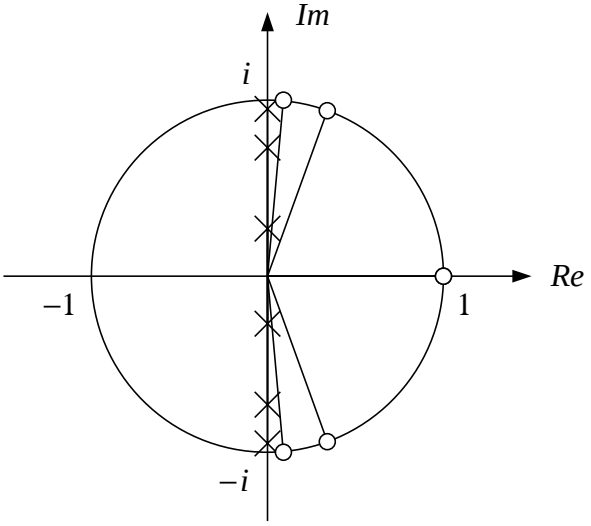

Figure 4. The pole–zero diagram of the unidirectional prototype sharp filter of order $n = 6$ with a nominal cut-off point of 90 degrees.

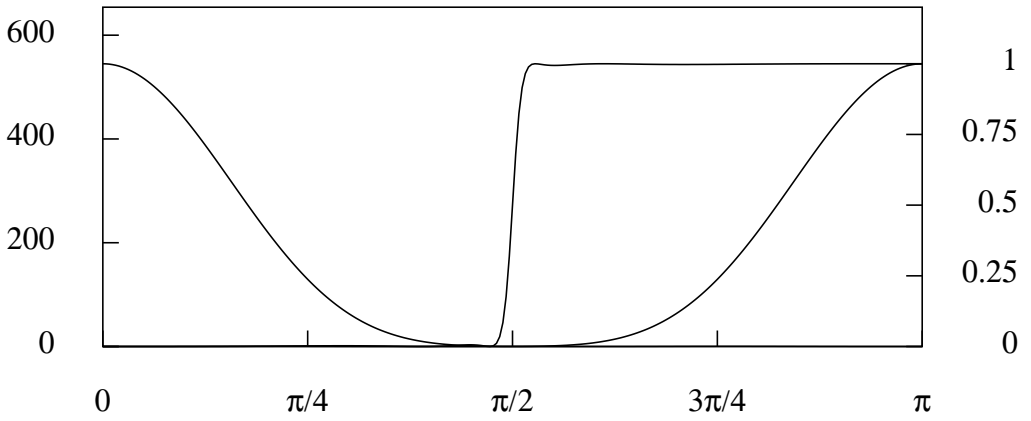

Figure 5. The gain of the bidirectional prototype sharp filter of order $n = 6$ with a nominal cut-off point of 90 degrees.

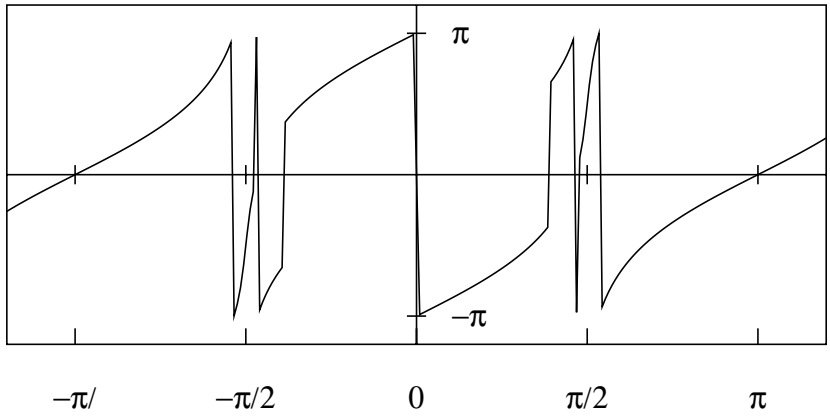

Figure 6. The phase response of the unidirectional prototype sharp filter of order $n = 6$ with a nominal cut-off point of 90 degrees.

with the gain of the resulting highpass bidirectional filter. Figure 3 shows the phase response of the unidirectional filter.

The transition in the gain of this highpass filter occurs in the middle of the frequency range where the low-frequency spectrum of $\xi(t)$ gives way to the high-frequency spectrum of $\eta(t)$, which is its mirror image.

The advantage of the Butterworth filter is the ease with which analytic expressions can be obtained for the poles of the filter. The availability of these expressions means that the Crámer–Wold factorisation $\gamma_D(z) = \phi(z^{-1})\phi(z)$, upon which implementation of the filter depends, is readily available. In the case of the more general filters which we shall propose, there are no available analytic expressions for the poles, which means that the Crámer–Wold factorisation must be obtained by a numerical method.

The algorithm of Wilson (1969), which is based on the Newton–Raphson procedure, is an effective way of achieving the factorisation; and versions which are coded in *C* and in *Pascal* have been provided by Pollock (1999)—see, also, Laurie (1989, 1982).

The disadvantage of the digital Butterworth filter is in the relatively slow rate of transition between the pass band and the stop band, which is due to the substantial overlap of the spectra of the low-frequency and the high-frequency components of $y(t)$. The greater the overlap of the spectra, the more gradual will be the transition.

The overlap can be reduced, and the rate of transition increased, by increasing the order of the Butterworth filter, which means increasing the number of zeros located at $z = 1$, in the case of the highpass filter, or at $z = -1$, in the case of the lowpass filter. However, the rate of transition can be increased more effectively by locating the additional zeros at other points within the stopband which are nearer to the cut-off point.

Consider the case of a high pass filter which is designed to fulfil the conditions of complementarity and symmetry. Then it is easy to see that placing a zero on the unit circle at the frequency angle of $\omega = \frac{1}{2}\pi - \epsilon$ will ensure that the transition from stop band to pass band will occur within a band of 2ϵ radians centered on the cut-off frequency of $\omega_c = \pi/2$.

Figure 4 shows the pole-zero diagram of a unidirectional sixth-order digital filter in which two zeros are located on the unit circle where it intersects the horizontal axis and in which the remaining four zeros are in conjugate pairs located at angles of 70 degrees and 85 degrees from the horizontal. Figure 5 show the gain of the bidirectional filter which is formed by compounding the causal filter with the reverse-time filter. It is notable that the rate of transition is far more rapid than it is for the Butterworth filter.

If a zero of the prototype filter is located on the unit circle, then its argument is bound to be less than the cut-off frequency of $\omega_c = \pi/2$. A zero of $\delta_L(z)$ at $z = e^{-\pi/2}$ would be accompanied by a zero in $\delta_H(z)$ at the same location. In that case, the denominator polynomial $\gamma_D(z) = \delta_L(z^{-1})\delta_L(z) + \delta_H(z^{-1})\delta_H(z)$ would also have a zero of unit modulus at $\omega = \pi/2$. This is in violation of the stability condition which requires that the poles of the causal filter must

fall outside the unit circle and that the poles of the reverse-time filter must fall inside. It is clear that the narrowing of the transition band is inevitably accompanied by a worsening problem of dynamic instability.

There are two recourses for mitigating the problem of instability. The first is to place the zeros on the unit circle but to constrain their arguments to keep some distance from the critical value of $\omega = \pi/2$. This may result in a transition band which is wider than is desired. Another recourse is to allow the arguments of the zeros to come close to the critical value, but to reduce the values of their moduli as they do so, causing them to retreat from the unit circle.

Amongst the desiderata that affect the design of the filter is the question of the precision with which the filter coefficients and the filter output can be represented. Low-precision arithmetic can lead to the imprecise location of the poles and zeros of the filter. It can also lead to errors in the filter output which will be propagated via the feedback, and it invites the risk of numerical overflow. The precise location of the zeros that bound the transition band is a matter for experimentation. The optimal design is elusive, but satisfactory designs are easy to come by.

Having determined the width of the transition band, there remains the task of ensuring that the gain of the filter is close to zero in the stop band. According to the condition of symmetry, this will also guarantee that the gain is close to unity in the passband. The objective may be achieved with a single, carefully placed, zero. Its placement in the interval $(0, \frac{1}{2}\pi - \epsilon)$ can be governed by a formal mathematical criterion, such as the minimisation of the integral of the difference, or the squared difference, of the gain of the practical filter and that of the ideal square-wave filter. In practice, the location may be determined by interacting with a computer program that provides a visual representation of the gain of the filter.

6. Frequency Transformations

The object of the highpass filter $\beta_H(z)$ is to remove from a time series the components whose frequencies range from $\omega = 0$ to a cut-off value of $\omega = \omega_c$. The prototype version of the filter has a cut-off at the frequency $\omega = \pi/2$. In order to convert the prototype filter to one which will serve the purpose, a means must be found for mapping the frequency interval $[0, \pi/2]$ into the interval $[0, \omega_c]$. This can be achieved by replacing the argument z, wherever it occurs in the filter formula, by the argument

$$g(z) = \frac{z - \alpha}{1 - \alpha z}, \tag{26}$$

where $\alpha = \alpha(\omega_c)$ is an appropriately specified parameter.

The function $g(z)$ fulfils the following conditions:

$$\begin{aligned}
&\text{(i)} && g(z)g(z^{-1}) = 1,\\
&\text{(ii)} && g(z) = z \quad \text{if} \quad \alpha = 0,\\
&\text{(iii)} && g(1) = 1 \quad \text{and} \quad g(-1) = -1,\\
&\text{(iv)} && \text{Arg}\big\{g(z)\big\} \geq \text{Arg}\{z\} \quad \text{if} \quad \alpha > 1,\\
&\text{(v)} && \text{Arg}\big\{g(z)\big\} \leq \text{Arg}\{z\} \quad \text{if} \quad \alpha < 1.
\end{aligned} \tag{27}$$

The conditions (i) and (ii) indicate that the modulus of the function is invariably unity. Thus, as z encircles the origin, $g = g(z)$ travels around the unit circle. The conditions of (iii) indicate that, if $z = e^{i\omega}$ travels around the unit circle, then g and z will coincide when $\omega = 0$ and when $\omega = \pi$—which are the values which bound the positive frequency range over which the transfer function of the filter is defined. Finally, conditions (iv) and (v) indicate that, if $g \neq z$, then g either leads or lags behind z uniformly as the two travel around the unit circle from $z = 1$ to $z = -1$.

The value of α is completely determined by any pair of corresponding values for g and z. Thus, from (26), it follows that

$$\begin{aligned}
\alpha &= \frac{z - g}{1 - gz}\\
&= \frac{g^{1/2}z^{-1/2} - g^{-1/2}z^{1/2}}{g^{1/2}z^{1/2} - g^{-1/2}z^{-1/2}}.
\end{aligned} \tag{28}$$

Imagine that the cut-off of a prototype filter is at $\omega = \theta$ and that it is desired to shift it to $\omega = \kappa$. Then $z = e^{i\theta}$ and $g = e^{i\kappa}$ will be corresponding values; and the appropriate way of shifting the frequency would be to replace the argument z within the filter formula by the function $g(z)$ wherein the parameter α is specified by

$$\begin{aligned}
\alpha &= \frac{e^{i(\kappa-\theta)/2} - e^{-i(\kappa-\theta)/2}}{e^{i(\kappa+\theta)/2} - e^{-i(\kappa+\theta)/2}}\\
&= \frac{\sin\{(\kappa - \theta)/2\}}{\sin\{(\kappa + \theta)/2\}}.
\end{aligned} \tag{29}$$

In general, the application of the frequency transformation to a filter with p poles and q zeros will result in filter $r = \max(p, q)$ poles and zeros. In the case of a Wiener–Kolmogorov filter with p poles and p zeros, the frequency transformation will leave the filter orders unchanged. It follows that, in this case, one can find the specification of the transformed filter by transforming each of the poles and zeros in isolation. The filter coefficients can be found by knitting together the transformed poles and zeros.

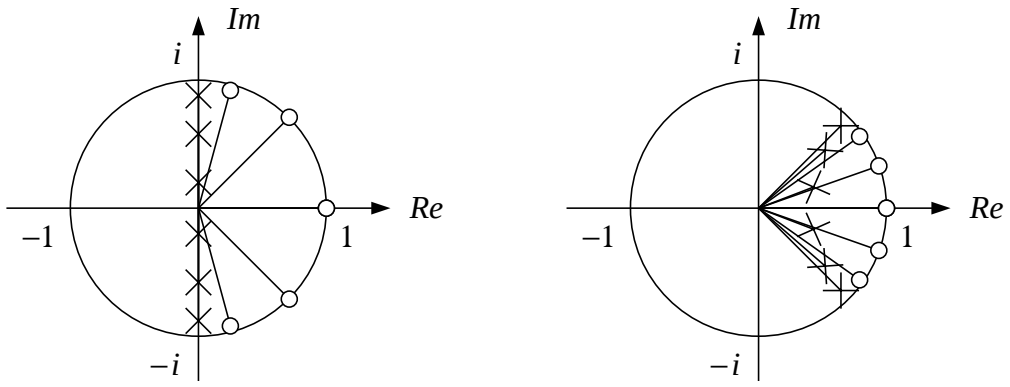

Figure 7. The pole-zero diagrams of the unidirectional highpass filters of order $r = 6$ when the cut-off point is at $\omega = \pi/2$ (left) and $\omega = \pi/4$ (right).

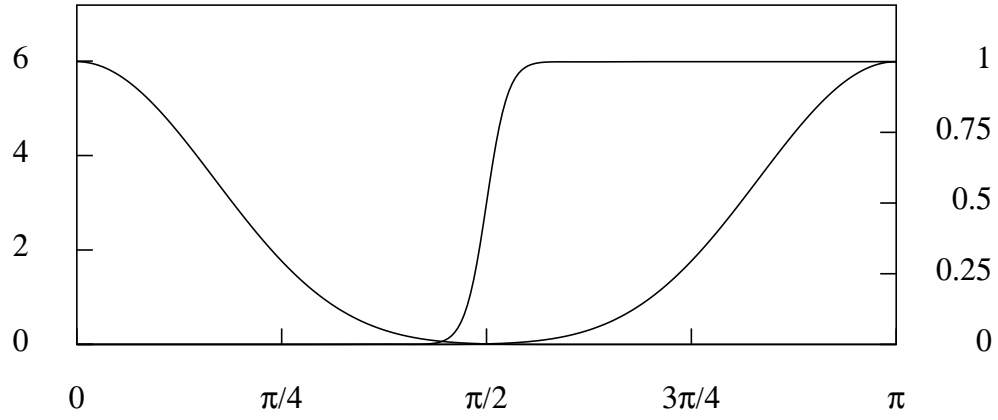

Figure 8. The gain of the bidirectional prototype filter of order $r = 6$ with a nominal cut-off point at $\omega = \pi/2$.

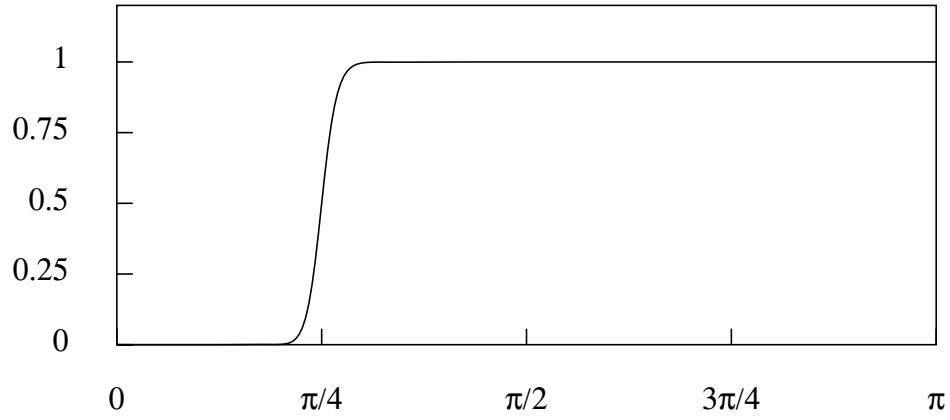

Figure 9. The gain of the bidirectional filter of order $r = 6$ with a nominal cut-off point of $\omega = \pi/4$.

Consider the generic factor within the denominator of the prototype. This is $z - \rho$, where ρ is one of the poles. Replacing z by $g(z)$ and setting the result to zero gives the following condition:

$$\frac{z - \alpha}{1 - \alpha z} - \rho = 0. \tag{30}$$

This indicates that the pole at $z = \rho$ will be replaced by a pole at

$$z = \frac{\alpha + \rho}{1 + \alpha\rho} = \frac{\alpha + \rho + \alpha^2\rho^* - \alpha\rho\rho^*}{1 + \alpha(\rho + \rho^*) + \alpha^2\rho\rho^*}. \tag{31}$$

The final expression, which comes from multiplying top and bottom of its predecessor by $1 + \alpha\rho^*$, where ρ^* is the conjugate of ρ, has a real valued denominator.

Figure 7, displays the pole-zero diagram of the unidirectional prototype filter and of a filter with a cut-off frequency of $\pi/4$, obtained by applying a frequency transformation to the prototype. The prototype filter has two zeros located on the unit circle where it intersects the positive horizontal axis. The remaining roots, which are in conjugate pairs, are located on radii at 45 degrees and 75 degrees from the horizontal. Those at 45 degrees lie on the unit circle whereas those at 75 degrees are at a distance of 0.95 from the origin. Figures 8 and 9 show the gain functions of the filters.

The comparison of the two filters suggests that one of the effects of applying a frequency transformation to the prototype filter is to bring some of poles closer to the perimeter of the unit circle. The example highlights the need to check the values of the moduli of the poles of the transformed filter to confirm that they remain sufficiently remote from unity.

7. Filtering Short Nonstationary Sequences

In this section, we shall describe how the finite-sample version of the rational filter may be implemented. For this purpose, we shall adopt the heuristic assumption that the data vector y, which has T elements, is generated by a process that is described by the equation

$$\begin{aligned}(1 - L)^d y(t) &= (1 + L)^n \prod_{i=1}^{m}(1 + \lambda_i L)\nu(t) + (1 - L)^n \prod_{i=1}^{m}(1 - \lambda_i L)\varepsilon(t) \\ &= (1 - L)^d \xi(t) + (1 - L)^d \eta(t) \\ &= \zeta(t) + \kappa(t) = g(t),\end{aligned} \tag{32}$$

where $n \geq d \geq 0$. Here $\nu(t)$ and $\varepsilon(t)$ are assumed to be Gaussian white-noise processes of unit variance. The autocovariance generating functions of $\zeta(t)$ and $\eta(t)$ are

$$\begin{aligned}\gamma^{\zeta\zeta}(z) &= (1 + z^{-1})^n (1 + z)^n \prod_{i=1}^{m}(1 + \lambda_i z^{-1})(1 + \lambda_i z), \\ \gamma^{\eta\eta}(z) &= (1 - z^{-1})^{n-d} (1 - z)^{n-d} \prod_{i=1}^{m}(1 - \lambda_i z^{-1})(1 - \lambda_i z).\end{aligned} \tag{33}$$

To find the finite-sample counterpart of equation (32), we need to represent the d-th difference operator $(1 - L)^d$ in terms of a matrix. Therefore, let the identity matrix of order T be denoted by $I_T = [e_0, e_1, \ldots, e_{T-1}]$, where e_j represents a column vector with a unit in the j-th position, counting from zero, and with zeros elsewhere. Then, the finite-sample lag operator is the matrix $L_T = [e_1, \ldots, e_{T-1}, 0]$, which has units on the first subdiagonal and zeros elsewhere. This matrix is formed by deleting the leading vector of the identity matrix and by appending a zero vector to the end of the array.

The matrix which takes the d-th difference of a vector of order T is obtained from $\Delta = (I - L_T)^d = [Q_*, Q]'$. The submatrix Q', which comprises all but the first d rows of Δ, serves to find the differenced version $q = Q'y$ of the data vector y. If $d = 0$, then $Q = \Delta = I$, and so our subsequent formulations will cover both the nonstationary case and the stationary case where there are no differencing operations.

Using this notation, the differenced data can be represented by

$$\begin{aligned} Q'y &= Q'\xi + Q'\eta \\ &= \zeta + \kappa = g. \end{aligned} \tag{34}$$

Let $D(\zeta) = \Gamma_\zeta$ and $D(\eta) = \Gamma_\eta$ be the variance–covariance matrices of the vectors ζ and η respectively for which the generating functions are given by (33). Then, to signify that the vectors have normal distributions, we may write

$$Q'\xi = \zeta \sim N(0, \Gamma_\zeta) \qquad \text{and} \qquad \eta \sim N(0, \Gamma_\eta); \tag{35}$$

and the joint probability density function of these statistically independent vectors will be given by

$$N(\zeta, \eta) = (2\pi)^{-(2T-d)/2} |\Gamma_\zeta \Gamma_\eta|^{-1/2} \exp\Big\{ -\frac{1}{2}(\xi' Q \Gamma_\zeta^{-1} Q'\xi + \eta' \Gamma_\eta^{-1} \eta) \Big\}. \tag{36}$$

The maximum-likelihood estimate x of the signal vector ξ can be found by minimising the following function, which is obtained from the exponent of the density function by setting $\eta = y - \xi$:

$$S(\xi) = \xi' Q \Gamma_\zeta^{-1} Q'\xi + (y - \xi)' \Gamma_\eta^{-1} (y - \xi). \tag{37}$$

The minimising value is

$$x = (Q \Gamma_\zeta^{-1} Q' + \Gamma_\eta^{-1})^{-1} \Gamma_\eta^{-1} y. \tag{38}$$

The matrix inversion lemma, which has been expounded, for example, by Rao (1973, p. 33) and by Pollock (1999, p. 228), indicates that

$$(Q \Gamma_\zeta^{-1} Q' + \Gamma_\eta^{-1})^{-1} = \Gamma_\eta - \Gamma_\eta Q (Q' \Gamma_\eta Q + \Gamma_\zeta)^{-1} Q' \Gamma_\eta; \tag{39}$$

and it follows that

$$\begin{aligned} x &= \left\{I - \Gamma_\eta Q(Q'\Gamma_\eta Q + \Gamma_\zeta)^{-1}Q'\right\} y \\ &= y - h, \end{aligned} \tag{40}$$

where h is the maximum-likelihood estimate of η.

The task of computing h can be accomplished by a handful of direct multiplications and recursions. Consider

$$\begin{aligned} h &= \Gamma_\eta Q(Q'\Gamma_\eta Q + \Gamma_\zeta)^{-1} g \\ &= \Gamma_\eta Q b. \end{aligned} \tag{41}$$

The first task is to calculate b by solving the equation

$$(Q'\Gamma_\eta Q + \Gamma_\zeta)b = g. \tag{42}$$

The solution is found via a Cholesky decomposition which sets $Q'\Gamma_\eta Q + \Gamma_\zeta = GG'$, where G is a lower-triangular matrix with $m+n$ nonzero diagonal bands, which is a small number equal to the recursive order of the filter. The system $GG'b = g$ can be cast in the form of $Gp = q$ and solved for p. Then $G'b = p$ can be solved for b. These are the recursive operations. Finding $h = \Gamma_\eta Q b$ thereafter entails only direct multiplications.

Notice that, in the case where $d = 0$ and where $Q = I$, the resulting filter matrix $B = \Gamma_\eta(\Gamma_\eta + \Gamma_\zeta)^{-1}$ has a form which is evidently related to the filter function $\beta(z) = \gamma^{\eta\eta}(z)\{\gamma^{\eta\eta}(z) + \gamma^{\zeta\zeta}(z)\}^{-1}$, which is of the form specified under (9). The Cholesky factorisation of $\Gamma_\eta + \Gamma_\zeta = GG'$ is analogous to the Cramer–Wold factorisation of $\gamma^{\eta\eta}(z) + \gamma^{\zeta\zeta}(z) = \phi(z^{-1})\phi(z)$.

We should comment briefly on the methods that are available for filtering short nonstationary sequences that rely upon the Kalman filter. The two principal methods are due to Ansley and Kohn (1985) and to de Jong (1991). In so far as its treatment of the starting-value problem is concerned, the method of Ansley and Kohn, which relies upon a prior transformation of the data, resembles the method proposed in this section. The method of de Jong involves an extension of the Kalman filter which is known as the diffuse Kalman filter.

The complexities of handling the starting-value problem by the method of de Jong are due in part to the essential nature of the Kalman filter algorithm, which is designed to absorb the data points one after another as they arrive in real time. This means that, in setting the starting value, the algorithm cannot afford to look ahead to the end of the sample. Therefore, the initial starting-value solution is inevitably subject to iterative updating as the sample unfolds.

If one has the advantage of working off-line, then it is unnecessary to proceed in this way. In that case, the start-up problem can be handled at the end of the sequence of calculations. However, as equation (41) indicates, the estimate h of the stationary component η can be found directly; and by concentrating on this component, the starting-value problem, which affects nonstationary components, can be circumvented.

Another way of avoiding the complexities of the Kalman filter whilst maintaining a high degree of computational efficiency is to adopt the method proposed by Burman (1980). The method is exploited in the TRAM0-SEATS econometric package (see Caporello, Maravall, and Sanchez 2001), which is aimed at trend estimation and deseasonalisation.

A Computer Program

The computer program that has been used in implementing the methods described in this paper, together with its code, is available from the author who may be contacted via the email address

stephen_pollock@sigmapi.u-net.com

References

Ansley, C.F., and R. Kohn, Estimation, Filtering and Smoothing in State Space Models with Incompletely Specified Initial Conditions, *The Annals of Statistics,* **13,** (1985), 1286–1316.

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

Beveridge, S. and C.R. Nelson, 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,** (1981), 151–174.

Bomhoff, E.J., *Financial Forecasting for Business and Economics*, The Dryden Press, London, (1994).

Brown, R.G. and P.Y.C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering: Second Edition,* John Wiley and Sons, New York, (1992).

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

Caporello, G., A, Maravall, and F.J. Sanchez. Program TSW Reference Manual, Banco de España, Madrid, (2001).

de Jong, P., The Diffuse Kalman Filter, *The Annals of Statistics,* **19,** (1991), 1073–1083.

Frances, H.P., *Periodicity and Stochastic Trends in Time Series,* Oxford University Press, Oxford, (1996).

Gómez, V., Three Equivalent Methods for Filtering Nonstationary Time Series, *Journal of Business and Economic Statistics,* **17,** (1999), 109–116.

Gómez, V. and A. Maravall, Seasonal Adjustment and Signal Extraction in Economic Time Series, chapter 8 in D. Peña, G.C. Tiao, and R.S. Tsay (eds.), *A Course in Time Series Analysis,* John Wiley and Sons, New York, (2001).

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

Hodrick, R., and E. Prescott, Post-war U.S. Business Cycles: An Empirical Investigation, Working Paper, Carnegie–Mellon University, Pittsburgh, Pennsylvania, (1980).

Hylleberg, S., (ed.), *Modelling Seasonality,* Oxford University Press, Oxford, (1992).

Kaiser, R. and A. Maravall, *Measuring Business Cycles in Economic Time Series,* Lecture Notes in Statistics 154, Springer-Verlag, New York, (2001).

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

Koopman, S.J., N. Shephard and J.A. Doornick, Statistical Algorithms for Models in State-space Form Using SsfPack 2.2, *Journal of Econometrics,* **2**, (1999), 113–166.

Laurie, D.P., Efficient Implementation of Wilson's Algorithm for Factorising a Self-Reciprocal Polynomial, *BIT,* **20,** (1980), 257–259.

Laurie, D.P., Cramér–Wold Factorisation, Algorithm AS 175, *Applied Statistics,* **31,** (1982), 86–90.

Maravall, A., Unobserved Components in Economic Times Series, in H. Pesaran and M. Wickens (eds.), *The Handbook of Applied Econometrics, volume 2,* Basil Blackwell, Oxford, (1995).

Morley, J.C., C.R. Nelson and E. Zivot, Why are Beveridge–Nelson and Unobserved Component Decompositions of GDP So Different?, Discussion Paper, Department of Economics, University of Washington, Seattle, (2001).

Pollock, D.S.G., Data Transformations and Detrending in Econometrics, Chapter 11 in C. Heij, H. Schumacher, B. Hanzon and K. Praagmam (eds.), *System Dynamics in Economic and Financial Models,* John Wiley and Sons, Chichester (1997).

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

Rao, C.R., *Linear Statistical Inference and its Applications,* John Wiley and Sons, New York, (1973).

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

Roberts, R.A., and C.T. Mullis, *Digital Signal Processing,* Addison Wesley, Reading, Massachusetts, (1987).

Stier, W., *Verfahren zur Analyse Saisonaler Schwankungen in Ökonomischen Zeitreihen,* Springer-Verlag, Berlin, (1980).

Wiener, N., *Extrapolation, Interpolation and Smoothing of Stationary Time Series,* MIT Technology Press and John Wiley and Sons, New York, (1950).

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

Wilson, G.T., Factorisation of the Covariance Generating Function of a Pure Moving Average Process, *SIAM Journal of Numerical Analysis,* **6,** (1969), 1–7.