Forecasting with Unobserved Components Time Series Models

Andrew Harvey
Faculty of Economics, University of Cambridge

Prepared for Handbook of Economic Forecasting

Contents

Abstract

Structural time series models are formulated in terms of components, such as trends, seasonals and cycles, that have a direct interpretation. As well as providing a framework for time series decomposition by signal extraction, they can be used for forecasting and for 'nowcasting' . The structural interpretation allows extensions to classes of models that are able to deal with various issues in multivariate series and to cope with non-Gaussian observations and nonlinear models. The statistical treatment is by the state space form and hence data irregularites such as missing observations are easily handled. Continuous time models offer further flexibility in that they can handle irregular spacing. The paper compares the forecasting performance of structural time series models with ARIMA and autoregressive models. Results are presented showing how observations in linear state space models are implicitly weighted in making forecasts and hence how autoregressive and vector error correction representations can be obtained. The use of an auxiliary series in forecasting and nowcasting is discussed. A final section compares stochastic volatility models with GARCH.

KEYWORDS: Cycles; continuous time; non-Gaussian models; state space; stochastic trend; stochastic volatility .

1 Introduction

The fundamental reason for building a time series model for forecasting is that it provides a way of weighting the data that is determined by the properties of the time series. Structural time series models (STMs) are formulated in terms of unobserved components, such as trends and cycles, that have a direct interpretation. Thus they are designed to focus on the salient features of the series and to project these into the future. They also provide a way of weighting the observations for signal extraction, so providing a description of the series; see Harvey and Koopman (2000). This article concentrates on prediction, though signal extraction at the end of the period - that is filtering - comes within our remit under the heading of 'nowcasting'. In summary an STM attempts to answer three questions: Where we have come from? Where we are now? and Where are we going ?

In an autoregression the past observations, up to a given lag, receive a weight obtained by minimising the sum of squares of one step ahead prediction errors. As such they form a good baseline for comparing models in terms of one step ahead forecasting performance. They can be applied directly to nonstationary time series, though imposing unit roots by differencing may be desirable to force the eventual forecast function to be a polynomial. The motivation for extending the class of models to allow moving average terms is one of parsimony. Long, indeed infinite, lags can be captured by a small number of parameters. The book by Box and Jenkins (1976) describes a model selection strategy for this class of autoregressive-integrated-moving average (ARIMA) processes. Linear STMs have reduced forms belonging to the ARIMA class. The issue for forecasting is whether the implicit restrictions they place on the ARIMA models help forecasting performance by ruling out models that have unattractive properties.

1.1 Historical background

Structural time series models developed from *ad hoc* forecasting procedures, the most basic of which is the exponentially weighted moving average (EWMA). The EWMA was generalised by Holt (1957) and Winters (1960). They introduced a slope component into the forecast function and allowed for seasonal effects. A somewhat different approach to generalising the EWMA was taken by Brown (1963), who set up forecasting procedures in a regression framework and adopted the method of discounted least squares. These methods became very popular with practitioners and are still widely used as they are simple and transparent.

Muth (1960) was the first to provide a rationale for the EWMA in terms of a properly specified statistical model, namely the random walk plus noise. Nerlove and Wage (1964) extended the model to include a slope term. These are the simplest examples of structural time series models. However, the technology of the sixties was such that further development along these lines was not pursued at the time. It was some time before statisticians became acquainted with the paper in the engineering literature by Schweppe (1965) which showed how a likelihood function could be evaluated from the Kalman filter *via* the prediction error decomposition. More significantly, even if this result had been known, it could not have been properly exploited because of the lack of computing power.

The most influential work on time series forecasting in the sixties was carried out by Box and Jenkins (1976). Rather than rationalising the EWMA by a structural model as Muth had done, Box and Jenkins observed that it could also be justified by a model in which the first differences of the variable followed a first-order moving average process. Similarly they noted that a rationale for the local linear trend extension proposed by Holt and Winters was given by a model in which second differences followed a second-order moving average process. A synthesis with the theory of stationary stochastic processes then led to the formulation of the class of ARIMA models, and the development of a model selection strategy. The estimation of ARIMA models proved to be a viable proposition at this time provided it was based on an approximate, rather than the exact, likelihood function.

Harrison and Stevens (1976) continued the work within the framework of structural time series models and were able to make considerable progress by exploiting the Kalman filter. Their response to the problems posed by parameter estimation was to adopt a Bayesian approach in which knowledge of certain key parameters was assumed. This led them to consider a further class of models in which the process generating the data switches between a finite number of regimes. This line of research has proved to be somewhat tangential to the main developments in the subject, although it is an important precursor to the econometric literature on regime switching.

Although the ARIMA approach to time series forecasting dominated the statistical literature in the 1970s and early 1980s, the structural approach was prevalent in control engineering. This was partly because of the engineers' familiarity with the Kalman filter which has been a fundamental algorithm in control engineering since its appearance in Kalman (1960). However, in a typical engineering situation there are fewer parameters to estimate and there may be a very large number of observations. The work carried out in engineering therefore tended to place less emphasis on maximum likelihood estimation and the development of a model selection methodology.

The potential of the Kalman filter for dealing with econometric and statistical problems began to be exploited in the 1970s, an early example being the work by Rosenberg (1973) on time-varying parameters. The subsequent development of a structural time series methodology began in the 1980s; see the books by Young (1984), Harvey (1989), West and Harrison (1989), Jones (1993) and Kitagawa and Gersch (1996). The book by Nerlove, Grether and Carvalho (1979) was an important precursor, although the authors did not use the Kalman filter to handle the unobserved components models that they fitted to various data sets.

The work carried out in the 1980s, and implemented in the STAMP package of Koopman et al (2000), concentrated primarily on linear models. In the 1990s, the rapid developments in computing power led to significant advances in non-Gaussian and nonlinear modelling. Furthermore, as Durbin and Koopman (2000) have emphasised, it brought classical and Bayesian approaches closer together because both draw on computer intensive techniques such as Markov chain Monte Carlo and importance sampling. The availability of these methods tends to favour the use of unobserved component models because of their flexibility in being able to capture the features highlighted by the theory associated with the subject matter.

1.2 Forecasting performance

Few studies deal explicitly with the matter of comparing the forecasting performance of STMs with other methods over a wide range of series. A notable exception is Andrews (1994). In his abstract, he concludes: ‘The structural approach appears to perform quite well on annual, quarterly, and monthly data, especially for long forecasting horizons and seasonal data. Of the more complex forecasting methods, structural models appear to be the most accurate.’ There are also a number of illustrations in Harvey (1989) and Harvey and Todd (1983). However, the most compelling evidence is indirect and comes from the results of the M3 forecasting competitions; the most recent of these is reported in Makridakis and Hibon (2000). They conclude (on p 460) as follows: ‘This competition has confirmed the original conclusions of M-competition using a new and much enlarged data set. In addition, it has demonstrated, once more, that simple methods developed by practicing forecasters (e.g., Brown’s Simple and Gardner’s Dampen (sic) Trend Exponential Smoothing) do as well, or in many cases better, than statistically sophisticated ones like ARIMA and ARARMA models’. Although Andrews seems to class structural models as complex, the fact is that they include most of the simple methods as special cases. The apparent complexity comes about because estimation is (explicitly) done by maximum likelihood and diagnostic checks are performed.

Although the links between exponential smoothing methods and STMs have been known for a long time, and were stressed in Harvey (1984, 1989), this point has not always been appreciated in the forecasting literature. Section 2 of this article sets out the models for EWMA, double exponential smoothing and damped trend exponential smoothing. Structural time series models therefore give many ad hoc procedures a firm theoretical underpinning. This is reinforced by a careful look at the so-called ‘theta method’, a new technique, introduced recently by Assimakopoulos and Nikolopoulos (2000), which did rather well in the last M3 competition. Makridakis and Hibon (2000, p 460) concluded that: ‘Although this method seems simple to use....and is not based on strong statistical theory, it performs remarkably well across different types of series, forecasting horizons and accuracy measures’. However, Hyndman and Billah (2003) show that the underlying model is just a random walk with drift plus noise. Hence it is easily handled by a program such as STAMP and there is no need to delve into the details of a method the description of which is, in the opinion of Hyndman and Billah (2003, p 287), ‘complicated, potentially confusing and involves several pages of algebra’.

1.3 State space and beyond

The state space form (SSF) allows a general treatment of virtually any linear time series models through the general algorithms of the Kalman filter and the associated smoother. Furthermore it permits the likelihood function to be computed. Section 6 reviews the SSF and presents some results that may not be well known but are relevant for forecasting. In particular it gives the ARIMA and AR representations of models in SSF. For multivariate series this leads to a method of computing the VECM representation of an unobserved component model with common trends.

The most striking benefits of the structural approach to time series modelling only become apparent when we start to consider more complex problems. The direct interpretation of the components allows parsimonious multivariate models to be set up and considerable insight can be obtained into the value of, for example, using auxiliary series to improve the efficiency of forecasting a target series. Furthermore the SSF offers enormous flexibility with regard to dealing with data irregularities, such as missing observations and observations at mixed frequencies. The study by Harvey and Chung (2000) on the measurement of British unemployment provides a nice illustration of how STMs are able to deal with forecasting and nowcasting when the series are subject to data irregularities. The challenge is how to obtain timely estimates of the underlying change in unemployment. Estimates of the numbers of unemployed according to the ILO definition have been published on a quarterly basis since the spring of 1992. From 1984 to 1991 estimates were published for the spring quarter only. The estimates are obtained from the Labour Force Survey (LFS), which consists of a rotating sample of approximately 60,000 households. Another measure of unemployment, based on administrative sources, is the number of people claiming unemployment benefit. This measure, known as the claimant count (CC), is available monthly, with very little delay and is an exact figure. It does not provide a measure corresponding to the ILO definition, but as figure 1 shows it moves roughly in the same way as the LFS figure. There are thus two issues to be addressed. The first is how to extract the best estimate of the underlying monthly change in a series which is subject to sampling error and which may not have been recorded every month. The second is how to use a related series to improve this estimate. These two issues are of general importance, for example in the measurement of the underlying rate of inflation or the way in which monthly figures on industrial production might be used to produce more timely estimates of national income. The STMs constructed by Harvey and Chung (2000) follow Pfeffermann (1991) in making use of the SSF to handle the rather complicated error structure coming from the rotating sample. Using CC as an auxiliary series halves the RMSE of the estimator of the underlying change in unemployment.

STMs can also be formulated in continuous time. This has a number of advantages, one of which is to allow irregularly spaced observations to be handled. The SSF is easily adapted to cope with this situation. Continuous time modelling of flow variables offers the possibility of certain extensions such as

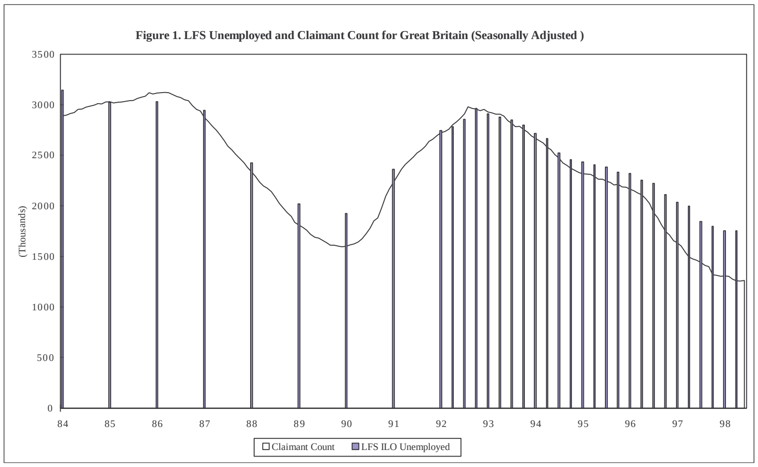

Figure 1: Annual and quarterly observations from the British labour force survey and the monthly claimant count

making cumulative predictions over a variable lead time.

Some of the most exciting recent developments in time series have been in nonlinear and non-Gaussian models. The final part of this survey provides an introduction to some of the models that can now be handled. Most of the emphasis is on what can be achieved by computer intensive methods. For example, it is possible to fit STMs with heavy-tailed distributions on the disturbances, thereby making them robust with respect to outliers and structural breaks. Similarly, non-Gaussian model with stochastic components can be set up. However, for modelling an evolving mean of a distribution for count data or qualitative observations, it is interesting that the use of conjugate filters leads to simple forecasting procedures based around the EWMA.

2 Structural time series models

The simplest structural time series models are made up of a *stochastic trend* component, μ_t, and a random irregular term. The stochastic trend evolves over time and the practical implication of this is that past observations are discounted when forecasts are made. Other components may be added. In particular a cycle is often appropriate for economic data. Again this is stochastic, thereby giving the flexibility needed to capture the type of movements that occur in practice. The statistical formulations of trends and cycles are described in the sub-sections below. A convergence component is also considered and it is shown how the model may be extended to include explanatory variables and interventions. Seasonality is discussed in a later section. The general statistical treatment is by the state space form described in section 6.

2.1 Exponential smoothing

Suppose that we wish to estimate the current level of a series of observations. The simplest way to do this is to use the sample mean. However, if the purpose of estimating the level is to use this as the basis for forecasting future observations, it is more appealing to put more weight on the most recent observations. Thus the estimate of the *current* level of the series is taken to be

$$m_T = \sum_{j=0}^{T-1} w_j y_{T-j} \tag{1}$$

where the w_j's are a set of weights that sum to unity. This estimate is then taken to be the forecast of future observations, that is

$$\hat{y}_{T+l|T} = m_T, \quad l = 1, 2, ... \tag{2}$$

so the *forecast function* is a horizontal straight line. One way of putting more weight on the most recent observations is to let the weights decline exponentially. Thus

$$m_T = \lambda \sum_{j=0}^{T-1} (1-\lambda)^j \, y_{T-j} \tag{3}$$

where λ is a *smoothing constant* in the range $0 < \lambda \leqslant 1$. The attraction of exponential weighting is that estimates can be updated by a simple recursion. If expression (3) is defined for any value of t from $t = 1$ to T, it can be split into two parts to give

$$m_t = (1-\lambda)\, m_{t-1} + \lambda y_t, \qquad t = 1, ..., T \tag{4}$$

with $m_0 = 0$. Since m_t is the forecast of y_{t+1}, the recursion is often written with $\hat{y}_{t+1|t}$ replacing m_t so that next period's forecast is a weighted average of the current observation and the forecast of the current observation made in the previous time period.

This method of constructing and updating forecasts of a level is known as an *exponentially weighted moving average* (EWMA) or *simple exponential smoothing*. The *smoothing constant*, λ, can be chosen so as to minimise the sum of squares of the one-step-ahead forecast errors, that is $S(\lambda) = \sum \hat{v}_t^2$.

The EWMA is also obtained if we take as our starting point the idea that we want to form an estimate of the mean by minimising a discounted sum of squares. Thus m_T is chosen by minimising $S(\omega) = \sum \omega^j (y_{T-j} - m_T)^2$ where $0 < \omega \leq 1$. It is easily established that $\omega = 1 - \lambda$.

The forecast function for the EWMA procedure is a horizontal straight line. Bringing a slope, b_T, into the forecast function gives

$$\hat{y}_{T+l|T} = m_T + b_T l, \quad l = 1, 2, ... \tag{5}$$

Holt (1957) and Winters (1960) introduced an updating scheme for calculating m_T and b_T in which past observations are discounted by means of two smoothing constants, λ_0 and λ_1, in the range $0 < \lambda_0, \lambda_1 < 1$. Let m_{t-1} and b_{t-1} denote the estimates of the level and slope at time $t-1$. The one-step-ahead forecast is then

$$\hat{y}_{t|t-1} = m_{t-1} + b_{t-1} \tag{6}$$

As in the EWMA, the updated estimate of the level, m_t, is a linear combination of $\hat{y}_{t|t-1}$ and y_t. Thus

$$m_t = \lambda_0 y_t + (1 - \lambda_0)(m_{t-1} + b_{t-1}) \tag{7}$$

From this new estimate of m_t, an estimate of the slope can be constructed as $m_t - m_{t-1}$ and this is combined with the estimate in the previous period to give

$$b_t = \lambda_1 (m_t - m_{t-1}) + (1 - \lambda_1) b_{t-1} \tag{8}$$

Together these equations form Holt's recursions. Following the argument given for the EWMA, starting values may be constructed from the initial observations as $m_2 = y_2$ and $b_2 = y_2 - y_1$. Hence the recursions run from $t = 3$ to $t = T$. The closer λ_0 is to zero, the less past observations are discounted in forming a current estimate of the level. Similarly, the closer λ_1 is to zero, the less they are discounted in estimating the slope. As with the EWMA, these smoothing constants can be fixed *a priori* or estimated by minimising the sum of squares of forecast errors.

2.2 Local level model

The local level model consists of a random walk plus noise,

$$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim NID\left(0, \sigma_\varepsilon^2\right), \quad t = 1, ..., T \tag{9}$$

$$\mu_t \quad = \quad \mu_{t-1} + \eta_t, \qquad \eta_t \sim NID(0, \sigma_\eta^2), \tag{10}$$

where the irregular and level disturbances, ε_t and η_t, respectively, are mutually independent and the notation $NID\left(0, \sigma^2\right)$ denotes normally and independently distributed with mean zero and variance σ^2. When σ_η^2 is zero, the level is constant. The signal-noise ratio, $q = \sigma_\eta^2 / \sigma_\varepsilon^2$, plays the key role in determining how observations should be weighted for prediction and signal extraction. The higher is q, the more past observations are discounted in forecasting the future.

Suppose that we know the mean and variance of μ_{t-1} conditional on observations up to and including time $t-1$, that is $\mu_{t-1} \mid Y_{t-1} \sim N(m_{t-1}, p_{t-1})$. Then, from (10), $\mu_t \mid Y_{t-1} \sim N(m_{t-1}, p_{t-1} + \sigma_\eta^2)$. Furthermore $y_t \mid Y_{t-1} \sim N(m_{t-1}, p_{t-1} + \sigma_\eta^2 + \sigma_\varepsilon^2)$. The information in y_t can be taken on board by a standard update of a prior distribution giving the conditional distribution at time t as $\mu_t \mid Y_t \sim N(m_t, p_t)$, where

$$m_t = m_{t-1} + [(p_{t-1} + \sigma_\eta^2) / \left(p_{t-1} + \sigma_\eta^2 + \sigma_\varepsilon^2\right)](y_t - m_{t-1}) \tag{11}$$

and

$$p_t = p_{t-1} + \sigma_\eta^2 - \left[(p_{t-1} + \sigma_\eta^2)^2 / \left(p_{t-1} + \sigma_\eta^2 + \sigma_\varepsilon^2\right)\right] \tag{12}$$

This process can be repeated as new observations become available. As we will see later this is a special case of the Kalman filter. But how should the filter be started? One possibility is to let $m_1 = y_1$, in which case $p_1 = \sigma_\varepsilon^2$. Another possibility is a diffuse prior in which the lack of information at the beginning of the series is reflected in an infinite value of p_0. However, if we set $\mu_0 \sim N(0, \kappa)$, update to get the mean and variance of μ_1 given y_1 and let $\kappa \to \infty$, the result is exactly the same as the first suggestion.

If the updating is applied repeatedly, then p_t becomes time invariant, that is $p_t \to p$. If we define $p_t^* = \sigma_\varepsilon^{-2} p_t$, divide both sides of (12) by σ_ε^2 and set $p_t^* = p_{t-1}^* = p^*$ we obtain

$$p^* = \left(-q + \sqrt{q^2 + 4q}\right)/2 \tag{13}$$

and it is clear that (11) leads to the EWMA with[1]

$$\lambda = (p^* + q)/(p^* + q + 1) = \left[\left(q^2 + 4q\right)^{1/2} - q\right]/2 \tag{14}$$

Although the discussion has so far been in terms of conditional distributions, the Gaussian assumption means that the conditional mean, m_t, is the minimum mean square error estimate (MMSE) of μ_t. As such its MSE is p_t; this does not depend on the observations and so it is the unconditional as well as the conditional MSE. Hence the updating recursions produce an estimat*or* of μ_t, which because it is a linear combination of the observations, we write as m_t. If the normality assumption is dropped, m_t is the minimum mean square error linear estimator (MMSLE). Similarly $\tilde{y}_{t|t-1} = m_{t-1}$ is the MMS(L)E of y_t.

As regards multi-step prediction at the end of the sample, the conditional distribution of y_{T+l} is easily obtained by writing

$$y_{T+l} = \mu_T + \sum_{j=1}^{l} \eta_{T+j} + \varepsilon_{T+l} = m_T + (\mu_T - m_T) + \sum_{j=1}^{l} \eta_{T+j} + \varepsilon_{T+l}.$$

Thus the MMSE predictor is $\tilde{y}_{T+l|T} = m_T, \quad l = 1, 2, ...$ and so the forecast function is a horizontal straight line which passes through the final estimator of the level. The forecast MSE, the conditional variance of y_{T+l}, is

$$MSE\left(\tilde{y}_{T+l|T}\right) = p_T + l\sigma_\eta^2 + \sigma_\varepsilon^2 = \sigma_\varepsilon^2(p_T^* + lq + 1), \quad l = 1, 2, . \tag{15}$$

This increases linearly with the forecast horizon, with p_T being the price paid for not knowing the starting point, μ_T. If T is reasonably large, then $p_T \simeq p$. Assuming σ_η^2 and σ_ε^2 to be known, a 95% prediction interval for y_{T+l} is given by $\tilde{y}_{T+l|T} \pm 1.96\sigma_{T+l|T}$ where $\sigma_{T+l|T}^2 = MSE(\tilde{y}_{T+l|T}) = \sigma_\varepsilon^2 p_{T+l|T}$.

[1] If $q = 0$, then $\lambda = 0$ so there is no updating if we switch to the steady-state filter or use the EWMA.

When a series has been subject to an instantaneous data transformation, the conditional distribution of a future value of the original series, $y^{\dagger}_{T+l}$, will no longer be normal. If logarithms have been taken, the MMSE is given by the mean of the conditional distribution of $y^{\dagger}_{T+l}$ which, being lognormal, yields

$$E\left(y^{\dagger}_{T+l} \mid Y_T\right) = \exp\left(\tilde{y}_{T+l|T} + 0.5\tilde{\sigma}^2_{T+l|T}\right), \quad l = 1, 2, ... \tag{16}$$

where $\tilde{\sigma}^2_{T+l|T} = \sigma^2_{\varepsilon} p_{T+l|T}$ is the conditional variance. A 95% prediction interval for $y^{\dagger}_{T+l}$, on the other hand, is straightforwardly computed as

$$\exp\left(\tilde{y}_{T+l|T} - 1.96\tilde{\sigma}^2_{T+l|T}\right) \leqslant y^{\dagger}_{T+l} \leqslant \exp\left(\tilde{y}_{T+l|T} + 1.96\tilde{\sigma}^2_{T+l|T}\right)$$

The model also provides the basis for using all the observations in the sample to calculate a MMSE of μ_t at all points in time. If μ_t is near the middle of a large sample then it turns out that

$$m_{t|T} \backsimeq \frac{\lambda}{2-\lambda} \sum_j (1-\lambda)^{|j|} y_{t+j}$$

Thus there is exponential weighting on either side with a higher q meaning that the closest observations receive a higher weight. This is signal extraction. A full discussion would go beyond the remit of this survey.

As regards estimation of q, the recursions deliver the mean and variance of the one-step ahead predictive distribution of each observation. Hence it is possible to construct a likelihood function in terms of the prediction errors, or *innovations*, $\nu_t = y_t - \tilde{y}_{t|t-1}$. Omce q has been estimated the innovations can be used for diagnostic checking.

2.3 Trends

The *local linear trend* model generalises the local level by introducing into (9) a stochastic slope, β_t, which itself follows a random walk. Thus

$$\begin{array}{rcll} \mu_t & = & \mu_{t-1} + \beta_{t-1} + \eta_t, & \eta_t \sim NID(0, \sigma^2_{\eta}), \\ \beta_t & = & \beta_{t-1} + \zeta_t, & \zeta_t \sim NID(0, \sigma^2_{\zeta}), \end{array} \tag{17}$$

where the irregular, level and slope disturbances, ε_t, η_t and ζ_t, respectively, are mutually independent. If both variances σ^2_{η} and σ^2_{ζ} are zero, the trend is deterministic. When only σ^2_{ζ} is zero, the slope is fixed and the trend reduces to a random walk with drift. Allowing σ^2_{ζ} to be positive, but setting σ^2_{η} to zero gives an *integrated random walk* trend, which when estimated tends to be relatively smooth. This model is often referred to as the '*smooth trend*' model.

Provided σ^2_{ζ} is strictly positive, we can generalise the argument used to obtain the local level filter and show that the recursion is as in (7) and (8) with the smoothing constants defined by

$$q_{\eta}=\left(\lambda_0^2+\lambda_0^2\lambda_1-2\lambda_0\lambda_1\right)/\left(1-\lambda_0\right) \quad and \quad q_{\zeta}=\lambda_0^2\lambda_1^2/\left(1-\lambda_0\right)$$

where q_η and q_ζ are the relative variances $\sigma_\eta^2/\sigma_\varepsilon^2$ and $\sigma_\zeta^2/\sigma_\varepsilon^2$ respectively; see Harvey (1989, ch4). Discounted least squares is a special case of double exponential smoothing; it is obtained by setting $q_\zeta=\left(q_\eta/2\right)^2$. For a smooth trend $\lambda_0=2\lambda_1/(1+\lambda_1)$.

Given the conditional means of the level and slope, that is m_T and b_T, it is not difficult to see from (17) that the forecast function for MMSE prediction is

$$\tilde{y}_{T+l|T}=m_T+b_Tl, \quad l=1,2,...$$

The *damped trend* model is a modification of (17) in which

$$\beta_t \quad = \quad \rho\beta_{t-1}+\zeta_t, \qquad \zeta_t\sim NID(0,\sigma_\zeta^2), \tag{18}$$

with $0<\rho\leq 1$. As regards forecasting

$$\widetilde{y}_{T+l|T}=m_T+b_T+\rho b_T+\cdots+\rho^{l-1}b_T=m_T+\left[\left(1-\rho^l\right)/\left(1-\rho\right)\right]b_T$$

so the final forecast function is a horizontal line at a height of $m_T+b_T/\left(1-\rho\right)$. The model could be extended by adding a constant, $\overline{\beta}$, so that

$$\beta_t \quad = \quad (1-\rho)\overline{\beta}+\rho\beta_{t-1}+\zeta_t.$$

2.4 Nowcasting

The forecast function for local linear trend starts from the current, or '*real time*', estimate of the level and increases according to the current estimate of the slope. Reporting these estimates is an example of what is sometimes called '*nowcasting*'. As with forecasting, a UC model provides a way of weighting the observations that is consistent with the properties of the series and enables MSEs to be computed.

The underlying change at the end of a series - the *growth rate* for data in logarithms - is usually the focus of attention since it is the direction in which the series is heading. It is instructive to compare model-based estimators with simple, more direct, measures. The latter have the advantage of transparency, but may entail a loss of information. For example, the first difference at the end of a series, $\Delta y_T=y_T-y_{T-1}$, may be a very poor estimator of underlying change. This is certainly the case if y_t is the logarithm of the monthly price level: its difference is the rate of inflation and this 'headline' figure is known to be very volatile. A more stable measure of change is the $r-th$ difference divided by r, that is

$$b_T^{(r)}=(1/r)\,\Delta_r y_T=(y_T-y_{T-r})/r. \tag{19}$$

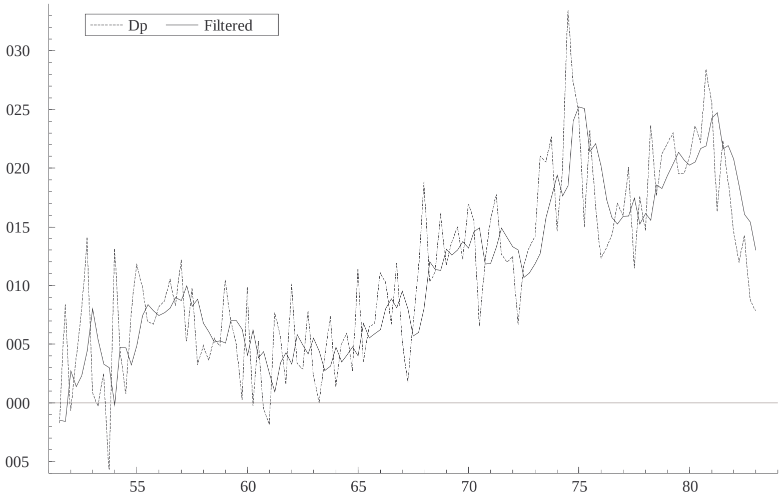

Figure 2: Quarterly rate of US inflation and filtered estimate

It is not unusual to measure the underlying monthly rate of inflation by subtracting the price level a year ago from the current price level and dividing by twelve. Note that since

$$\Delta_r y_t = \sum_{j=0}^{r-1} \Delta y_{t-j},$$

$b_T^{(r)}$ is the average of the last r first differences.

Figure 2 shows the quarterly rate of inflation in the US together with the filtered estimator obtained from a local level model with q estimated to be 0.22. At the end of the series, in the first quarter of 1983, the underlying level was 0.011, corresponding to an annual rate of 4.4%. The RMSE was one fifth of the level. The headline figure is 3.1%, but at the end of the year it was back up to 4.6%.

The effectiveness of these simple measures of change depends on the properties of the series. If the observations are assumed to come from a local linear trend model with (for algebraic convenience) the current slope in the level equation, then

$$\Delta y_t = \beta_t + \eta_t + \Delta\varepsilon_t, \qquad t = 2, \ldots T$$

and it can be seen that taking Δy_T as an estimator of current underlying change, β_T, implies a MSE of $\sigma_\eta^2 + 2\sigma_\varepsilon^2$. Further manipulation shows that the MSE of $b_T^{(r)}$ as an estimator of β_T is

$$MSE(b_T^{(r)}) = Var\left\{b_T^{(r)} - \beta_T\right\} = \frac{(r-1)\,(2r-1)}{6r}\sigma_\zeta^2 + \frac{\sigma_\eta^2}{r} + \frac{2\sigma_\varepsilon^2}{r^2} \qquad (20)$$

When $\sigma_\varepsilon^2 = 0$, the irregular component is not present and so the trend is observed directly. In this case the first differences follow a local level model and the filtered estimate $\tilde{\beta}_T$ is an EWMA of the Δy_t's. In the steady-state, $MSE\left(\tilde{\beta}_T\right)$ is as in (15) with σ_ε^2 replaced by σ_η^2 and $q = \sigma_\zeta^2/\sigma_\eta^2$. Table 1 shows some comparisons.

Table 1: RMSEs of $r-th$ differences, $b_T^{(r)}$, as estimators of underlying change, relative to RMSE of corresponding estimator from the local linear trend model

	$q = \sigma_\zeta^2/\sigma_\eta^2$			
r	0.1	0.5	1	10
1	1.92	1.41	1.27	1.04
3	1.20	1.10	1.20	2.54
12	1.27	1.92	2.41	6.20
Mean lag	2.70	1	0.62	0.09

Measures of change are sometimes based on differences of rolling (moving) averages. The rolling average, Y_t, over the previous δ time periods is

$$Y_t = (1/\delta)\sum_{j=0}^{\delta-1} y_{t-j}. \tag{21}$$

and the estimator of underlying change from $r-th$ differences is

$$B_T^{(r)} = (1/r)\,\Delta_r Y_T, \quad r = 1, 2, ... \tag{22}$$

This estimator can also be expressed as a weighted average of current and past first differences. For example, if $r = 3$, then

$$B_T^{(3)} = (1/9)\Delta y_T + (2/9)\Delta y_{T-1} + (1/3)\Delta y_{T-2} + (2/9)\Delta y_{T-3} + (1/9)\Delta y_{T-4}.$$

The series of $B_T^{(3)\prime}s$ is quite smooth but it can be slow to respond to changes. An expression for the MSE of $B_T^{(r)}$ can be obtained using the same approach as for $b_T^{(r)}$. Some comparisons of $MSEs$ can be found in Harvey and Chung (2000). As an example, in table 1 the figures for $r = 3$ for the four different values of q are 1.17, 1.35, 1.61 and 3.88.

A change in the sign of the slope may indicate a *turning point*. The $RMSE$ attached to a model-based estimate at a particular point in time gives some idea of significance. As new observations become available, the estimate and its (decreasing) $RMSE$ may be monitored by a smoothing algorithm. Planas and Rossi (2004) give an expression for the variance of the revision error - the difference between the initial filtered estimate and a subsequent smoothed estimate - and note that, because the revisions are independent of past filtered estimates, it is possible to anticipate the region within which subsequent estimates are likely to fall.

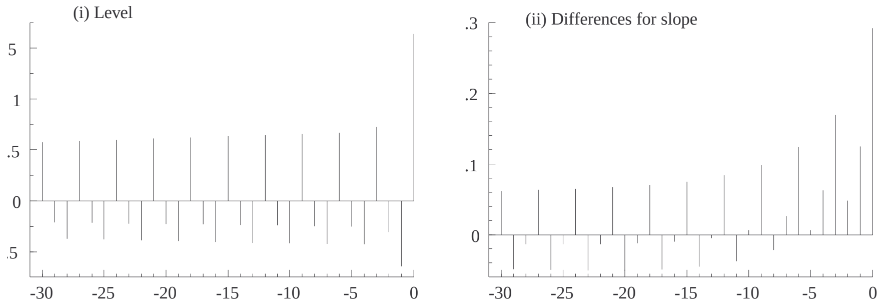

Figure 3: Weights used to construct estimates of the current level and slope of the LFS series

2.5 Surveys and measurement error

Structural time series models can be extended to take account of sample survey error from a rotational design. The statistical treatment using the state space form is not difficult; see Pfeffermann (1991). Furthermore it permits changes over time that might arise, for example, from an increase in sample size or a change in survey design.

UK Labour force survey - Harvey and Chung (2000) model quarterly LFS as a stochastic trend but with a complex error coming from the rotational survey design. The implied weighting pattern of first differences for the estimator of the underlying change, computed from the SSF by the algorithm of Koopman and Harvey (2003), is shown in figure 3 together with the weights for the level itself. It is interesting to contrast the weights for the slope with those of $B_T^{(3)}$ above.

2.6 Cycles

The stochastic cycle is

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda_c & \sin\lambda_c \\ -\sin\lambda_c & \cos\lambda_c \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix}, \quad t = 1, ..., T, \qquad (23)$$

where λ_c is frequency in radians, ρ is a damping factor and κ_t and κ_t^* are two mutually independent Gaussian white noise disturbances with zero means and common variance σ_κ^2. Given the initial conditions that the vector $(\psi_0, \psi_0^*)'$ has zero mean and covariance matrix $\sigma_\psi^2 \mathbf{I},$ it can be shown that for $0 \leq \rho < 1$, the process ψ_t is stationary and indeterministic with zero mean, variance σ_ψ^2 $= \sigma_\kappa^2/(1-\rho^2)$ and autocorrelation function (ACF)

$$\rho(\tau) = \rho^{\tau} \cos \lambda_c \tau, \quad \tau = 0, 1, 2, ... \tag{24}$$

For $0 < \lambda_c < \pi$, the spectrum of ψ_t displays a peak, centered around λ_c, which becomes sharper as ρ moves closer to one; see Harvey (1989, p60). The period corresponding to λ_c is $2\pi/\lambda_c$.

Higher order cycles have been suggested by Harvey and Trimbur (2003). The *nth order stochastic cycle*, $\psi_{n,t}$, for positive integer n, is

$$\begin{bmatrix} \psi_{1,t} \\ \psi^*_{1,t} \end{bmatrix} = \rho \begin{bmatrix} \cos \lambda_c & \sin \lambda_c \\ -\sin \lambda_c & \cos \lambda_c \end{bmatrix} \begin{bmatrix} \psi_{1,t-1} \\ \psi^*_{1,t-1} \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa^*_t \end{bmatrix}, \tag{25}$$

$$\begin{bmatrix} \psi_{i,t} \\ \psi^*_{i,t} \end{bmatrix} = \rho \begin{bmatrix} \cos \lambda_c & \sin \lambda_c \\ -\sin \lambda_c & \cos \lambda_c \end{bmatrix} \begin{bmatrix} \psi_{i,t-1} \\ \psi^*_{i,t-1} \end{bmatrix} + \begin{bmatrix} \psi_{i-1,t-1} \\ \psi^*_{i-1,t-1} \end{bmatrix}, \quad i = 2, ..., n$$

The variance of the cycle for $n = 2$ is $\sigma^2_\psi = \{(1 + \rho^2)/(1 - \rho^2)^3\}\sigma^2_\kappa$, while the ACF is

$$\rho(\tau) = \rho^{\tau} \cos(\lambda_c \tau)[1 + \{(1 - \rho^2)/(1 + \rho^2)\}\tau], \qquad \tau = 0, 1, 2, ... \tag{26}$$

The derivation and expressions for higher values of n are in Trimbur (2004).

For very short term forecasting, transitory fluctuations may be captured by a local linear trend. However, it is usual better to separate out such movements by including a stochastic cycle. Combining the components in an additive way, that is

$$y_t = \mu_t + \psi_t + \varepsilon_t, \quad t = 1, .., T, \tag{27}$$

provides the usual basis for trend-cycle decompositions. The cycle may be regarded as measuring the output gap. Extracted higher order cycles tend to be smoother with more noise consigned to the irregular.

The *cyclical trend* model incorporates the cycle into the slope by moving it from (27) to the equation for the level:

$$\mu_t \quad = \quad \mu_{t-1} + \psi_{t-1} + \beta_{t-1} + \eta_t \tag{28}$$

The damped trend is a special case corresponding to $\lambda_c = 0$.

2.7 Forecasting components

A UC model not only yields forecasts of the series itself, it also provides forecasts for the components and their MSEs.

US GDP A trend plus cycle model, (27), was fitted to the logarithm of quarterly seasonally adjusted real per capita US GDP using STAMP. illustrates how a UC can capture the salient features of a time series. Fig 4 shows the forecasts for the series itself with one $RMSE$ on either side, while figures 5 and 6 show the forecasts for the logarithms of the cycle and the trend together with their smoothed values since 1975. Figure 7 shows the annualised underlying growth rate (the estimate of the slope times four) and the fourth differences of

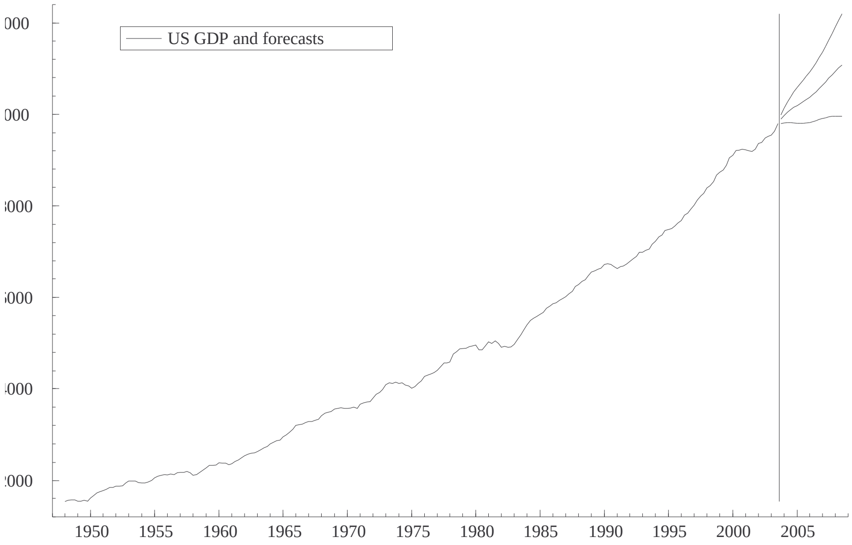

Figure 4: US GDP per capita and forecasts with 68% prediction interval

the (logarithms of the) series. The latter is fairly noisy, though much smoother than first differences, and it includes the effect of temporary growth emanating from the cycle. The growth rate from the model, on the other hand, shows the long term growth rate and indicates how the prolonged upswings of the 1960s and 1990s are assigned to the trend rather than to the cycle. (Indeed it might be interesting to consider fitting a cyclical trend model with an additive cycle). The estimate of the growth rate at the end of the series is 2.5%, with a RMSE of 1.2%, and this is the growth rate that is projected into the future.

Fitting a trend plus cycle model provides more scope for identifying turning points and assessing their significance. Different definitions of turning points might be considered, for example a change in sign of the cycle, a change in sign of its slope or a change in sign of the slope of the cycle and the trend together.

2.8 Convergence models

Long-run movements often have a tendency to converge to an equilibrium level. In an autoregressive framework this is captured by an error correction model (ECM). The UC approach is to add cycle and irregular components to an ECM so as to avoid confounding the transitional dynamics of convergence with short-term steady-state dynamics. Thus

$$y_t = \alpha + \mu_t + \psi_t + \varepsilon_t, \qquad t = 1, ..., T \tag{29}$$

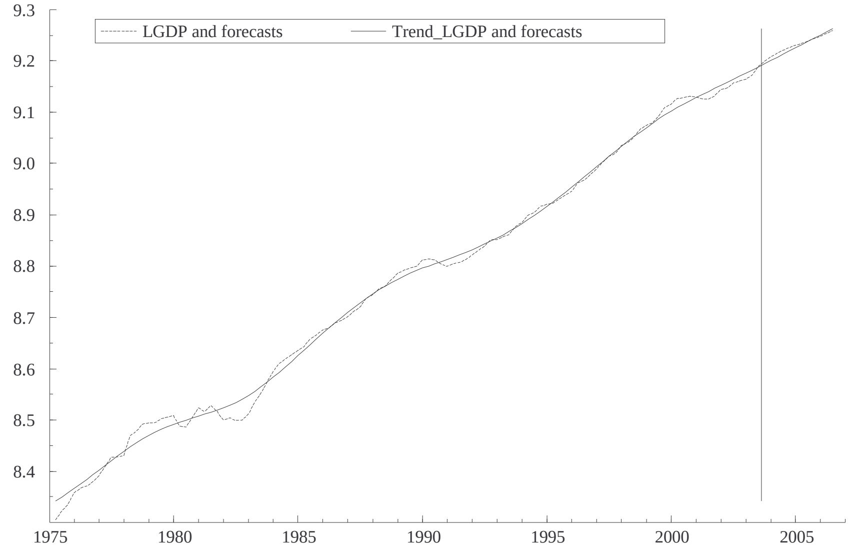

Figure 5: Trend in US GDP

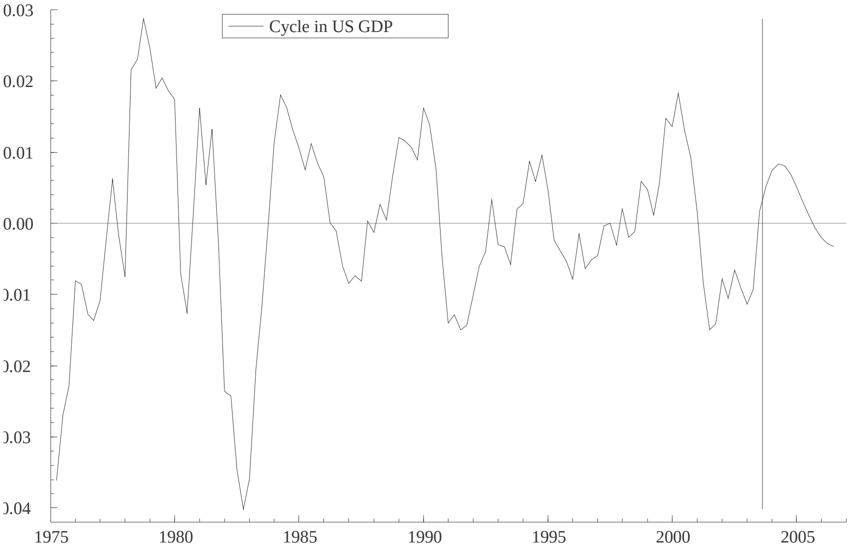

Figure 6: Cycle in US GDP

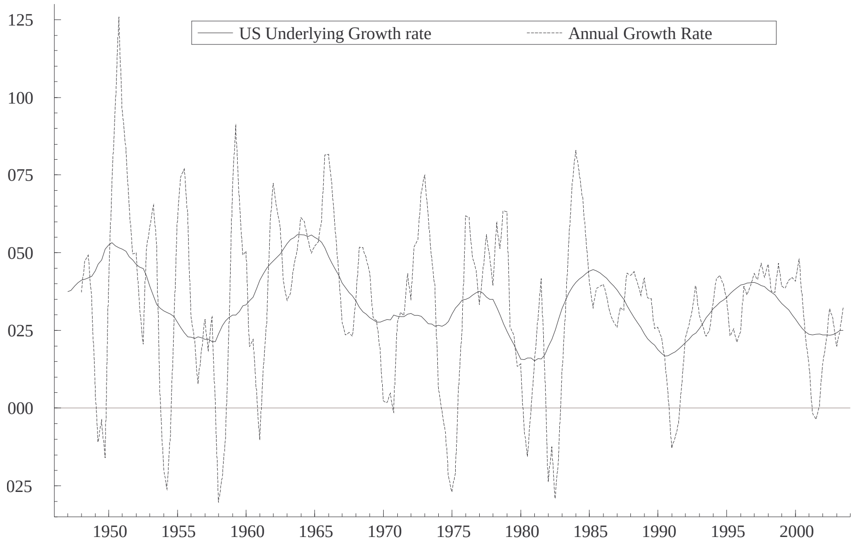

Figure 7: Smoothed estimates of slope of US per capita GDP and annual differences.

with

$$\mu_t = \phi\mu_{t-1} + \eta_t, \quad or \qquad \Delta\mu_t = (\phi - 1)\mu_{t-1} + \eta_t,$$

Smoother transitional dynamics, and hence a better separation into convergence and short-term components, can be achieved by specifying μ_t in (29) as

$$\begin{aligned} \mu_t &= \phi\mu_{t-1} + \beta_{t-1}, \quad t = 1, ..., T, \\ \beta_t &= \phi\beta_{t-1} + \zeta_t, \end{aligned} \tag{30}$$

where $0 \leq \phi \leq 1$; the smooth trend model is obtained when $\phi = 1$. This second-order ECM can be expressed as

$$\Delta\mu_t = -(1-\phi)^2\mu_{t-1} + \phi^2\Delta\mu_{t-1} + \zeta_t$$

showing that the underlying change depends not only on the gap but also on the change in the previous time period. The variance and ACF can be obtained from the properties of an AR(2) process or by noting that the model is a special case of the second order cycle with $\lambda_c = 0$.

For the smooth convergence model the $\ell-$step ahead forecast function, standardised by dividing by the current value of the gap, is $(1 + c\ell)\phi^{\ell}, \ell = 0, 1, 2, ..$ where c is a constant that depends on the ratio, λ, of the gap in the previous time period to the current one, that is $\lambda = \widetilde{\mu}_{T-1|T}/\widetilde{\mu}_{T|T}$. Since the one-step ahead forecast is $2\phi - \phi^2\lambda$, it follows that $c = 1 - \phi\lambda$, so

$$\widetilde{\mu}_{T+\ell|T} = (1 + (1 - \phi\lambda)\ell)\phi^{\ell}\widetilde{\mu}_T, \qquad \ell = 0, 1, 2, ..$$

If $\lambda = 1/\phi$, the expected convergence path is the same as in the first order model. If λ is set to $2/(1+\phi^2)$, the convergence path is the same way as the ACF. In this case, the slower convergence can be illustrated by noting, for example, that with $\phi = 0.96$, 39% of the gap can be expected to remain after 50 time periods as compared with only 13% in the first-order case. The most interesting aspect of the second-order model is that if the convergence process stalls sufficiently, the gap can be expected to widen in the short run as shown later in figure 10.

3 ARIMA and AR models

The reduced forms of the principal structural time series models are ARIMA processes. The relationship between the structural and reduced forms gives considerable insight into the potential effectiveness of different ARIMA models for forecasting and the possible shortcomings of the approach.

From the theoretical point of view, the autoregressive representation of STMs is useful in that it shows how the observations are weighted when forecasts are made. From the practical point of view it indicates the kind of series for which autoregressions are unlikely to be satisfactory.

After discussing the ways in which ARIMA and autoregressive model selection methodologies contrast with the way in which structural time series models are chosen, we examine the rationale underlying single source of error STMs.

3.1 ARIMA models and the reduced form

An *autoregressive-integrated-moving average* model of order (p, d, q) is one in which the observations follow a stationary and invertible $ARMA(p, q)$ process after they have been differenced d times. It is often denoted by writing, $y_t \sim ARIMA(p, d, q)$. If a constant term, θ_0, is included we may write

$$\Delta y_t = \theta_0 + \phi_1 \Delta y_{t-1} + \cdots + \phi_p \Delta y_{t-p} + \xi_t + \theta_1 \xi_{t-1} + \cdots + \theta_q \xi_{t-q} \tag{31}$$

where $\phi_1, ..., \phi_p$ are the autoregressive parameters, $\theta_1, ..., \theta_q$ are the moving average parameters and $\xi_t \sim NID\left(0, \sigma^2\right)$. By defining polynomials in the lag operator, L,

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p \tag{32}$$

and

$$\theta(L) = 1 - \theta_1 L + \cdots + \theta_q L^q \tag{33}$$

the model can be written more compactly as

$$\phi(L) \Delta^d y_t = \theta_0 + \theta(L) \xi_t \tag{34}$$

A structural time series model normally contains several disturbance terms. Provided the model is linear, the components driven by these disturbances can be combined to give a model with a single disturbance. This is known as the *reduced form*. The reduced form is an ARIMA model, and the fact that it is derived from a structural form will typically imply restrictions on the parameter space. If these restrictions are not imposed when an ARIMA model of the implied order is fitted, we are dealing with the *unrestricted* reduced form.

The reduced forms of the principal structural models are set out below, and the restrictions on the ARIMA parameter space explored. Expressions for the reduced form parameters may, in principle, be determined by equating the autocovariances in the structural and reduced forms. In practice this is rather complicated except in the simplest cases. An algorithm is given in Nerlove *et al.* (1979, pp. 70-78). General results for finding the reduced form for any model that can be put in state space form are given in section 6.

Local level/random walk plus noise models The reduced form is ARIMA(0,1,1). Equating the autocorrelations of first differences at lag one gives

$$\theta = \left[\left(q^2 + 4q \right)^{1/2} - 2 - q \right] /2 \tag{35}$$

where $q = \sigma_\eta^2/\sigma_\varepsilon^2$. Since $0 \leqslant q \leqslant \infty$ corresponds to $-1 \leqslant \theta \leqslant 0$, the MA parameter in the reduced form covers only half the usual parameter space. Is this a disadvantage or an advantage? The forecast function is an EWMA with $\lambda = 1+\theta$ and if θ is positive the weights alternate between positive and negative values. This may be unappealing.

Local linear trend The reduced form of the local linear trend is an ARIMA(0,2,2) process. The restrictions on the parameter space are more severe than in the case of the random walk plus noise model; see Harvey (1989, p. 69).

Cycles The cycle has an ARMA(2,1) reduced form. The MA part is subject to restrictions but the more interesting constraints are on the AR parameters. The roots of the AR polynomial are $\rho^{-1} \exp \left(\pm i\lambda_c \right)$. Thus, for $0 < \lambda_c < \pi$, they are a pair of complex conjugates with modulus ρ^{-1} and phase λ_c, and when $0 \leqslant \rho < 1$ they lie outside the unit circle. Since the roots of an AR(2) polynomial can be either real or complex, the formulation of the cyclical model effectively restricts the admissible region of the autoregressive coefficients to that part which is capable of giving rise to pseudo-cyclical behaviour. When a cycle is added to noise the reduced form is ARMA(2,2).

Models constructed from several components may have quite complex reduced forms but with strong restrictions on the parameter space. For example the reduced form of the trend plus cycle model is $ARIMA(2,2,4)$. Unrestricted estimation of high order ARIMA models may not be possible. Indeed such models are unlikely to be selected by the ARIMA methodology. In the case of US GDP, for example, $ARIMA(1,1,0)$ with drift gives a similar fit to the trend plus cycle model and hence will yield a similar one-step ahead forecasting performance; see Harvey and Jaeger (1993). The structural model may, however, forecast better several steps ahead.

3.2 Autoregressive models

The autoregressive representation may be obtained from the ARIMA reduced form or computed directly from the SSF as described in the next section. For more complex models computation from the SSF may be the only feasible option.

For the local level model, it follows from the ARIMA(0,1,1) reduced form that the first differences have a stationary autoregressive representation

$$\Delta y_t = -\sum_{j=1}^{\infty}(-\theta)^j \Delta y_{t-j} + \xi_t \tag{36}$$

Expanding the difference operator gives

$$y_t = (1+\theta)\sum_{j=1}^{\infty}(-\theta)^{j-1} y_{t-j} + \xi_t \tag{37}$$

from which it is immediately apparent that the MMSE forecast of y_t at time $t-1$ is an EWMA. If changes in the level are dominated by the irregular, the signal-noise ratio is small and θ is close to minus one. As a result the weights decline very slowly and a low order autoregression may not give a satisfactory approximation. This issue becomes more acute in a local linear trend model as the slope will typically change rather slowly. One consequence of this is that unit root tests rarely point to autoregressive models in second differences as being appropriate; see Harvey and Jaeger (1993).

3.3 Model selection in ARIMA, AR and structural time series models

An STM sets out to capture the salient features of a time series. These are often apparent from the nature of the series - an obvious example is seasonal data - though with many macroeconomic series there are strong reasons for wanting to fit a cycle. While the STM should be consistent with the correlogram, this typically plays a minor role. Indeed many models are selected without consulting it. Once a model has been chosen, diagnostic checking is carried out in the same way as for an ARIMA model.

ARIMA models are typically more parsimonious model than autoregressions. The MA terms are particularly important when differencing has taken place. Thus an ARIMA(0,1,1) is much more satisfactory than an autoregression if the true model is a random walk plus noise with a small signal-noise ratio. However, one of the drawbacks of ARIMA models as compared with STMs is that a parsimonious model may not pick up some of the more subtle features of a time series. As noted earlier, ARIMA model selection methodology will usually lead to an ARIMA(1,1,0) specification, with constant, for US GDP. For the data in sub-section 2.7, the constant term indicates a growth rate of 3.4%. This is bigger than the estimate for the structural model at the end of the series, one reason being that, as figure 7 makes clear, the long-run growth rate has been slowly declining over the last fifty years.

ARIMA model selection is based on the premise that the ACF and related statistics can be accurately estimated and are stable over time. Even if this is the case, it can be difficult to identify moderately complex models with the result that important features of the series may be missed. In practice, the sampling error associated with the correlogram may mean that even simple ARIMA models are difficult to identify, particularly in small samples. STMs are more robust as the choice of model is not dependent on correlograms. ARIMA model selection becomes even more problematic with missing observations and other data irregularities. See Durbin and Koopman (2001, pp 51-3) and Harvey (1989, pp 80-1) for further discussion.

Autoregressive models can always be fitted to time series and will usually provide a decent baseline for one-step ahead prediction. Model selection is relatively straightforward. Unit root tests are usually used to determine the degree of differencing and lags are included in the final model according to statistical significance or a goodness of fit criterion.[2] The problems with this strategy are that unit root tests often have poor size and power properties and may give a result that depends on how serial correlation is handled. Once decisions about differencing have been made, there are different views about how best to select the lags to be included. Should gaps be allowed for example? It is rarely the case that 't-statistics' fall monotonically as the lag increases, but on the other hand creating gaps is often arbitrary and is potentially distorting. Perhaps the best thing is to do is to fix the lag length according to a goodness of fit criterion, in which case autoregressive modelling is effectively nonparametric.

Tests that are implicitly concerned with the order of differencing can also be carried out in a UC framework. They are stationarity rather than unit root tests, testing the null hypothesis that a component is deterministic. The statistical theory is actually more unified with the distributions under the null hypothesis coming from the family of Cramér-von Mises distributions; see Harvey (2001).

Finally, the forecasts from an ARIMA model that satisfies the reduced form restrictions of the STM will be identical to those from the STM and will have the same MSE. For nowcasting, Box, Pierce and Newbold (1987) show how the estimators of the level and slope can be extracted from the ARIMA model. These will be the same as those obtained from the STM. However, an MSE can only be obtained for a specified decomposition.

3.4 Correlated components

Single source of error (SSOE) models are a compromise between ARIMA and STMs in that they retain the structure associated with trends, seasonals and other components while easing the restrictions on the reduced form. For example for a local level we may follow Ord *et al* (1997) in writing

$$y_t = \mu_{t-1} + \xi_t, \quad t = 1, ..., T \tag{38}$$

$$\mu_t = \mu_{t-1} + k\xi_t, \qquad \xi_t \sim NID\left(0, \sigma^2\right). \tag{39}$$

[2] With US GDP, for example, this methodology again leads to ARIMA(1,1,0).

Substituting for μ_t leads straight to an $ARIMA(0,1,1)$ model, but one in which θ is no longer constrained to take only negative values, as in (35). However, invertibility requires that k lie between zero and two, corresponding to $|\theta| < 1$. For more complex models imposing the invertibility restriction[3] may not be quite so straightforward.

As already noted, using the full invertible parameter space of the $ARIMA(0,1,1)$ model means that the weights in the EWMA can oscillate between positive and negative values. Chatfield *et al* (2001) prefer this greater flexibility, while I would argue that it can often be unappealing. The debate raises the more general issue of why UC models are usually specified to have uncorrelated components. Harvey and Koopman (2000) point out that one reason is that this produces symmetric filters for signal extraction, while in SSOE models smoothing and filtering are the same. This argument may carry less weight for forecasting. However, the MSE attached to a filtered estimate in an STM is of some value for nowcasting; in the local level model, for example, the MSE in (15) can be interpreted as the contribution to the forecast MSE that arises from not knowing the starting value for the forecast function.

In the local level model, an assumption about the correlation between the disturbances - zero or one in the local level specifications just contrasted - is needed for identifiability. This is not always the case. For example, Morley, Nelson and Zivot (2003) estimate the correlation in a model with trend and cycle components.

4 Explanatory variables and interventions

Explanatory variables can be added to unobserved components, thereby providing a bridge between regression and time series models. Thus

$$y_t = \mu_t + \mathbf{x}_t' \boldsymbol{\delta} + \varepsilon_t, \quad t = 1, ..., T \tag{40}$$

where $\mathbf{x}_t$ is a $k \times 1$ vector of observable exogenous variables, some of which may be lagged values, and $\boldsymbol{\delta}$ is a $k \times 1$ vector of parameters. In a model of this kind the trend is allowing for effects that cannot be measured. If the stochastic trend is a random walk with drift, then first differencing yields a regression model with a stationary disturbance; with a stochastic drift, second differences are needed. However, using the state space form allows the variables to remain in levels and this is a great advantage as regards interpretation; compare the transfer function models of Box and Jenkins (1976).

Spirits -The data set of annual observations on the per capita consumption of spirits in the UK, together with the explanatory variables of per capita income and relative price, is a famous one, having been used as a testbed for the Durbin-Watson statistic in 1951. The observations run from 1870 to 1938 and are in logarithms. A standard econometric approach would be to include a linear or

[3] In the STM invertibility of the reduced form is automatically ensured by the requirement that variances are not allowed to be negative.

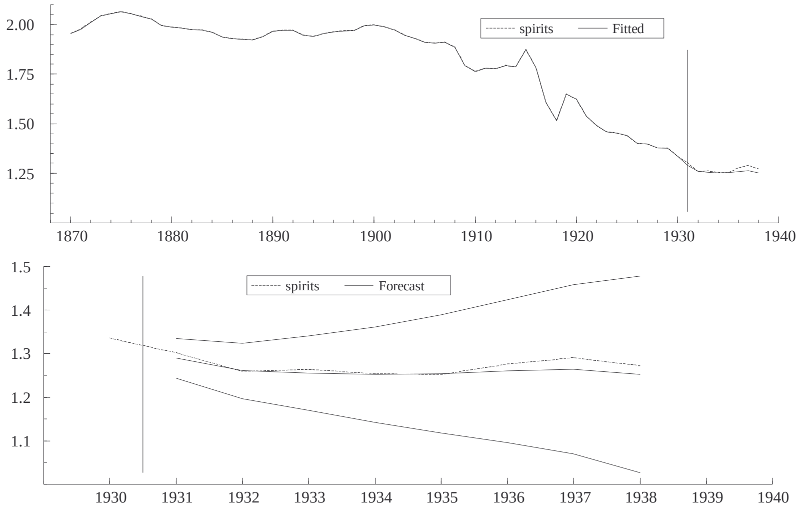

Figure 8: Mult-step forecasts for UK spirits from 1930

quadratic time trend in the model with an AR(1) disturbance; see Fuller (1996, p 522). The structural time series approach is simply to use a stochastic trend with the explanatory variables. The role of the stochastic trend is to pick up changes in tastes and habits that cannot be explicitly measured. Such a model gives a better fit and produces better forecasts. Figure 8 shows the multi-step forecasts produced from 1930 onwards, using the observed values of the explanatory variables. The lower graph shows a 68% prediction interval (± one $RMSE$). Further details on this example can be found in the STAMP manual, Koopman *et al* (2000, p64-70).

US Teenage Unemployment In a study of the relationship between teenage employment and minimum wages in the US, Bazen and Marimoutou (2002, p 699) show that a structural time series model estimated up to 1979 '...accurately predicts what happens to teenage unemployment subsequently, when the minimum wage was frozen after 1981 and then increased quite substantially in the 1990s.' They note that ..' previous models break down due to their inability to capture changes in the trend, cyclical and seasonal components of teenage employment.'

4.1 Interventions

Intervention variables may be introduced into a model. Thus in a simple stochastic trend plus error model

$$y_t = \mu_t + \lambda w_t + \varepsilon_t, \quad t = 1, ..., T \tag{41}$$

If an unusual event is to be treated as an outlier, it may be captured by a *pulse* dummy variable, that is

$$w_t = \begin{cases} 0 & \text{for } t \neq \tau \\ 1 & \text{for } t = \tau \end{cases} \tag{42}$$

A structural break in the level at time τ may be modelled by a level shift dummy,

$$w_t = \begin{cases} 0 & \text{for} \quad t < \tau \\ 1 & \text{for} \quad t \geq \tau \end{cases}$$

or by a pulse in the level equation, that is

$$\mu_t = \mu_{t-1} + \lambda w_t + \beta_{t-1} + \eta_t$$

where w_t is given by (42). Similarly a change in the slope can be modelled in (41) by defining

$$w_t = \begin{cases} 0 & \text{for} \quad t \leq \tau \\ t - \tau & \text{for} \quad t > \tau \end{cases}$$

or by putting a pulse in the equation for the slope. Note that a piecewise linear trend emerges as a special case when there are no disturbances in the level and slope equations.

Modelling structural breaks by dummy variables is appropriate when they are associated with a change in policy or a specific event. The interpretation of structural breaks as large stochastic shocks to the level or slope will prove to be a useful way of constructing a robust model when their timing is unknown; see sub-section 9.4.

4.2 Time-varying parameters

A time-varying parameter model may be set up by letting the coefficients in (40) follow random walks, that is

$$\boldsymbol{\delta}_t = \boldsymbol{\delta}_{t-1} + \boldsymbol{\upsilon}_t, \quad \boldsymbol{\upsilon}_t \sim NID(\mathbf{0}, \mathbf{Q})$$

The effect of $\mathbf{Q}$ being p.d. is to discount the past observations in estimating the latest value of the regression coefficient. Models in which the parameters evolve as stationary autoregressive processes have also been considered; see, for example, Rosenberg (1973). Chow (1984) and Nicholls and Pagan (1985) give surveys.

5 Seasonality

A seasonal component, γ_t, may be added to a model consisting of a trend and irregular to give

$$y_t = \mu_t + \gamma_t + \varepsilon_t \tag{43}$$

A fixed seasonal pattern may be modelled as

$$\gamma_t = \sum_{j=1}^{s} \gamma_j z_{jt}$$

where s is the number of seasons and the dummy variable z_{jt} is one in season j and zero otherwise. In order not to confound trend with seasonality, the coefficients, γ_j, $j = 1, ..., s$, are constrained to sum to zero. The seasonal pattern may be allowed to change over time by letting the coefficients evolve as random walks as in Harrison and Stevens (1976, pp. 217-18). If γ_{jt} denotes the effect of season j at time t, then

$$\gamma_{jt} = \gamma_{j,t-1} + \omega_{jt}, \quad \omega_t \sim NID(0, \sigma_\omega^2), \quad j = 1, ..., s \tag{44}$$

Although all s seasonal components are continually evolving, only one affects the observations at any particular point in time, that is $\gamma_t = \gamma_{jt}$ when season j is prevailing at time t. The requirement that the seasonal components evolve in such a way that they always sum to zero is enforced by the restriction that the disturbances sum to zero at each point in time. This restriction is implemented by the correlation structure in

$$Var\left(\boldsymbol{\omega}_t\right) = \sigma_\omega^2 \left(\mathbf{I} - s^{-1}\mathbf{ii}'\right) \tag{45}$$

where $\boldsymbol{\omega}_t = (\omega_{1t}, ..., \omega_{st})'$, coupled with initial condition requiring that the seasonals sum to zero at $t = 0$. It can be seen from (45) that $Var\left(\mathbf{i}'\boldsymbol{\omega}_t\right) = 0$.

In the *basic structural model* (BSM), μ_t in (43) is the local linear trend of (17), the irregular component, ε_t, is assumed to be random, and the disturbances in all three components are taken to be mutually uncorrelated. The signal noise ratio associated with the seasonal, that is $q_\omega = \sigma_\omega^2/\sigma_\varepsilon^2$, determines how rapidly the seasonal changes relative to the irregular. Figure 9 shows the forecasts, made using the STAMP package of Koopman *et al* (2000), for a quarterly series on the consumption of gas in the UK by 'Other final users'. The forecasts for the seasonal component are made by projecting the estimates of the $\gamma_{jT}'s$ into the future. As can be seen, the seasonal pattern repeats itself over a period of one year and sums to zero. Another example of how the BSM successfully captures changing seasonality can be found in the study of alcoholic beverages by Lenten and Moosa (1999).

5.1 Trigonometric seasonal

Instead of using dummy variables, a fixed seasonal pattern may by modelled by a set of trigonometric terms at the seasonal frequencies, $\lambda_j = 2\pi j/s, \quad j = 1, ..., [s/2]$, where [.] denotes rounding down to the nearest integer. The seasonal effect at time t is then

$$\gamma_t = \overset{[s/2]}{\underset{j=1}{\Sigma}} \left(\alpha_j \cos \lambda_j t + \beta_j \sin \lambda_j t\right) \tag{46}$$

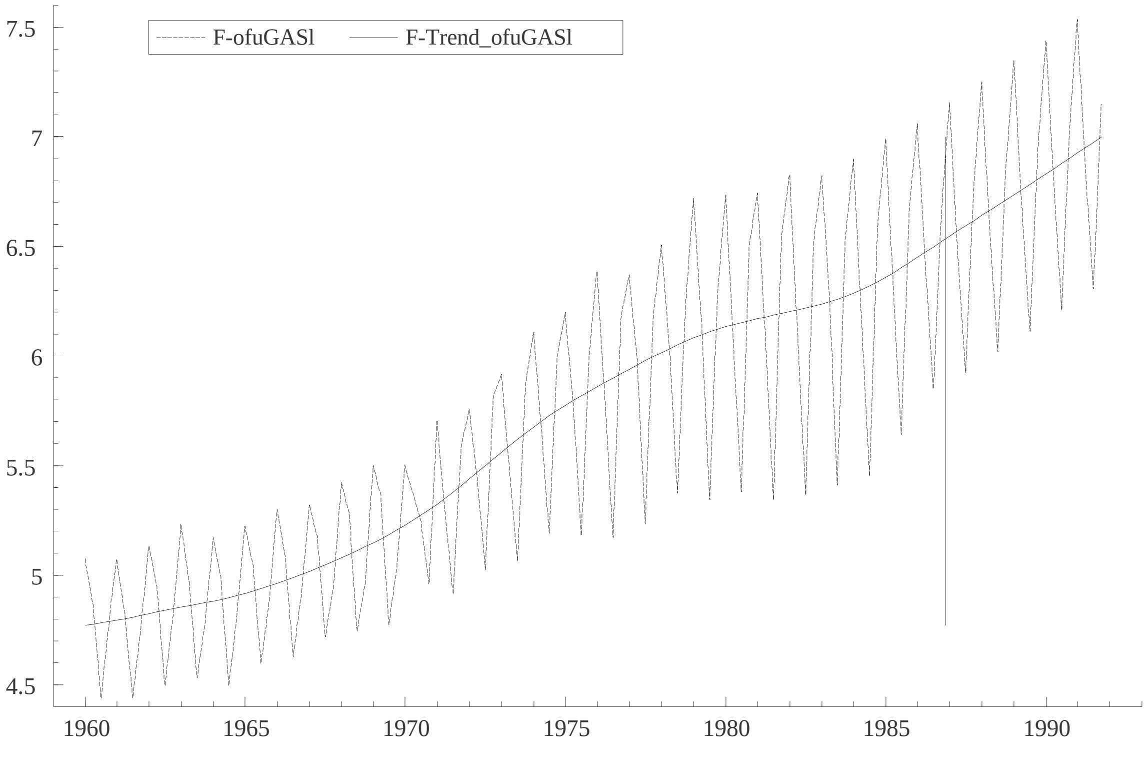

Figure 9: Trend and forecasts for 'Other final users' of gas in the UK

When s is even, the sine term disappears for $j = s/2$ and so the number of trigonometric parameters, the α_j's and β_j's, is always $s-1$. Provided that the full set of trigonometric terms is included, it is straightforward to show that the estimated seasonal pattern is the same as the one obtained with dummy variables.

The trigonometric components may be allowed to evolve over time in the same way as the stochastic cycle, (23). Thus

$$\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{jt} \tag{47}$$

with

$$\left.\begin{array}{r} \gamma_{jt} = \gamma_{j,t-1} \cos\lambda_j + \gamma^*_{j,t-1} \sin\lambda_j + \omega_{jt} \\ \gamma^*_{jt} = -\gamma_{j,t-1} \sin\lambda_j + \gamma^*_{j,t-1} \cos\lambda_j + \omega^*_{jt} \end{array}\right\}, \quad j = 1, ..., [(s-1)/2] \tag{48}$$

where ω_{jt} and ω^*_{jt} are zero mean white-noise processes which are uncorrelated with each other with a common variance σ^2_j for $j = 1, ..., [(s-1)/2]$. The larger these variances, the more past observations are discounted in estimating the seasonal pattern. When s is even, the component at $j = s/2$ reduces to

$$\gamma_t = \gamma_{j,t-1} \cos\lambda_j + \omega_{jt}, \qquad j = s/2 \tag{49}$$

The seasonal model proposed by Hannan, Terrell and Tuckwell (1970), in which α_j and β_j in (46) evolve as random walks, is effectively the same as the model above.

Assigning different variances to each harmonic allows them to evolve at varying rates. However, from a practical point of view it is usually desirable[4] to let these variances be the same except at $j = s/2$. Thus, for s even, $Var(\omega_{jt}) = Var(\omega^*_{jt}) = \sigma^2_j = \overline{\sigma}^2_\omega, \quad j = 1, ..., [(s-1)/2]$ and $Var(\omega_{s/2,t}) = \overline{\sigma}^2_\omega/2$. As shown in Proietti (2000), this is equivalent to the dummy variable seasonal model, with $\sigma^2_\omega = 2\overline{\sigma}^2_\omega/s$ for s even and $\sigma^2_\omega = 2\overline{\sigma}^2_\omega/(s-1)$ for s odd.

A damping factor could very easily be introduced into the trigonometric seasonal model, just as in (23). However, since the forecasts would gradually die down to zero, such a seasonal component is not capturing the persistent effects of seasonality. In any case the empirical evidence, for example in Canova and Hansen (1995), clearly points to nonstationary seasonality.

5.2 Reduced form

The reduced form of the stochastic seasonal model is

$$\gamma_t = -\sum_{j=1}^{s-1} \gamma_{t-j} + \omega_t \tag{50}$$

[4] As a rule, very little is lost in terms of goodness of fit by imposing this restriction. Although the model with different seasonal variances is more flexible, Bruce and Jurke (1996) show that it can lead to a significant increase in the roughness of the seasonal factors.

with ω_t following an $MA(s-2)$ process. Thus the expected value of the seasonal effects over the previous year is zero. The simplicity of a *single shock* model, in which ω_t is white noise, can be useful for pedagogic purposes. The relationship between this model and the *balanced dummy variable* model based on (44) is explored in Proietti (2000). In practice, it is usually preferable to work with the latter.

Given (50), it is easy to show that the reduced form of the BSM is such that $\Delta\Delta_s y_t \sim MA\,(s+1)\,.$

5.3 Nowcasting

When data are seasonally adjusted, revisions are needed as new observations become available and the estimates of the seasonal effects near the end of the series change. Often the revised figures are published only once a year and the changes to the adjusted figures can be quite substantial. For example, in the LFS, Harvey and Chung (2000) note that the figures for $b_T^{(3)}$ for February, March and April of 1998 were originally -6.4, 1.3 and -1.0 but using the revised data made available in early 1999 they became 7.9, 22.3 and -16.1 respectively. It appears that even moderate revisions in levels can translate into quite dramatic changes in differences, thereby rendering measures like $b_T^{(3)}$ virtually useless as a current indicator of change. Overall, the extent and timing of revisions casts doubt on the wisdom of estimating change from adjusted data, whatever the method used. Fitting models to unadjusted data has the attraction that the resulting estimates of change not only take account of seasonal movements but also reflect these movements in their RMSEs.

5.4 Holt-Winters

In the BSM the state vector is of length $s+1$, and it is not easy to obtain analytic expressions for the steady-state form of the filtering equations. On the other hand, the extension of the Holt-Winters local linear trend recursions to cope with seasonality involves only a single extra equation. However, the component for each season is only updated every s periods and an adjustment has to be made to make the seasonal factors sum to zero. Thus there is a price to be paid for having only three equations because when the Kalman filter is applied to the BSM, the seasonal components are updated in every period and they automatically sum to zero. The Holt-Winters procedure is best regarded as an approximation to the Kalman filter applied to the BSM; why anyone would continue to use it is something of a mystery. Further discussion on different forms of additive and multiplicative Holt-Winters recursions can be found in Ord, Kohler and Snyder (1997).

5.5 Seasonal ARIMA models

For modelling seasonal data, Box and Jenkins (1976, ch. 9) proposed a class of multiplicative seasonal ARIMA models; see also the chapter by Ghysels et al in this volume. The most important model within this class has subsequently become known as the 'airline model' since it was originally fitted to a monthly series on UK airline passenger totals. The model is written as

$$\Delta\Delta_s y_t = (1+\theta L)\,(1+\Theta L^s)\,\xi_t \tag{51}$$

where $\Delta_s = 1 - L^s$ is the seasonal difference operator and θ and Θ are MA parameters which, if the model is to be invertible, must have modulus less than one. Box and Jenkins (1976, pp. 305-6) gave a rationale for the airline model in terms of EWMAs at monthly and yearly intervals.

Maravall (1985), compares the autocorrelation functions of $\Delta\Delta_s y_t$ for the BSM and airline model for some typical values of the parameters and finds them to be quite similar, particularly when the seasonal MA parameter, Θ, is close to minus one. In fact in the limiting case when Θ is equal to minus one, the airline model is equivalent to a BSM in which σ_ζ^2 and σ_ω^2 are both zero. The airline model provides a good approximation to the reduced form when the slope and seasonal are close to being deterministic. If this is not the case the implicit link between the variability of the slope and that of the seasonal component may be limiting.

The plausibility of other multiplicative seasonal ARIMA models can, to a certain extent, be judged according to whether they allow a canonical decomposition into trend and seasonal components; see Hillmer and Tiao (1982). Although a number of models fall into this category the case for using them is unconvincing. It is hardly surprising that most procedures for ARIMA model-based seasonal adjustment are based on the airline model. However, although the airline model may often be perfectly adequate as a vehicle for seasonal adjustment, it is of limited value for forecasting many economic time series. For example, it cannot deal with business cycle effects.

Pure AR models can be very poor at dealing with seasonality since seasonal patterns typically change rather slowly and this may necessitate the use of long seasonal lags. However, it is possible to combine an autoregression with a stochastic seasonal component as in Harvey and Scott (1994).

Consumption A model for aggregate consumption provides a nice illustration of the way in which a simple parsimonious STM that satisfies economic considerations can be constructed. Using UK data from 1957q3 to 1992q2, Harvey and Scott (1994) show that a special case of the BSM consisting of a random walk plus drift, β, and a stochastic seasonal not only fits the data but yields a seasonal martingale difference that does little violence to the forward-looking theory of consumption. The unsatisfactory nature of an autoregression is illustrated in the paper by Osborn and Smith (1989) where sixteen lags are required to model seasonal differences. As regards ARIMA models, Osborn and Smith (1989) select a special case of the airline model in which $\theta = 0$. This contrasts with the reduced form for the structural model which has $\Delta_s c_t$ following an $MA(s-1)$ process (with non-zero mean). The seasonal ARIMA model matches the ACF but does not yield forecasts satisfying a seasonal martingale, that is $E[\Delta_s c_{t+s}] = s\beta$.

5.6 Extensions

It is not unusual for the level of a monthly time series to be influenced by *calendar effects.* Such effects arise because of changes in the level of activity resulting from variations in the composition of the calendar between years. The two main sources of calendar effects are trading day variation and moving festivals. They may both be introduced into a structural time series model and estimated along with the other components in the model. The state space framework allows them to change over time as in Dagum, Quenneville and Sutradhar (1992). Methods of detecting calendar effects are discussed in Busetti and Harvey (2003). As illustrated by Hillmer (1982, p. 388), failure to realise that calendar effects are present can distort the correlogram of the series and lead to inappropriate ARIMA models being chosen.

The treatment of *weekly, daily or hourly* observations raises a host of new problems. The structural approach offers a means of tackling them. Harvey, Koopman and Riani (1996) show how to deal with a weekly seasonal pattern by constructing a parsimonious but flexible model for the UK money supply based on time-varying splines and incorporating a mechanism to deal with moving festivals such as Easter. Harvey and Koopman (1993) also use time-varying splines to model and forecast hourly electricity data.

Periodic or *seasonal specific* models were originally introduced to deal with certain problems in environmental science, such as modelling river flows; see Hipel and McLeod (1994, ch. 14). The key feature of such models is that separate stationary AR or ARMA model are constructed for each season. Econometricians have developed periodic models further to allow for nonstationarity within each season and constraints across the parameters in different seasons; see Franses and Papp (2004). These approaches are very much within the autoregressive/unit root paradigm. The structural framework offers a more general way of capturing periodic features by allowing periodic components to be combined with components common to all seasons. These common components may exhibit seasonal heteroscedasticity, that is they may have different values for the parameters in different seasons. Such models have a clear interpretation and make explicit the distinction between an evolving seasonal pattern of the kind typically used in a structural time series model and genuine periodic effects. Proietti (1998) discusses these issues and gives the example of Italian industrial production where August behaves so differently from the other months that it is worth letting it have its own trend. There is further scope for work along these lines.

Krane and Wascher (1999) use state space methods to explore the interaction between *seasonality and business cycles.* They apply their methods to US employment and conclude that seasonal movements can be affected by business cycle developments.

Stochastic seasonal components can be combined with *explanatory variables* by introducing them into regression models in the same way as stochastic trends. The way in which this can give insight into the specification of dynamic regression models is illustrated in the paper by Harvey and Scott (1994) where it is suggested that seasonality in an error correction model be captured by a stochastic seasonal component. The model provides a good fit to UK consumption and casts doubt on the specification adopted in the influential paper of Davidson *et al* (1978). Moosa and Kennedy (1998) reach the same conclusion using Australian data.

6 State space form

The statistical treatment of unobserved components models can be carried out efficiently and in great generality by using the state space form (SSF) and the associated algorithms of the Kalman filter and smoother.

The general linear state space form applies to a multivariate time series, $\mathbf{y}_t$, containing N elements. These observable variables are related to an $m \times 1$ vector, $\boldsymbol{\alpha}_t$, known as the *state vector*, through a *measurement equation*

$$\mathbf{y}_t = \mathbf{Z}_t\boldsymbol{\alpha}_t + \mathbf{d}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, ..., T \tag{52}$$

where $\mathbf{Z}_t$ is an $N \times m$ matrix, $\mathbf{d}_t$ is an $N \times 1$ vector and $\boldsymbol{\varepsilon}_t$ is an $N \times 1$ vector of serially uncorrelated disturbances with mean zero and covariance matrix $\mathbf{H}_t$, that is $E(\boldsymbol{\varepsilon}_t) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}_t) = \mathbf{H}_t$.

In general the elements of $\boldsymbol{\alpha}_t$ are not observable. However, they are known to be generated by a first-order Markov process,

$$\boldsymbol{\alpha}_t = \mathbf{T}_t\boldsymbol{\alpha}_{t-1} + \mathbf{c}_t + \mathbf{R}_t\boldsymbol{\eta}_t, \quad t = 1, ..., T \tag{53}$$

where $\mathbf{T}_t$ is an $m \times m$ matrix, $\mathbf{c}_t$ is an $m \times 1$ vector, $\mathbf{R}_t$ is an $m \times g$ matrix and $\boldsymbol{\eta}_t$ is a $g \times 1$ vector of serially uncorrelated disturbances with mean zero and covariance matrix, $\mathbf{Q}_t$, that is $E(\boldsymbol{\eta}_t) = \mathbf{0}$ and $Var(\boldsymbol{\eta}_t) = \mathbf{Q}_t$. Equation (53) is the *transition equation*.

The specification of the state space system is completed by assuming that the initial state vector, $\boldsymbol{\alpha}_0$, has a mean of $\mathbf{a}_0$ and a covariance matrix $\mathbf{P}_0$, that is $E(\boldsymbol{\alpha}_0) = \mathbf{a}_0$ and $Var(\boldsymbol{\alpha}_0) = \mathbf{P}_0$, where $\mathbf{P}_0$ is positive semi-definite, and that the disturbances $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are uncorrelated with the initial state, that is $E(\boldsymbol{\varepsilon}_t\boldsymbol{\alpha}_0') = \mathbf{0}$ and $E(\boldsymbol{\eta}_t\boldsymbol{\alpha}_0') = \mathbf{0}$ for $t = 1, , ..., T$. In what follows it will be assumed that the disturbances are uncorrelated with each other in all time periods, that is $E(\boldsymbol{\varepsilon}_t\boldsymbol{\eta}_s') = \mathbf{0}$ for all $s, t = 1, ..., T$, though this assumption may be relaxed, the consequence being a slight complication in some of the filtering formulae.

It is sometimes convenient to use the future form of the transition equation,

$$\boldsymbol{\alpha}_{t+1} = \mathbf{T}_t\boldsymbol{\alpha}_t + \mathbf{c}_t + \mathbf{R}_t\boldsymbol{\eta}_t, \quad t = 1, ..., T, \tag{54}$$

as opposed to the contemporaneous form of (53). The corresponding filters are the same unless $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are correlated.

6.1 Kalman filter

The Kalman filter is a recursive procedure for computing the optimal estimator of the state vector at time t, based on the information available at time t. This information consists of the observations up to and including $\mathbf{y}_t$. The system matrices together with $\mathbf{a}_0$ and $\mathbf{P}_0$ are assumed to be known in all time periods and so do not need to be explicitly included in the information set.

In a Gaussian model, the disturbances $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$, and the initial state, are all normally distributed. Because a normal distribution is characterised by its first two moments, the Kalman filter can be interpreted as updating the mean and covariance matrix of the conditional distribution of the state vector as new observations become available. The conditional mean is the minimum mean square error estimate, and because the MSE matrix does not depend on the observations, it is unconditional and the conditional mean is also optimal in the sense that it is a minimum mean square error estimator. When the normality assumption is dropped, the Kalman filter is still optimal estimator in that it minimises the mean square error within the class of all linear estimators; see Anderson and Moore (1979, 29-32).

Consider the Gaussian state space model with observations available up to and including time $t-1$. Given this information set, let $\boldsymbol{\alpha}_{t-1}$ be normally distributed with known mean, $\mathbf{a}_{t-1}$, and $m \times m$ covariance matrix, $\mathbf{P}_{t-1}$. Then it follows from (53) that $\boldsymbol{\alpha}_t$ is normal with mean

$$\mathbf{a}_{t|t-1} = \mathbf{T}_t \mathbf{a}_{t-1} + \mathbf{c}_t \tag{55}$$

and covariance matrix

$$\mathbf{P}_{t|t-1} = \mathbf{T}_t \mathbf{P}_{t-1} \mathbf{T}_t' + \mathbf{R}_t \mathbf{Q}_t \mathbf{R}_t', \quad t = 1, ..., T$$

These two equations are known as the *prediction equations.* The predictive distribution of the next observation, $\mathbf{y}_t$, is normal with mean

$$\widetilde{\mathbf{y}}_{t|t-1} = \mathbf{Z}_t \mathbf{a}_{t|t-1} + \mathbf{d}_t \tag{56}$$

and covariance matrix

$$\mathbf{F}_t = \mathbf{Z}_t \mathbf{P}_{t|t-1} \mathbf{Z}_t' + \mathbf{H}_t, \quad t = 1, ..., T \tag{57}$$

Once the new observation becomes available, a standard result on the multivariate normal distribution yields the *updating equations,*

$$\mathbf{a}_t = \mathbf{a}_{t|t-1} + \mathbf{P}_{t|t-1} \mathbf{Z}_t' \mathbf{F}_t^{-1} (\mathbf{y}_t - \mathbf{Z}_t \mathbf{a}_{t|t-1} - \mathbf{d}_t) \tag{58}$$

and

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1} \mathbf{Z}_t' \mathbf{F}_t^{-1} \mathbf{Z}_t \mathbf{P}_{t|t-1},$$

as the mean and variance of the conditional distribution of $\boldsymbol{\alpha}_t$; see Harvey (1989, p 109).

Taken together (55) and (58) make up the Kalman filter. If desired they can be written as a single set of recursions going directly from $\mathbf{a}_{t-1}$ to $\mathbf{a}_t$ or, alternatively, from $\mathbf{a}_{t|t-1}$ to $\mathbf{a}_{t+1|t}$. We might refer to these as, respectively, the *contemporaneous* and *predictive filter.* In the latter case

$$\mathbf{a}_{t+1|t} = \mathbf{T}_{t+1}\mathbf{a}_{t|t-1} + \mathbf{c}_{t+1} + \mathbf{K}_t\boldsymbol{\nu}_t \tag{59}$$

or

$$\mathbf{a}_{t+1|t} = \left(\mathbf{T}_{t+1} - \mathbf{K}_t\mathbf{Z}_t\right)\mathbf{a}_{t|t-1} + \mathbf{K}_t\mathbf{y}_t + \left(\mathbf{c}_{t+1} - \mathbf{K}_t\mathbf{d}_t\right) \tag{60}$$

where the gain matrix, $\mathbf{K}_t$, is given by

$$\mathbf{K}_t = \mathbf{T}_{t+1}\mathbf{P}_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{-1}, \quad t = 1, ..., T \tag{61}$$

The recursion for the covariance matrix,

$$\mathbf{P}_{t+1|t} = \mathbf{T}_{t+1}(\mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t\mathbf{P}_{t|t-1})\mathbf{T}_{t+1}' + \mathbf{R}_{t+1}\mathbf{Q}_{t+1}\mathbf{R}_{t+1}', \tag{62}$$

is a *Riccati equation.*

The starting values for the Kalman filter may be specified in terms of $\mathbf{a}_0$ and $\mathbf{P}_0$ or $\mathbf{a}_{1|0}$ and $\mathbf{P}_{1|0}$. Given these initial conditions, the Kalman filter delivers the optimal estimator of the state vector as each new observation becomes available. When all T observations have been processed, the filter yields the optimal estimator of the current state vector, and/or the state vector in the next time period, based on the full information set. A diffuse prior corresponds to setting $\mathbf{P}_0 = \kappa\mathbf{I}$, and letting the scalar κ go to infinity.

6.2 Prediction

In the Gaussian model, (52) and (53), the Kalman filter yields $\mathbf{a}_T$, the MMSE of $\boldsymbol{\alpha}_T$ based on all the observations. In addition it gives $\mathbf{a}_{T+1|T}$ and the one-step-ahead predictor, $\widetilde{\mathbf{y}}_{T+1|T}$. As regards multi-step prediction, taking expectations, conditional on the information at time T, of the transition equation at time $T + \ell$ yields the recursion

$$\mathbf{a}_{T+l|T} = \mathbf{T}_{T+l}\mathbf{a}_{T+l-1|T} + \mathbf{c}_{T+l} \quad l = 1, 2, 3, ... \tag{63}$$

with initial value $\mathbf{a}_{T|T} = \mathbf{a}_T$. Similarly

$$\mathbf{P}_{T+l|T} = \mathbf{T}_{T+l}\mathbf{P}_{T+l-1|T}\mathbf{T}_{T+l}' + \mathbf{R}_{T+l}\mathbf{Q}_{T+l}\mathbf{R}_{T+l}', \quad l = 1, 2, 3, ... \tag{64}$$

with $\mathbf{P}_{T|T} = \mathbf{P}_T$. Thus $\mathbf{a}_{T+l|T}$ and $\mathbf{P}_{T+l|T}$ are evaluated by repeatedly applying the Kalman filter prediction equations. The MMSE of $\mathbf{y}_{T+l}$ can be obtained directly from $\mathbf{a}_{T+l|T}$. Taking conditional expectations in the measurement equation for $\mathbf{y}_{T+l}$ gives

$$E\left(\mathbf{y}_{T+l} \mid \mathbf{Y}_T\right) = \widetilde{\mathbf{y}}_{T+l|T} = \mathbf{Z}_{T+l}\mathbf{a}_{T+l|T} + \mathbf{d}_{T+l}, \quad l = 1, 2, ... \tag{65}$$

with MSE matrix

$$MSE\left(\widetilde{\mathbf{y}}_{T+l|T}\right) = \mathbf{Z}_{T+l}\mathbf{P}_{T+l|T}\mathbf{Z}'_{T+l} + \mathbf{H}_{T+l}, \quad l = 1, 2, ... \tag{66}$$

When the normality assumption is relaxed, $\mathbf{a}_{T+l|T}$ and $\tilde{\mathbf{y}}_{T+l|T}$ are still minimum mean square *linear* estimators.

It is often of interest to see how past observations are weighted when forecasts are constructed: Koopman and Harvey (2003) give an algorithm for computing weights for $\mathbf{a}_T$ and weights for $\widetilde{\mathbf{y}}_{T+l|T}$ are then obtained straightforwardly.

6.3 Innovations

The joint density function for the T sets of observations, $\mathbf{y}_1, ..., \mathbf{y}_T$, is

$$p\left(\mathbf{Y}; \boldsymbol{\psi}\right) = \prod_{t=1}^{T} p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}\right) \tag{67}$$

where $p\left(\mathbf{y}_t \mid \mathbf{Y}_{t-1}\right)$ denotes the distribution of $\mathbf{y}_t$ conditional on the information set at time $t-1$, that is $\mathbf{Y}_{t-1} = \left\{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_1\right\}$. In the Gaussian state space model, the conditional distribution of $\mathbf{y}_t$ is normal with mean $\widetilde{\mathbf{y}}_{t|t-1}$ and covariance matrix $\mathbf{F}_t$. Hence the $N \times 1$ vector of prediction errors or *innovations,*

$$\boldsymbol{\nu}_t = \mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1}, \quad t = 1, ..., T, \tag{68}$$

is serially independent with mean zero and covariance matrix $\mathbf{F}_t$, that is $\boldsymbol{\nu}_t \sim NID(\mathbf{0}, \mathbf{F}_t)$.

Re-arranging (68), (56) and (59) gives the *innovations form* representation[5]

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Z}_t\mathbf{a}_{t|t-1} + \mathbf{d}_t + \boldsymbol{\nu}_t \\ \mathbf{a}_{t+1|t} &= \mathbf{T}_t\mathbf{a}_{t|t-1} + \mathbf{c}_t + \mathbf{K}_t\boldsymbol{\nu}_t \end{aligned} \tag{69}$$

This mirrors the original SSF, with the transition equation as in (54), except that $\mathbf{a}_{t|t-1}$ appears in the place of the state and the disturbances in the measurement and transition equations are perfectly correlated. Since the model contains only one disturbance vector, it may be regarded as a reduced form with $\mathbf{K}_t$ subject to restrictions coming from the original structural form. The SSOE models discussed in sub-section 3.4 are effectively in innovations form but if this is the starting point of model formulation some way of putting constraints on $\mathbf{K}_t$ has to be found.

6.4 Time-invariant models

In many applications the state space model is time-invariant. In other words the system matrices $\mathbf{Z}_t, \mathbf{d}_t, \mathbf{H}_t, \mathbf{T}_t, \mathbf{c}_t, \mathbf{R}_t$ and $\mathbf{Q}_t$ are all independent of time and so can be written without a subscript. However, most of the properties in which

[5] The fact that the observations are a linear function of serially independent disturbances means that the model, as opposed to the SSF, is linear.

we are interested apply to a system in which $\mathbf{c}_t$ and $\mathbf{d}_t$ are allowed to change over time and so the class of models under discussion is effectively

$$\mathbf{y}_t = \mathbf{Z}\boldsymbol{\alpha}_t + \mathbf{d}_t + \boldsymbol{\varepsilon}_t, \quad Var\left(\boldsymbol{\varepsilon}_t\right) = \mathbf{H} \tag{70}$$

and

$$\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \mathbf{c}_t + \mathbf{R}\boldsymbol{\eta}_t, \quad Var\left(\boldsymbol{\eta}_t\right) = \mathbf{Q} \tag{71}$$

with $E\left(\boldsymbol{\varepsilon}_t\boldsymbol{\eta}_s'\right) = \mathbf{0}$ for all s, t and $\mathbf{P}_{1|0}$, $\mathbf{H}$ and $\mathbf{Q}$ p.s.d.

The principal STMS are time invariant and easily put in SSF with a measurement equation that, for univariate models, will be written

$$y_t = \mathbf{z}'\boldsymbol{\alpha}_t + \varepsilon_t, \quad t = 1, ..., T \tag{72}$$

with $Var(\varepsilon_t) = H = \sigma_\varepsilon^2$. Thus state space form of the damped trend model, (18) is:

$$y_t = \begin{bmatrix}1 & 0\end{bmatrix}\boldsymbol{\alpha}_t + \varepsilon_t \tag{73}$$

$$\boldsymbol{\alpha}_t = \begin{bmatrix}\mu_t \\ \beta_t\end{bmatrix} = \begin{bmatrix}1 & 1 \\ 0 & \rho\end{bmatrix}\begin{bmatrix}\mu_{t-1} \\ \beta_{t-1}\end{bmatrix} + \begin{bmatrix}\eta_t \\ \zeta_t\end{bmatrix} \tag{74}$$

The local linear trend is the same but with $\rho = 1$.

The Kalman filter applied to the model in (70) is in a steady state if the error covariance matrix is time-invariant, that is $\mathbf{P}_{t+1|t}= \mathbf{P}$. This implies that the covariance matrix of the innovations is also time-invariant, that is $\mathbf{F}_t = \mathbf{F} = \mathbf{ZPZ}'+\mathbf{H}$. The recursion for the error covariance matrix is therefore redundant in the steady state, while the recursion for the state becomes

$$\mathbf{a}_{t+1|t}= \mathbf{La}_{t|t-1}+\mathbf{Ky}_t + \left(\mathbf{c}_{t+1} - \mathbf{Kd}_t\right) \tag{75}$$

where the transition matrix is defined by

$$\mathbf{L} = \mathbf{T} - \mathbf{KZ} \tag{76}$$

and $\mathbf{K} = \mathbf{TPZ}'\mathbf{F}^{-1}$.

Under what conditions will the KF converge to a steady state? Letting $\mathbf{P}_{t+1|t}= \mathbf{P}_{t|t-1}= \mathbf{P}$ in (62) yields the algebraic Riccati equation

$$\mathbf{P} - \mathbf{TPT}' + \mathbf{TPZ}'\mathbf{F}^{-1}\mathbf{ZPT}'-\mathbf{RQR}'= \mathbf{0} \tag{77}$$

and the Kalman filter has a steady-state solution if there exists a time-invariant error covariance matrix, $\mathbf{P}$, that satisfies this equation. Although the solution to the Riccati equation was obtained for the local level model in (13), it is usually difficult to obtain an explicit solution. A discussion of various algorithms can be found in Ionescu, Oara and Weiss (1997). Even if it is known that a solution exists, it is not immediately apparent whether the $\mathbf{P}$ matrix will be unique or whether it will be p.s.d.

The model is stable if the roots of $\mathbf{T}$ are less than one in absolute value, that is $\left|\lambda_i\left(\mathbf{T}\right)\right| < 1, i = 1, ..., m$ and it can be shown that

$$\lim_{t\to\infty}\mathbf{P}_{t+1|t} = \mathbf{P} \tag{78}$$

with $\mathbf{P}$ independent of $\mathbf{P}_{1|0}$. Convergence to $\mathbf{P}$ is exponentially fast provided that $\mathbf{P}$ is the only p.s.d. matrix satisfying the algebraic Riccati equation. Note that with $\mathbf{d}_t$ time invariant and $\mathbf{c}_t$ zero the model is stationary. The stability condition can be readily checked but it is stronger than is necessary. It is apparent from (75) that what is needed is $|\lambda_i(\mathbf{L})| < 1, \quad i = 1, ..., m$, but, of course, $\mathbf{L}$ depends on $\mathbf{P}$. However, it is shown in the engineering literature that the result in (78) holds if the system is detectable and stabilisable. Further discussion can be found in Anderson and Moore (1979, section 4.4), Burridge and Wallis (1988) and Harvey (1989, ch 3).

6.4.1 Filtering weights

If the filter is in a steady-state, the recursion for the predictive filter in (75) can be solved to give

$$\mathbf{a}_{t+1|t} = \sum_{j=0}^{\infty} \mathbf{L}^j \mathbf{K} \mathbf{y}_{t-j} + \sum_{j=0}^{\infty} \mathbf{L}^j \mathbf{c}_{t+1-j} + \sum_{j=0}^{\infty} \mathbf{L}^j \mathbf{K} \mathbf{d}_{t-j} \tag{79}$$

Thus it can be seen explicitly how the filtered estimator is a weighted average of past observations. The one-step ahead predictor, $\widetilde{\mathbf{y}}_{t+1|T}$, can similarly be expressed in terms of current and past observations by shifting (56) forward one time period and substituting from (79). Note that when $\mathbf{c}_t$ and $\mathbf{d}_t$ are time-invariant, we can write

$$\mathbf{a}_{t+1|t} = (\mathbf{I} - \mathbf{L}L)^{-1} \mathbf{K} \mathbf{y}_t + (\mathbf{I} - \mathbf{L})^{-1} (\mathbf{c} - \mathbf{K}\mathbf{d}) \tag{80}$$

If we are interested in the weighting pattern for the current filtered estimator, as opposed to one-step ahead, the Kalman filtering equations need to be combined as

$$\mathbf{a}_t = \mathbf{L}^{\dagger} \mathbf{a}_{t-1} + \mathbf{K}^{\dagger} \mathbf{y}_t + \left(\mathbf{c}_t - \mathbf{K}^{\dagger} \mathbf{d}_t\right) \tag{81}$$

where $\mathbf{L}^{\dagger} = (\mathbf{I} - \mathbf{K}^{\dagger}\mathbf{Z})\mathbf{T}$ and $\mathbf{K}^{\dagger} = \mathbf{P}\mathbf{Z}'\mathbf{F}^{-1}$. An expression analogous to (80) is then obtained.

6.4.2 ARIMA representation

The ARIMA representation for any model in SSF can be obtained as follows. Suppose first that the model is stationary. The two equations in the steady-state innovations form may be combined to give

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{Z}(\mathbf{I} - \mathbf{T}L)^{-1} \mathbf{K} \boldsymbol{\nu}_{t-1} + \boldsymbol{\nu}_t \tag{82}$$

The *(vector) moving-average representation* is therefore

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Psi}(L) \boldsymbol{\nu}_t \tag{83}$$

where $\boldsymbol{\Psi}(L)$ is a matrix polynomial in the lag operator

$$\boldsymbol{\Psi}(L) = \mathbf{I} + \mathbf{Z}(\mathbf{I} - \mathbf{T}L)^{-1} \mathbf{K} L \tag{84}$$

Thus, given the steady-state solution, we can compute that MA coefficients.

If the stationarity assumption is relaxed, we can write

$$|\mathbf{I}-\mathbf{T}L|\,\mathbf{y}_t= \left[|\mathbf{I}-\mathbf{T}L|\,\mathbf{I}+\mathbf{Z}\,(\mathbf{I}-\mathbf{T}L)^{\dagger}\,\mathbf{K}L\right]\boldsymbol{\nu}_t \tag{85}$$

where $|\mathbf{I}-\mathbf{T}L|$ may contain unit roots. If, in a univariate model, there are d such unit roots, then the reduced form is an $ARIMA\,(p,d,q)$ model with $p+d \leqslant m$. Thus in the local level model, we find, after some manipulation of (85), that

$$\Delta y_t = \nu_t - \nu_{t-1} + k\nu_{t-1} = \nu_t - (1+p)^{-1}\nu_{t-1} = \nu_t + \theta\nu_{t-1} \tag{86}$$

confirming that the reduced form is $ARIMA\,(0,1,1)$.

6.4.3 Autoregressive representation

Recalling the definition of an innovation vector in (68) we may write

$$\mathbf{y}_t = \mathbf{Z}\mathbf{a}_{t|t-1} + \mathbf{d} + \boldsymbol{\nu}_t$$

Substituting for $\mathbf{a}_{t|t-1}$ from (80), lagged one time period, gives

$$\mathbf{y}_t = \boldsymbol{\delta} + \mathbf{Z}\sum_{j=1}^{\infty}\mathbf{L}^{j-1}\mathbf{K}\mathbf{y}_{t-j} + \boldsymbol{\nu}_t, \qquad Var(\boldsymbol{\nu}_t) = \mathbf{F} \tag{87}$$

where

$$\boldsymbol{\delta} = (\mathbf{I}-\mathbf{Z}(\mathbf{I}-\mathbf{L})^{-1}\mathbf{K})\mathbf{d} + \mathbf{Z}(\mathbf{I}-\mathbf{L})^{-1}\mathbf{c} \tag{88}$$

The *(vector) autoregressive representation* is therefore

$$\boldsymbol{\Phi}(L)\mathbf{y}_t = \boldsymbol{\delta} + \boldsymbol{\nu}_t \tag{89}$$

where $\boldsymbol{\Phi}(L)$ is the matrix polynomial in the lag operator

$$\boldsymbol{\Phi}(L) = \mathbf{I} - \mathbf{Z}(\mathbf{I}-\mathbf{L}L)^{-1}\mathbf{K}L$$

and $\boldsymbol{\delta} = \boldsymbol{\Phi}(1)\mathbf{d} + \mathbf{Z}(\mathbf{I}-\mathbf{L})^{-1}\mathbf{c}$.

If the model is stationary, it may be written as

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}^{-1}(L)\boldsymbol{\nu}_t \tag{90}$$

where $\boldsymbol{\mu}$ is as in the moving-average representation of (83). This implies that $\boldsymbol{\Phi}^{-1}(L) = \boldsymbol{\Psi}(L)$: hence the identity

$$(\mathbf{I} - \mathbf{Z}(\mathbf{I}-\mathbf{L}L)^{-1}\mathbf{K}L)^{-1} = \mathbf{I} + \mathbf{Z}(\mathbf{I}-\mathbf{T}L)^{-1}\mathbf{K}L.$$

6.4.4 Forecast functions

Running the Kalman filter up to time T gives the current estimate of the state vector. This contains the starting values for the forecast functions of the various components and the series itself. The *multi-step predictor* can be written as

$$\widetilde{\mathbf{y}}_{T+l|T} = \mathbf{Z}\mathbf{a}_{T+l|T} = \mathbf{Z}\mathbf{T}^{l}\mathbf{a}_{T}, \quad l = 1, 2, ... \tag{91}$$

This is the MMSE of $\mathbf{y}_{T+\ell}$ in a Gaussian model. The weights assigned to current and past observations may be determined by substituting from (79). Substituting repeatedly from the recursion for the MSE of $\mathbf{a}_{T+l|T}$ gives

$$MSE\left(\widetilde{\mathbf{y}}_{T+l|T}\right) = \mathbf{Z}\mathbf{T}^{l}\mathbf{P}_{T}\mathbf{T}'^{l}\mathbf{Z}' + \mathbf{Z}\left(\sum\limits_{j=0}^{l-1}\mathbf{T}^{j}\mathbf{R}\mathbf{Q}\mathbf{R}'\mathbf{T}'^{j}\right)\mathbf{Z}' + \mathbf{H} \tag{92}$$

It is sometimes more convenient to express $\widetilde{\mathbf{y}}_{T+l|T}$ in terms of the predictive filter, that is as $\mathbf{Z}\mathbf{T}^{l-1}\mathbf{a}_{T+1|T}$. A corresponding expression for the MSE can be written down in terms of $\mathbf{P}_{T+1|T}$.

Local linear trend The forecast function is

$$\widetilde{y}_{T+l|T} = m_{T+l|T} = m_{T} + b_{T}\ell, \quad l = 1, 2, ...$$

while from (92), the MSE is

$$\left(p_{T}^{(1,1)} + 2lp_{T}^{(1,2)} + l^{2}p_{T}^{(2,2)}\right) + l\sigma_{\eta}^{2} + \frac{1}{6}l\left(l-1\right)\left(2l-1\right)\sigma_{\zeta}^{2} + \sigma_{\varepsilon}^{2}, \quad l = 1, 2, ... \tag{93}$$

where $p_{T}^{(i,j)}$ is the ij-th element of the matrix $\mathbf{P}_{T}$. The third term, which is the contribution arising from changes in the slope, leads to the most dramatic increases as l increases. If the trend model were completely deterministic both the second and third terms would disappear. In a model where some components are deterministic, including them in the state vector ensures that their contribution to the MSE of predictions is accounted for by the elements of $\mathbf{P}_{T}$ appearing in the first term.

6.5 Maximum likelihood estimation and the prediction error decomposition

A state space model will normally contain unknown parameters, or hyperparameters, that enter into the system matrices. The vector of such parameters will be denoted by $\boldsymbol{\psi}$. Once the observations are available, the joint density in (67) can be reinterpreted as a likelihood function and written $L\left(\boldsymbol{\psi}\right)$. The ML estimator of $\boldsymbol{\psi}$ is then found by maximising $L\left(\boldsymbol{\psi}\right)$. It follows from the discussion below (67) that the Gaussian likelihood function can be written in terms of the innovations, that is

$$\log L = -\frac{NT}{2}\log 2\pi - \frac{1}{2}\sum\limits_{t=1}^{T}\log\left|\mathbf{F}_{t}\right| - \frac{1}{2}\sum\limits_{t=1}^{T}\boldsymbol{\nu}_{t}'\mathbf{F}_{t}^{-1}\boldsymbol{\nu}_{t} \tag{94}$$

This is sometimes known as the *prediction error decomposition* form of the likelihood.

Once an algorithm for computing the likelihood function has been found, it must be maximised with respect to the unknown parameters $\boldsymbol{\psi}$. This will normally be carried out by some kind of numerical optimisation procedure. A univariate model can usually be reparameterised so that $\boldsymbol{\psi} = \left[\boldsymbol{\psi}_*' \quad \sigma_*^2\right]'$ where $\boldsymbol{\psi}_*$ is a vector containing $n-1$ parameters and σ_*^2 is one of the disturbance variances in the model. The Kalman filter can then be run independently of σ_*^2 and this allows it to be concentrated out of the likelihood function.

If prior information is available on all the elements of $\boldsymbol{\alpha}_0$, then $\boldsymbol{\alpha}_0$ has a proper prior distribution with known mean, $\mathbf{a}_0$, and bounded covariance matrix, $\mathbf{P}_0$. The Kalman filter then yields the exact likelihood function. Unfortunately, genuine prior information is rarely available. The solution is to start the Kalman filter at $t = 0$ with a diffuse prior. Suitable algorithms are discussed in Durbin and Koopman (2001, ch 5).

When parameters are estimated, the formula for $MSE\left(\tilde{\mathbf{y}}_{T+l|T}\right)$ in (66) will underestimate the true MSE because it does not take into account the extra variation, of $0\left(T^{-1}\right)$, due to estimating $\boldsymbol{\psi}$. Methods of approximating this additional variation are discussed in Quenneville and Singh (2000). Using the bootstrap is also a possibility; see Stoffer and Wall (2004).

Diagnostic tests can be based on the standardised innovations, $\mathbf{F}_t^{-1/2}\boldsymbol{\nu}_t$. These residuals are serially independent if $\boldsymbol{\psi}$ is known, but when parameters are estimated the distribution of statistics designed to test for serially correlation are affected just as they are when an ARIMA model is estimated. Auxiliary residuals based on smoothed estimates of the disturbances $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are also useful; Harvey and Koopman (1992) show how they can give an indication of outliers or structural breaks.

6.6 Missing observations, temporal aggregation and mixed frequency

Missing observations are easily handled in the SSF simply omitting the updating equations while retaining the prediction equations. Filtering and smoothing then go through automatically and the likelihood function is constructed using prediction errors corresponding to actual observations. When dealing with flow variables, such as income, the issue is one of temporal aggregation. This may be dealt with by the introduction of a cumulator variable into the state as described in Harvey (1989, sub-section 6.3). The ability to handle missing and temporally aggregated observations offers enormous flexibility, for example in dealing with observations at mixed frequencies. The unemployment series in figure 1 provide an illustration.

It is sometimes necessary to make predictions of the cumulative effect of a flow variable up to a particular lead time. This is especially important in stock or production control problems in operations research. Calculating the correct MSE may be ensured by augmenting the state vector by a cumulator variable and making predictions from the Kalman filter in the usual way; see Johnston and Harrison (1986) and Harvey (1989, pp 225-6). The continuous time solution described later in sub-section 8.3 is more elegant.

6.7 Bayesian Methods

Since the state vector is a vector of random variables, a Bayesian interpretation of the Kalman filter as a way of updating a Gaussian prior distribution on the state to give a posterior is quite natural. The mechanics of filtering, smoothing and prediction are the same irrespective of whether the overall framework is Bayesian or classical. As regards initialization of the Kalman filter for a non-stationary state vector, the use of a proper prior is certainly not necessary from the technical point of view and a diffuse prior provides the solution in a classical framework.

The Kalman filter gives the mean and variance of the distribution of future observations, conditional on currently available observations. For the classical statistician, the conditional mean is the MMSE of the future observations. For the Bayesian, the conditional mean minimises the expected loss for a symmetric loss function. With a quadratic loss function, the expected loss is given by the conditional variance. Further discussion can be found in the chapter by Geweke and Whitman.

The real differences in classical and Bayesian treatments arise when the parameters are unknown. In the classical framework these are estimated by maximum likelihood. Inferences about the state and predictions of future observations are then usually made conditional on the estimated values of the hyperparameters, though some approximation to the effect of parameter uncertainty can be made as noted at the end of sub-section 6.5. In a Bayesian set-up, on the other hand, the hyperparameters, as they are often called, are random variables. The development of simulation techniques based on Gibbs sampling or Markov chain Monte Carlo (MCMC) has now made a full Bayesian treatment a feasible proposition. This means that it is possible to simulate a predictive distribution for future observations that takes account of hyperparameter uncertainty; see, for example, Carter and Kohn (1994) and Frühwirth-Schnatter (2004). The computuations may be speeded up considerably by using the *simulation smoother* introduced by de Jong and Shephard (1995) and further developed by Durbin and Koopman (2002).

Prior distributions of variance parameters are best specified as inverted gamma distributions. This distribution allows a non-informative prior to be adopted as in Frühwirth-Schnatter (1994, p196). It is difficult to construct sensible informative priors for the variances themselves. Any knowledge we might have is most likely to be based on signal-noise ratios. Koop and van Dijk (2000) adopt an approach in which the signal-noise ratio in a random walk plus noise is transformed so as to be between zero and one. Harvey, Trimbur and van Dijk (2003) use non-informative inverted gamma priors on variances together with informative priors on the parameters λ_c and ρ in the stochastic cycle.

7 Multivariate models

The principal STMs can be extended to handle more than one series. Simply allowing for cross-correlations leads to the class of seemingly unrelated times series equation (SUTSE) models. Models with common factors emerge as a special case. As well as having a direct interpretation, multivariate structural time series models may provide more efficient inferences and forecasts. They are particularly useful when a target series is measured with a large error or is subject to a delay, while a related series does not suffer from these problems.

7.1 SUTSE models

Suppose we have N time series. Define the vector $\mathbf{y}_t = (y_{1t}, .., y_{Nt})'$ and similarly for $\boldsymbol{\mu}_t, \boldsymbol{\psi}_t$ and $\boldsymbol{\varepsilon}_t$. Then a multivariate UC model may be set up as

$$\mathbf{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\psi}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim NID(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon), \quad t = 1, ..., T, \tag{95}$$

where $\boldsymbol{\Sigma}_\varepsilon$ is an $N \times N$ positive semi-definite matrix. The trend is

$$\begin{aligned} \boldsymbol{\mu}_t &= \boldsymbol{\mu}_{t-1} + \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad && \boldsymbol{\eta}_t \sim NID(\mathbf{0}, \boldsymbol{\Sigma}_\eta) \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\zeta}_t, && \boldsymbol{\zeta}_t \sim NID(\mathbf{0}, \boldsymbol{\Sigma}_\zeta) \end{aligned} \tag{96}$$

The *similar cycle* model is

$$\begin{bmatrix} \boldsymbol{\psi}_t \\ \boldsymbol{\psi}_t^* \end{bmatrix} = \left[\rho \begin{bmatrix} \cos\lambda_c & \sin\lambda_c \\ -\sin\lambda_c & \cos\lambda_c \end{bmatrix} \otimes \mathbf{I}_N \right] \begin{bmatrix} \boldsymbol{\psi}_{t-1} \\ \boldsymbol{\psi}_{t-1}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\kappa}_t \\ \boldsymbol{\kappa}_t^* \end{bmatrix}, \quad t = 1, ..., T, \tag{97}$$

where $\boldsymbol{\psi}_t$ and $\boldsymbol{\psi}_t^*$ are $N \times 1$ vectors and $\boldsymbol{\kappa}_t$ and $\boldsymbol{\kappa}_t^*$ are $N \times 1$ vectors of the disturbances such that

$$E(\boldsymbol{\kappa}_t \boldsymbol{\kappa}_t') = E(\boldsymbol{\kappa}_t^* \boldsymbol{\kappa}_t^{*'}) = \boldsymbol{\Sigma}_\kappa, \quad E(\boldsymbol{\kappa}_t \boldsymbol{\kappa}_t^{*'}) = \mathbf{0}, \tag{98}$$

where $\boldsymbol{\Sigma}_\kappa$ is an $N \times N$ covariance matrix. The model allows the disturbances to be correlated across the series. Because the damping factor and the frequency, ρ and λ_c, are the same in all series, the cycles in the different series have similar properties; in particular their movements are centred around the same period. This seems eminently reasonable if the cyclical movements all arise from a similar source such as an underlying business cycle. Furthermore, the restriction means that it is often easier to separate out trend and cycle movements when several series are jointly estimated.

Homogeneous models are a special case when all the covariance matrices,$\boldsymbol{\Sigma}_\eta, \boldsymbol{\Sigma}_\zeta, \boldsymbol{\Sigma}_\varepsilon$, and $\boldsymbol{\Sigma}_\kappa$, are proportional; see Harvey (1989, ch 8, section 3). In this case, the same filter and smoother is applied to each series. Multivariate calculations are not required unless MSEs are needed.

7.2 Reduced form and VARMAs

The reduced form of a SUTSE model is a multivariate ARIMA(p, d, q) model with p, d and q taking the same values as in the corresponding univariate case. General expressions may be obtained from the state space form using (85). Similarly the VAR representation may be obtained from (87).

The disadvantage of a VAR is that long lags may be needed to give a good approximation and the loss in degrees of freedom is compounded as the number of series increases. For ARIMA models the restrictions implied by a structural form are very strong - and this leads one to question the usefulness of the whole class. The fact that VARIMA models are far more difficult to estimate than VARs means that they have not been widely used in econometrics - unlike the univariate case, there are few, if any compensating advantages.

The issues can be illustrated with the multivariate random walk plus noise. The reduced form is the multivariate ARIMA(0,1,1) model

$$\Delta \mathbf{y}_t = \boldsymbol{\xi}_t + \boldsymbol{\Theta}\boldsymbol{\xi}_{t-1}, \quad \boldsymbol{\xi}_t \sim NID(\mathbf{0}, \boldsymbol{\Sigma}) \tag{99}$$

In the univariate case, the structural form implies that θ must lie between zero and minus one in the reduced form ARIMA(0,1,1) model. Hence only half the parameter space is admissible. In the multivariate model, the structural form not only implies restrictions on the parameter space in the reduced form, but also reduces its dimension. The total number of parameters in the structural form is $N(N+1)$ while in the unrestricted reduced form, the covariance matrix of $\boldsymbol{\xi}_t$ consists of $N(N+1)/2$ different elements but the MA parameter matrix contains N^2. Thus if N is five, the structural form contains thirty parameters while the unrestricted reduced form has forty. The restrictions are even tighter when the structural model contains several components.[6]

The reduced form of a SUTSE model is always invertible although it may not always be strictly invertible. In other words some of the roots of the MA polynomial for the reduced form may lie on, rather than outside, the unit circle. In the case of the multivariate random walk plus noise, the condition for strict invertibility of the stationary form is that $\boldsymbol{\Sigma}_\eta$ should be p.d. However, the Kalman filter remains valid even if $\boldsymbol{\Sigma}_\eta$ is only p.s.d. On the other hand, ensuring that $\boldsymbol{\Theta}$ satisfies the conditions of invertibility is technically more complex.

In summary, while the multivariate RWN has a clear interpretation and rationale, the meaning of the elements of $\boldsymbol{\Theta}$ is unclear, certain values may be undesirable and invertibility is difficult to impose.

[6] No simple expressions are available for $\boldsymbol{\Theta}$ in terms of structural parameters in the multivariate case. However, its value may be computed from the steady-state by observing that $\mathbf{I} - \mathbf{T}L = (1-L)\mathbf{I}$ and so, proceeding as in (85), one obtains the symmetric $N \times N$ moving average matrix, $\boldsymbol{\Theta}$, as $\mathbf{K} - \mathbf{I} = -\mathbf{L} = -(\mathbf{P} + \mathbf{I})^{-1}$.

7.3 Dynamic common factors

Reduced rank disturbance covariance matrices in a SUTSE model imply common factors. The most important cases arise in connection with the trend and it is this aspect of dynamic factors that the section focusses on. However, it is possible to have common seasonal components and common cycles. The common cycle model is a special case of the similar cycle model and is an example of what Engle and Kozicki (1993) call a common feature.

7.3.1 Common trends and co-integration

With $\mathbf{\Sigma}_\zeta = 0$ the trend in (96) is a random walk plus deterministic drift, $\boldsymbol{\beta}$. If the rank of $\mathbf{\Sigma}_\eta$ is $K < N$, the model can be written in terms of K common trends, $\boldsymbol{\mu}_t^\dagger$, that is

$$\begin{aligned} \mathbf{y}_{1t} &= \boldsymbol{\mu}_t^\dagger + \boldsymbol{\varepsilon}_{1t} \\ \mathbf{y}_{2t} &= \mathbf{\Pi}\boldsymbol{\mu}_t^\dagger + \overline{\boldsymbol{\mu}} + \boldsymbol{\varepsilon}_{2t} \end{aligned} \tag{100}$$

where $\mathbf{y}_t$ is partitioned into a $K \times 1$ vector $\mathbf{y}_{1t}$ and an $R \times 1$ vector $\mathbf{y}_{2t}$, $\boldsymbol{\varepsilon}_\mathbf{t}$ is similarly partitioned, $\mathbf{\Pi}$ is an $R \times K$ matrix of coefficients and the $K \times 1$ vector $\boldsymbol{\mu}_t^\dagger$ follows a multivariate random walk with drift

$$\boldsymbol{\mu}_t^\dagger = \boldsymbol{\mu}_{t-1}^\dagger + \boldsymbol{\beta}^\dagger + \boldsymbol{\eta}_t^\dagger, \quad \boldsymbol{\eta}_t^\dagger \sim NID(\mathbf{0}, \mathbf{\Sigma}_\eta^\dagger), \tag{101}$$

with $\boldsymbol{\eta}_t^\dagger$ and $\boldsymbol{\beta}^\dagger$ being $K \times 1$ vectors and $\mathbf{\Sigma}_\eta^\dagger$ a $K \times K$ positive definite matrix.

The presence of common trends implies co-integration. In the local level model, (118), there exist $R = N - K$ co-integrating vectors. Let $\mathbf{A}$ be an $R \times N$ matrix partitioned as $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$. The common trend system in (118) can be transformed to an equivalent co-integrating system by pre-multiplying by an $N \times N$ matrix

$$\begin{bmatrix} \mathbf{I}_K & \mathbf{0} \\ \mathbf{A}_1 & \mathbf{A}_2 \end{bmatrix} \tag{102}$$

If $\mathbf{A} = (-\mathbf{\Pi}, \mathbf{I}_R)$ this is just

$$\mathbf{y}_{1t} = \boldsymbol{\mu}_t^\dagger + \boldsymbol{\varepsilon}_{1t},$$

$$\mathbf{y}_{2t} = \mathbf{\Pi}\mathbf{y}_{1t} + \overline{\boldsymbol{\mu}} + \boldsymbol{\varepsilon}_t, \tag{103}$$

where $\boldsymbol{\varepsilon}_\mathbf{t} = \boldsymbol{\varepsilon}_{2t} - \mathbf{\Pi}\boldsymbol{\varepsilon}_{1t}$. Thus the second set of equations consists of co-integrating relationships, $\mathbf{A}\mathbf{y}_t$, while the first set contains the common trends. This is a special case of the *triangular representation* of a co-integrating system.

The notion of co-breaking, as expounded in Clements and Hendry (1998), can be incorporated quite naturally into a common trends model by the introduction of a dummy variable, w_t, into the equation for the trend, that is

$$\boldsymbol{\mu}_t^{\dagger}{=}\,\boldsymbol{\mu}_{t-1}^{\dagger}{+}\boldsymbol{\beta}^{\dagger}+\boldsymbol{\lambda}w_t+\boldsymbol{\eta}_t^{\dagger},\quad \boldsymbol{\eta}_t^{\dagger}\sim NID(\mathbf{0},\boldsymbol{\Sigma}_{\eta}^{\dagger}), \tag{104}$$

where $\boldsymbol{\lambda}$ is a $K\times 1$ vector of coefficients. Clearly the breaks do not appear in the R stationary series in $\mathbf{Ay}_t$.

7.3.2 Representation of a common trends model by a vector error correction model (VECM)

The VECM representation of a VAR

$$\mathbf{y}_t=\boldsymbol{\delta}+\sum_{j=1}^{\infty}\boldsymbol{\Phi}_j\mathbf{y}_{t-j}{+}\boldsymbol{\xi}_t \tag{105}$$

is

$$\Delta\mathbf{y}_t=\boldsymbol{\delta}+\boldsymbol{\Phi}^{*}\mathbf{y}_{t-1}+\sum_{r=1}^{\infty}\boldsymbol{\Phi}_r^{*}\Delta\mathbf{y}_{t-r}+\boldsymbol{\xi}_t,\qquad Var(\boldsymbol{\xi}_t)=\boldsymbol{\Sigma} \tag{106}$$

where the relationship between the $N\times N$ parameter matrices, $\boldsymbol{\Phi}_r^{*}$, and those in the VAR model is

$$\boldsymbol{\Phi}^{*}=-\boldsymbol{\Phi}\,(1)=\sum_{k=1}^{\infty}\boldsymbol{\Phi}_k-\mathbf{I},\qquad \boldsymbol{\Phi}_j^{*}=-\sum_{k=j+1}^{\infty}\boldsymbol{\Phi}_k,\qquad j=1,2,\ldots \tag{107}$$

If there are R co-integrating vectors, contained in the $R\times N$ matrix $\mathbf{A}$, then $\boldsymbol{\Phi}^{*}$ contains K unit roots and $\boldsymbol{\Phi}^{*}=\boldsymbol{\Gamma}\mathbf{A}$, where $\boldsymbol{\Gamma}$ is $N\times R$.

If there are no restrictions on the elements of $\boldsymbol{\delta}$ they contain information on the $K\times 1$ vector of common slopes, $\boldsymbol{\beta}^{*}$, and on the $R\times 1$ vector of intercepts, $\boldsymbol{\mu}^{*}$, that constitutes the mean of $\mathbf{Ay}_t$. This is best seen by writing (106) as

$$\Delta\mathbf{y}_t=\mathbf{A}_{\perp}\boldsymbol{\beta}^{*}+\boldsymbol{\Gamma}(\mathbf{Ay}_{t-1}-\boldsymbol{\mu}^{*})+\sum_{r=1}^{p-1}\boldsymbol{\Phi}_r^{*}(\Delta\mathbf{y}_{t-r}-\mathbf{A}_{\perp}\boldsymbol{\beta}^{*})+\boldsymbol{\xi}_t, \tag{108}$$

where $\mathbf{A}_{\perp}$ is an $N\times K$ matrix such that $\mathbf{AA}_{\perp}=\mathbf{0}$, so that there are no slopes in the co-integrating vectors. The elements of $\mathbf{A}_{\perp}\boldsymbol{\beta}^{*}$ are the growth rates of the series. Thus[7]

$$\boldsymbol{\delta}\;{=}(\mathbf{I}{-}\sum_{j=1}^{p-1}\boldsymbol{\Phi}_j^{*})\mathbf{A}_{\perp}\boldsymbol{\beta}^{*}{-}\boldsymbol{\Gamma}\boldsymbol{\mu}^{*} \tag{109}$$

Structural time series models have an implied triangular representation as we saw in (103). The connection with VECMs is not so straightforward. The coefficients of the VECM represention for any UC model with common (random walk plus drift) trends can be computed numerically by using the algorithm of

[7] If we don't want time trends in the series, the growth rates must be set to zero so we must constrain $\boldsymbol{\delta}$ to depend only on the R parameters in $\boldsymbol{\mu}^{*}$ by setting $\boldsymbol{\delta}=-\boldsymbol{\Gamma}\boldsymbol{\mu}^{*}$. In the special case when $R=N$, there are no time trends and $\boldsymbol{\delta}=-\boldsymbol{\Gamma}\boldsymbol{\mu}^{*}$ is the unconditional mean.

Koopman and Harvey (2003). Here we derive analytic expressions for the VECM representation of a local level model, (100), noting that, in terms of the general state space model, $\mathbf{Z}=(\mathbf{I},\mathbf{\Pi}')'$. The coefficient matrices in the VECM depend on the $K\times N$ steady-state Kalman gain matrix, $\mathbf{K}$, as given from the algebraic Riccati equations. Proceeding in this way can give interesting insights into the structure of the VECM.

From the vector autoregressive form of the Kalman filter, (87), noting that $\mathbf{T}=\mathbf{I}_K$, so $\mathbf{L}=\mathbf{I}_K{-}\mathbf{KZ}$, we have

$$\mathbf{y}_t=\boldsymbol{\delta}+\mathbf{Z}(\mathbf{I}_K{-}(\mathbf{I}_K{-}\mathbf{KZ})L)^{-1}\mathbf{K}\mathbf{y}_{t-1}+\boldsymbol{\nu}_t,\qquad Var(\boldsymbol{\nu}_t)=\mathbf{F} \tag{110}$$

(Note that $\mathbf{F}$ and $\mathbf{K}$ depend on $\mathbf{Z},\mathbf{\Sigma}_\eta$ and $\mathbf{\Sigma}_\varepsilon$ via the steady-state covariance matrix, $\mathbf{P}$.) This representation corresponds to a VAR with $\boldsymbol{\nu}_t=\boldsymbol{\xi}_t$ and $\mathbf{F}=\mathbf{\Sigma}$. The polynomial in the infinite vector autoregression, (105), is therefore

$$\mathbf{\Phi}(L)=\mathbf{I}_N-\mathbf{Z}\left[\mathbf{I}_K-(\mathbf{I}_K-\mathbf{KZ})\,L\right]^{-1}\mathbf{K}L$$

The matrix

$$\mathbf{\Phi}(1)=\mathbf{I}_N-\mathbf{Z}\left(\mathbf{KZ}\right)^{-1}\mathbf{K} \tag{111}$$

has the property that $\mathbf{\Phi}(1)\mathbf{Z}=\mathbf{0}$ and $\mathbf{K\Phi}(1)=\mathbf{0}$. Its rank is easily seen to be R, as required by the Granger representation theorem; this follows because it is idempotent and so the rank is equal to the trace.

The expression linking $\boldsymbol{\delta}$ to $\overline{\boldsymbol{\mu}}$ and $\boldsymbol{\beta}^\dagger$ is obtained from (88) as

$$\boldsymbol{\delta}=\left[\mathbf{I}_N-\mathbf{Z}\left(\mathbf{KZ}\right)^{-1}\mathbf{K}\right]\left[\begin{array}{c}\mathbf{0}\\ \overline{\boldsymbol{\mu}}\end{array}\right]+\mathbf{Z}\left(\mathbf{KZ}\right)^{-1}\boldsymbol{\beta}^\dagger \tag{112}$$

since $\mathbf{d}=(\mathbf{0}',\overline{\boldsymbol{\mu}})'$. The vectors $\overline{\boldsymbol{\mu}}$ and $\boldsymbol{\beta}^\dagger$ contain N non-zero elements between them; thus the components of both level and growth are included in $\boldsymbol{\delta}$.

The coefficient matrices in the infinite VECM, (106), are $\mathbf{\Phi}^*=-\mathbf{\Phi}(1)$ and

$$\mathbf{\Phi}_j^*=-\mathbf{Z}\left[\mathbf{I}_K-\mathbf{KZ}\right]^j\left(\mathbf{KZ}\right)^{-1}\mathbf{K},\quad j=1,2,... \tag{113}$$

The VECM of (108) is given by setting $\mathbf{A}_\perp=\mathbf{Z}=(\mathbf{I},\mathbf{\Pi}')'$ and $\boldsymbol{\beta}^*=\boldsymbol{\beta}^\dagger$. The $\mathbf{A}$ matrix is not unique for $N-K=R>1$, but it can be set to $[-\mathbf{\Pi},\mathbf{I}_R]$ and the $\mathbf{\Gamma}$ matrix must then satisfy $\mathbf{\Gamma A}=\mathbf{\Phi}^*$. However, since $\mathbf{A}(\mathbf{0}',\overline{\boldsymbol{\mu}}')'=\boldsymbol{\mu}^*$, this choice of $\mathbf{A}$ implies $\overline{\boldsymbol{\mu}}=\boldsymbol{\mu}^*$. Hence it follows from (109) and (112) that $\mathbf{\Gamma}$ is given by the last R columns of $\mathbf{\Phi}^*$.

7.3.3 Single common trend

For a single common trend we may write

$$\mathbf{y}_t{=}\,\mathbf{z}\mu_t^\dagger{+}\boldsymbol{\varepsilon}_t,\qquad t=1,\ldots,T, \tag{114}$$

where $\mathbf{z}$ is a vector and $\mu_t^\dagger$ is a univariate random walk. It turns out that optimal filtering and smoothing can be carried out exactly as for a univariate local level model for $\overline{\overline{y}}_t = \overline{\sigma}_\varepsilon^2 \mathbf{z}'\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{y}_t$ with $\overline{q} = \sigma_\eta^2/\overline{\sigma}_\varepsilon^2$, where $\overline{\sigma}_\varepsilon^{-2} = \mathbf{z}^{'}\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{z}$. This result[8] is not entirely obvious since, unless the diagonal elements of $\boldsymbol{\Sigma}_\varepsilon$ are the same, univariate estimators would have different $q's$ and hence different smoothing constants. It has implications for estimating an underlying trend from a number of series. The result follows by applying a standard matrix inversion lemma, as in Harvey (1989, p108), to $\mathbf{F}_t^{-1}$ in the vector $\mathbf{k}_t = p_{t|t-1}\mathbf{z}'\mathbf{F}_t^{-1}$ to give

$$\mathbf{k}_t = [p_{t|t-1}^*/(p_{t|t-1}^* + 1)]\overline{\sigma}_\varepsilon^2\mathbf{z}'\boldsymbol{\Sigma}_\varepsilon^{-1} \tag{115}$$

where $p_{t|t-1}^* = \overline{\sigma}_\varepsilon^{-2}p_{t|t-1}$ Thus the Kalman filter can be run as a univariate filter for $\overline{\overline{y}}_t$. In the steady state, $\overline{p}^*$ is as in (13) but using $\overline{q}$ rather than q. Then from (115) we get $\mathbf{k} = [(\overline{p}^* + \overline{q})/(\overline{p}^* + \overline{q} + 1)]\overline{\sigma}_\varepsilon^2\mathbf{z}'\boldsymbol{\Sigma}_\varepsilon^{-1}$.

As regards the VECM representation, $\mathbf{I}_K - \mathbf{KZ} = 1 - \mathbf{k}'\mathbf{z}$ is a scalar and the coefficients of the lagged differences, the elements of the $\boldsymbol{\Phi}_j^{*\prime}s$, all decay at the same rate. Since $\mathbf{k}'\mathbf{z} = (\overline{p}^* + \overline{q})/(\overline{p}^* + \overline{q} + 1)$

$$\boldsymbol{\Phi}_j^* = -(1/\mathbf{k}'\mathbf{z})(1-\mathbf{k}'\mathbf{z})^j\mathbf{z}\mathbf{k}' = -\left(\overline{p}^* + \overline{q} + 1\right)^{-j}\overline{\sigma}_\varepsilon^2\mathbf{z}\mathbf{z}'\boldsymbol{\Sigma}_\varepsilon^{-1}, \quad j = 1, 2, ...$$

Furthermore

$$\boldsymbol{\Phi}(1) = -\boldsymbol{\Phi}^* = \mathbf{I}-(1/\mathbf{k}'\mathbf{z})\mathbf{z}\mathbf{k}' = \mathbf{I}-\overline{\sigma}_\varepsilon^2\mathbf{z}\mathbf{z}'\boldsymbol{\Sigma}_\varepsilon^{-1}. \tag{116}$$

If w_k is the weight attached to y_k in forming the mean, that is w_k is the $k-th$ element of the vector $\overline{\sigma}_\varepsilon^2\mathbf{z}'\boldsymbol{\Sigma}_\varepsilon^{-1}$, the $i-$th equation in the VECM can be expressed[9] as

$$\Delta y_{it} = \delta_i - \left(y_{i,t-1} - z_i\overline{\overline{y}}_{t-1}\right) - z_i\sum_{k=1}^{N} w_k \sum_{j=1}^{\infty}\left(-\overline{\theta}\right)^j \Delta y_{k,t-j} + v_{it}, \tag{117}$$

where δ_i is a constant, $\overline{\theta} = -1/(\overline{p}^* + \overline{q} + 1)$ depends on $\overline{q}$ and the $v_{it}'s$ are serially uncorrelated disturbances. The terms $y_{i,t-1} - z_i\overline{\overline{y}}_{t-1}$ can also be expressed as $N-1$ co-integrating vectors weighted by the elements of the last $N-1$ columns of $\boldsymbol{\Phi}^*$. *The most interesting point to emerge from this representation is that the (exponential) decay of the weights attached to lagged differences is the same for all variables in each equation.*

The single common trends model illustrates the implications of using a VAR or VECM as an approximating model. It has already been noted that an autoregression can be a very poor approximation to a random walk plus noise model, particularly if the signal-noise ratio, q, is small. In a multivariate model the problems are compounded. Thus, ignoring $\overline{\boldsymbol{\mu}}$ and $\beta^\dagger$, a model with a single common trend contains N parameters in addition to the parameters in $\boldsymbol{\Sigma}_\varepsilon$. The VECM has a disturbance covariance matrix with the same number of parameters as $\boldsymbol{\Sigma}_\varepsilon$. However the error correction matrix $\boldsymbol{\Phi}^*$ is $N \times N$ and on top of this a sufficient number of lagged differences, with $N \times N$ parameter matrices, $\boldsymbol{\Phi}_j^*$, must be used to give a reasonable approximation.

[8] Kozicki (1999) gives a related result

[9] In the univariate case $\overline{\overline{y}}_t = y_t$ and so (117) reduces to the (unstandardised) EWMA of differences, (36).

7.4 Convergence

STMs have recently been adapted to model converging economies and to produce forecasts that take account of convergence. Before describing these models it is first necessary to discuss balanced growth.

7.4.1 Balanced growth, stability and convergence

The *balanced growth* UC model is a special case of (95):

$$\mathbf{y}_t = \mathbf{i}\mu_t^{\dagger}+\boldsymbol{\alpha} + \boldsymbol{\psi}_t+\boldsymbol{\varepsilon}_t, \ , \qquad t = 1, ..., T, \tag{118}$$

where $\mu_t^{\dagger}$ is a univariate local linear trend, $\mathbf{i}$ is a vector of ones, and $\boldsymbol{\alpha}$ is an $N \times 1$ vector of constants. Although there may be differences in the level of the trend in each series, the slopes are the same, irrespective of whether they are fixed or stochastic.

A balanced growth model implies that the series have a stable relationship over time. This means that there is a full rank $(N-1) \times N$ matrix, $\mathbf{D}$, with the property that $\mathbf{Di} = \mathbf{0}$, thereby rendering $\mathbf{Dy}_t$ jointly stationary. If the series are stationary in first differences, balanced growth may be incorporated in a vector error correction model (VECM) of the form (108) by letting $\mathbf{A} = \mathbf{D}$ and $\mathbf{A}_{\perp} = \mathbf{i}$. The system has a single unit root, guaranteed by the fact that $\mathbf{Di} = \mathbf{0}$. The constants in $\boldsymbol{\delta}$ contain information on the common slope, β, and on the differences in the levels of the series, as contained in the vector $\boldsymbol{\alpha}$. These differences might be parameterised with respect to the contrasts in $\mathbf{Dy}_{t-1}$. For example if $\mathbf{Dy}_t$ has elements $y_{it} - y_{i+1,t}, i = 1, .., N-1$, then α_i, the $i-th$ element of the $(N-1) \times 1$ vector $\boldsymbol{\alpha}$, is the gap between y_i and y_{i+1}. In any case, $\boldsymbol{\delta} = (\mathbf{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^*)\mathbf{i}\beta - \boldsymbol{\Gamma\alpha}$. The matrix $\boldsymbol{\Gamma}$ contains $N(N-1)$ free parameters and these may be estimated efficiently by OLS applied to each equation in turn. However, there is no guarantee that the estimate of $\boldsymbol{\Gamma}$ will be such that the model is stable.

7.4.2 Convergence models

A multivariate convergence model may be set up as

$$\mathbf{y}_t = \boldsymbol{\alpha}+\beta\mathbf{i}t + \boldsymbol{\mu}_t + \boldsymbol{\psi}_t+\boldsymbol{\varepsilon}_t, \quad t = 1, ..., T \tag{119}$$

with $\boldsymbol{\psi}_t$ and $\boldsymbol{\varepsilon}_t$ defined as (95) and

$$\boldsymbol{\mu}_t = \boldsymbol{\Phi\mu}_{t-1} + \boldsymbol{\eta}_t, \qquad Var(\boldsymbol{\eta}_t) = \boldsymbol{\Sigma}_{\eta} \tag{120}$$

Each row of $\boldsymbol{\Phi}$ sums to unity, $\boldsymbol{\Phi}\mathbf{i} = \mathbf{i}$. Thus setting λ to one in $(\boldsymbol{\Phi} - \lambda\mathbf{I})\mathbf{i} = \mathbf{0}$, shows that $\boldsymbol{\Phi}$ has an eigenvalue of one with a corresponding eigenvector consisting of ones. The other roots of $\boldsymbol{\Phi}$ are obtained by solving $|\boldsymbol{\Phi} - \lambda\mathbf{I}| = 0$; they should have modulus less than one for convergence.

If we write

$$\overline{\boldsymbol{\phi}}'\boldsymbol{\mu}_t = \overline{\boldsymbol{\phi}}'\boldsymbol{\Phi\mu}_{t-1} + \overline{\boldsymbol{\phi}}'\boldsymbol{\eta}_t$$

it is clear that the $N \times 1$ vector of weights, $\overline{\boldsymbol{\phi}}$, which gives a random walk must be such that $\overline{\boldsymbol{\phi}}^{\prime}(\mathbf{\Phi}-\mathbf{I})=\mathbf{0}^{\prime}$. Since the roots of $\mathbf{\Phi}^{\prime}$ are the same as those of $\mathbf{\Phi}$, it follows from writing $\mathbf{\Phi}\overline{\boldsymbol{\phi}}^{\prime}=\overline{\boldsymbol{\phi}}^{\prime}$ that $\overline{\boldsymbol{\phi}}$ is the eigenvector of $\mathbf{\Phi}^{\prime}$ corresponding to its unit root. This random walk, $\overline{\mu}_{\phi t}=\overline{\boldsymbol{\phi}}^{\prime}\boldsymbol{\mu}_t$, is a common trend in the sense that it yields the common growth path to which all the economies converge. This is because $\lim_{j\rightarrow\infty}\mathbf{\Phi}^j=\mathbf{i}\overline{\boldsymbol{\phi}}^{\prime}$. The common trend for the observations is a random walk with drift, β.

The *homogeneous* model has $\mathbf{\Phi}=\!\phi\mathbf{I}+(1\!-\!\phi)\mathbf{i}\overline{\boldsymbol{\phi}}^{\prime}$, where $\mathbf{i}$ is an $N\!\times\!1$ vector of ones, ϕ is a scalar convergence parameter and $\overline{\boldsymbol{\phi}}$ is an $N\!\times\!1$ vector of parameters with the property that $\overline{\boldsymbol{\phi}}^{\prime}\mathbf{i}=\!1$. (It is straightforward to confirm that $\overline{\boldsymbol{\phi}}$ is the eigenvector of $\mathbf{\Phi}^{\prime}$ corresponding to the unit root). The likelihood function is maximized numerically with respect to ϕ and the elements of $\overline{\boldsymbol{\phi}}$, denoted $\overline{\phi}_i, i=1,...,N$; the $\boldsymbol{\mu}_t$ vector is initialised with a diffuse prior. It is assumed that $0\leq\phi\leq1$, with $\phi=1$ indicating no convergence. The $\overline{\phi}_i^{\prime}s$ are constrained to lie between zero and one and to sum to one.

In a homogeneous model, each trend can be decomposed into the common trend and a convergence component. The vector of convergence components defined by is $\boldsymbol{\mu}_t^{\dagger}=\boldsymbol{\mu}_t-\mathbf{i}\overline{\mu}_{\phi t}$ and it is easily seen that

$$\boldsymbol{\mu}_t^{\dagger}=\phi\boldsymbol{\mu}_{t-1}^{\dagger}+\boldsymbol{\eta}_t^{\dagger},\qquad t=1,...,T. \tag{121}$$

where $\boldsymbol{\eta}_t^{\dagger}=\boldsymbol{\eta}_t-\mathbf{i}\overline{\eta}_{\phi t}$. The error correction form for each series

$$\Delta\mu_{it}^{\dagger}=(\phi-1)\mu_{i,t-1}^{\dagger}+\eta_{it}^{\dagger},\qquad i=1,...,N,$$

shows that its relative growth rate depends on the gap between it and the common trend. Substituting (121) into (119) gives

$$\mathbf{y}_t=\boldsymbol{\alpha}\!+\!\beta\mathbf{i}t+\mathbf{i}\overline{\mu}_{\phi t}+\boldsymbol{\mu}_t^{\dagger}+\boldsymbol{\psi}_t\!+\!\boldsymbol{\varepsilon}_t,\quad t=1,...,T$$

Once convergence has taken place, the model is of the balanced growth form, (118), but with an additional stationary component $\boldsymbol{\mu}_t^{\dagger}$.

The smooth homogeneous convergence model is

$$\mathbf{y}_t=\boldsymbol{\alpha}+\boldsymbol{\mu}_t+\boldsymbol{\psi}_t\!+\!\boldsymbol{\varepsilon}_t,\quad t=1,...,T \tag{122}$$

and

$$\boldsymbol{\mu}_t=\mathbf{\Phi}\boldsymbol{\mu}_{t-1}+\boldsymbol{\beta}_{t-1},\qquad\boldsymbol{\beta}_t=\mathbf{\Phi}\boldsymbol{\beta}_{t-1}+\boldsymbol{\zeta}_t,\qquad Var(\boldsymbol{\zeta}_t)=\mathbf{\Sigma}_{\zeta},$$

with $\mathbf{\Phi}=\!\phi\mathbf{I}+(1\!-\!\phi)\mathbf{i}\overline{\boldsymbol{\phi}}^{\prime}$ as before. The convergence components can now be given a second-order error correction representation as in sub-section 2.8. The forecasts converge to those of a smooth common trend, but in doing so they may exhibit temporary divergence. The extracted components change relatively smoothly thereby enabling them to be separated from transitory cycles.

Using scalar notation to write the smooth homogeneous model in terms of the common trend, $\overline{\mu}_{\phi,t}$, and convergence processes, $\mu^{\dagger}_{it} = \mu_{it} - \overline{\mu}_{\phi,t}, i = 1, ..., N$, we obtain

$$y_{it} = \alpha_i + \overline{\mu}_{\phi,t} + \mu^{\dagger}_{it} + \psi_{it} + \varepsilon_{it}, \quad i = 1, ..., N, \tag{124}$$

where $\Sigma\alpha_i = 0$, the common trend is

$$\overline{\mu}_{\phi t} = \overline{\mu}_{\phi t-1} + \overline{\beta}_{\phi t-1}, \qquad \overline{\beta}_{\phi t} = \overline{\beta}_{\phi,t-1} + \overline{\eta}_{\phi t}$$

and the convergence components are

$$\mu^{\dagger}_{it} = \phi\mu^{\dagger}_{i,t-1} + \beta^{\dagger}_{it}, \quad \beta^{\dagger}_{it} = \phi\beta^{\dagger}_{i,t-1} + \eta^{\dagger}_{it}, \quad i = 1, ..., N$$

US regions Carvalho and Harvey (2002) fit a smooth, homogeneous absolute convergence model, (124) with $\alpha_i = 0, i = 1, ..., N$ to annual series of six US regions. (NE and ME were excluded as they follow growth paths that, especially for the last two decades, seem to be diverging from the growth paths of the other regions.). The similar cycle parameters were estimated to be $\rho = 0.79$ and $2\pi/\lambda = 8.0$ years, while the estimate of ϕ was 0.889 and the weights, $\overline{\phi}_i$, were such that the common trend is basically constructed by weighting Great Lakes two-thirds and Plains one third. The model not only allows a separation into trends and cycles but also separates out the long-run balanced growth path from the transitional (converging) regional dynamics, thus permitting a characterisation of convergence stylised facts. Figure 10 shows the forecasts of the convergence components for the six regional series over a twenty year horizon (2000-2019). The striking feature of this figure is not the eventual convergence, but rather the prediction of divergence in the short run. Thus, although Plains and Great Lakes converge rapidly to the growth path of the common trend, which is hardly surprising given the composition of the common trend, the Far West, Rocky Mountains, South East and South West are all expected to widen their income gap, relative to the common trend, during the first five years of the forecast period. Only then do they resume their convergence towards the common trend and even then with noticeable differences in dynamics. This temporary divergence is a feature of the smooth convergence model; the second-order error correction specification not only admits slower changes but also, when the convergence process stalls, allows for divergence in the short run.

.

7.5 Forecasting and Nowcasting with Auxiliary Series

The use of an auxiliary series that is a coincident or leading indicator yields potential gains for nowcasting and forecasting. Our analysis will be based on bivariate models. We will take one series, the first, to be the target series while the second is the related series. With nowcasting our concern is with the reduction in the MSE in estimating the level and the slope. We then examine how this translates into gains for forecasting. The emphasis is somewhat different from that in the chapter by Marcelino where the concern is with the information to be gleaned from a large number of series.

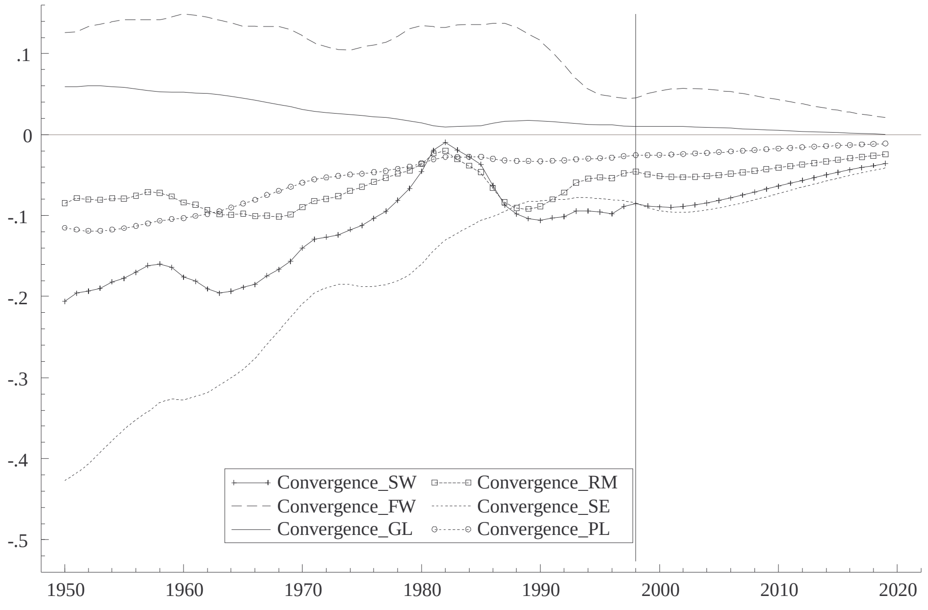

Figure 10: Forecasts for convergence components in US regions.

We will concentrate on the local linear trend model, that is

$$\mathbf{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\varepsilon}_t, \qquad t = 1, \ldots, T, \quad \boldsymbol{\varepsilon}_t \sim NID\left(0, \boldsymbol{\Sigma}_\varepsilon\right) \tag{125}$$

where $\mathbf{y}_t$ and all the other vectors are 2×1 and $\boldsymbol{\mu}_t$ is as in (96). It is useful to write the covariance matrices of $\boldsymbol{\eta}_t$ as

$$\boldsymbol{\Sigma}_\eta = \begin{bmatrix} \sigma_{1\eta}^2 & \rho_\eta \sigma_{1\eta}\sigma_{2\eta} \\ \rho_\eta \sigma_{1\eta}\sigma_{2\eta} & \sigma_{2\eta}^2 \end{bmatrix} \tag{126}$$

where ρ_η is the correlation and similarly for the other disturbance covariance matrices, where the correlations will be ρ_ε and ρ_ζ.

When $\rho_\zeta = \pm 1$ there is then only one source of stochastic movement in the two slopes. This is the *common slopes* model. We can write

$$\beta_{2t} = \bar{\beta} + \theta \beta_{1t}, \quad t = 1, ..., T \tag{127}$$

where $\theta = sgn(\rho_\zeta)\sigma_{2\zeta}/\sigma_{1\zeta}$ and $\bar{\beta}$ is a constant. When $\bar{\beta} = 0$, the model has *proportional slopes.* If, furthermore, θ is equal to one, that is $\sigma_{2\zeta} = \sigma_{1\zeta}$ and ρ_ζ positive, there are *identical slopes.*

The series in a common slopes model are *co-integrated* of order (2,1). Thus, although both y_{1t} and y_{2t} require second differencing to make them stationary, there is a linear combination of first differences which is stationary. If, in addition, $\rho_\eta = \pm 1$, and, furthermore, $\sigma_{2\eta}/\sigma_{1\eta} = \sigma_{2\zeta}/\sigma_{1\zeta}$, then the series are CI(2,2), meaning that there is a linear combination of the observations themselves which is stationary. These conditions mean that $\boldsymbol{\Sigma}_\zeta$ is proportional to $\boldsymbol{\Sigma}_\eta$, which is a special case of what Koopman *et al* (2000) call *trend homogeneity.*

7.5.1 Coincident (concurrent) indicators

In order to gain some insight into the potential gains from using a coincident indicator for nowcasting and forecasting, consider the local level model, that is (125) without the vector of slopes, $\boldsymbol{\beta}_t$. The MSE matrix of predictions is given by a straightforward generalisation of (15), namely

$$MSE\left(\widetilde{y}_{T+l|T}\right) = \mathbf{P}_T + l\boldsymbol{\Sigma}_\eta + \boldsymbol{\Sigma}_\varepsilon, \quad l = 1, 2, ...$$

The gains arise from $\mathbf{P}_T$ as the current level is estimated more precisely. However, $\mathbf{P}_T$ will tend to be dominated by the uncertainty in the level as the lead time increases.

Assuming the target series to be the first series, interest centres on RMSE($\widetilde{\mu}_{1T}$). It might be thought that high correlation between the disturbances in the two series necessarily leads to big reductions in this RMSE. However, this need not be the case. If $\boldsymbol{\Sigma}_\eta = q\boldsymbol{\Sigma}_\varepsilon$, where q is a positive scalar, the model as a whole is homogeneous, and there is no gain from a bivariate model (except in the estimation of the factors of proportionality). This is because the bivariate filter is the same as the univariate filter; see Harvey (1989, pp 435-42). As a simple illustration, consider a model with $\sigma_{2\varepsilon} = \sigma_{1\varepsilon}$ and $q = 0.5$. RMSEs were calculated from the steady-state $\mathbf{P}$ matrix for various combinations of ρ_{ε} and ρ_{η}. With $\rho_{\varepsilon} = 0.8$, RMSE$(\widetilde{\mu}_{1T})$ relative to that obtained in the univariate model is $0.94, 1$ and 0.97 for ρ_{η} equal to $0, 0.8$ and 1 respectively. Thus there is no gain under homogeneity and there is less reduction in RMSE when the levels are perfectly correlated compared with when they are uncorrelated. The biggest gain in precision is when $\rho_{\varepsilon} = -1$ and $\rho_{\eta} = 1$. In fact if the levels are identical, $(y_{1t} + y_{2t})/2$ estimates the level exactly. When $\rho_{\varepsilon} = 0$, the relative RMSEs are $1, 0.93$ and 0.80 for ρ_{η} equal to $0, 0.8$ and 1 respectively.

Observations on a related series can also be used to get more accurate estimates of the underlying growth rate in a target series and hence more accurate forecasts. For example, when the target series contains an irregular component but the related series does not, there is always a reduction in RMSE$(\widetilde{\beta}_{1T})$ from using the related series (unless the related series is completely deterministic). Further analysis of potential gains can be found in Harvey and Chung (2000)

Labour Force Survey- The challenge posed by combining quarterly survey data on unemployment with the monthly claimant count was described in the introduction. The appropriate model for the monthly CC series, y_{2t}, is a local linear trend with no irregular component. The monthly model for the LFS series is similar, except that the observations contain a survey sampling error as described in sub-section 2.5. A bivariate model with these features can be handled within the state space framework even if the LFS observations are only available every quarter or, as was the case before 1992, every year. A glance at figure 1 suggests that the underlying trends in the two series are not the same. However, such divergence does not mean that the CC series contains no usable information. For example it is plausible that the underlying slopes of the two series move closely together even though the levels show a tendency to drift apart. In terms of model (125) this corresponds to a high correlation, ρ_{ζ}, between the stochastic slopes, accompanied by a much lower correlation for the levels, ρ_{η}. The analysis at the start of this sub-section indicates that such a combination could lead to a considerable gain in the precision with which the underlying change in ILO unemployment is estimated. Models were estimated using monthly CC observations from 1971 together with quarterly LFS observations from May 1992 and annual observations from 1984. The last observations are in August 1998. The proportional slopes model is the preferred one. The weighting functions are shown in figure 11.

Output gap - Kuttner (1994) uses a bivariate model for constructing a timely and economically sensible estimate of potential output by exploiting the cyclical relationship between inflation and the output gap. Planas and Rossi (2004) extend this idea further and examine the implications for detecting turning points. Kuttner's model combines the equation for the trend-cycle decomposition of GDP, y_t, in (27) with a Phillips curve effect that relates inflation to the change in GDP and its cycle, ψ_t, that is

$$\Delta p_t = \mu_t^p + \gamma \Delta y_t + \beta \psi_t + u_t,$$

where p_t is the logarithm of the price level, μ_t^p is the trend in inflation and u_t

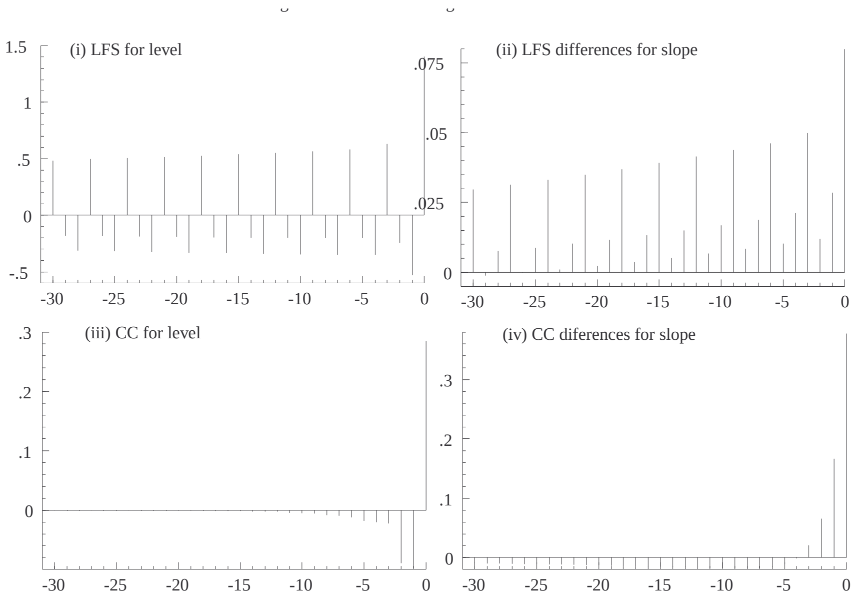

Figure 11: Weights applied to levels and differences of LFS and CC in estimating the current underlying change in LFS

is a moving average disturbance.

7.5.2 Delayed observations and leading indicators

Suppose that the first series is observed with a delay. We can then use the second series to get a better estimate of the first series and its underlying level than could be obtained by univariate forecasting. For the local level, the measurement equation at time T is then

$$y_{2,T} = (0\ 1)\boldsymbol{\mu}_T + \varepsilon_{2,T}$$

and applying the KF we find

$$m_{1,T} = m_{1,T|T-1} + \frac{p_{1,2,T|T-1}}{p_{2,T|T-1} + \sigma^2_{\varepsilon 2}}(y_{2,T} - \widetilde{y}_{2,T|T-1})$$

where, for example, $p_{1,2,T|T-1}$ is the element of $\mathbf{P}_{T|T-1}$ in row one, column two. The estimator of $y_{1,T}$ is given by the same expression, though the MSE's are different. In the homogeneous case it can be shown that the MSE is multiplied by $1-\rho^2$, where ρ is the correlation between the disturbances; see Harvey (1989, p 467). The analysis of leading indicators is essentially the same.

7.5.3 Preliminary observations and data revisions

The optimal use of different vintages of observations in constructing the best estimate of a series, or its underlying level, at a particular date is an example of nowcasting; see Harvey (1989, 337-41) and the references therein and the chapter by Croushore. Using a state space approach, Patterson (1995) provides recent evidence on UK consumers' expenditure and concludes (p54) that '..preliminary vintages are not efficient forecasts of the final vintage.'

Benchmarking can be regarded as another example of nowcasting in which monthly or quarterly observations collected over the year are readjusted so as to be consistent with the annual total obtained from another source such as a survey; see Durbin and Quenneville (1997). The state space treatment is similar to that of data revisions.

8 Continuous time

A continuous time model is more fundamental than one in discrete time. For many variables, the process generating the observations can be regarded as a continuous one even though the observations themselves are only made at discrete intervals. Indeed a good deal of the theory in economics and finance is based on continuous time models.

There are also strong statistical arguments for working with a continuous time model. Apart from providing an elegant solution to the problem of irregularly spaced observations, a continuous time model has the attraction of not being tied to the time interval at which the observations happen to be made. One of the consequences is that, for flow variables, the parameter space is more extensive than it typically would be for an analogous discrete time model. The continuous time formulation is also attractive for forecasting flow variables, particularly when cumulative predictions are to be made over a variable lead time.

Only univariate time series will be considered here. We will suppose that observations are spaced at irregular intervals. The $\tau - th$ observation will be denoted y_τ, for $\tau = 1, ..., T$, and t_τ will denote the time at which it is made, with $t_0 = 0$. The time between observations will be denoted by $\delta_\tau = t_\tau - t_{\tau-1}$.

As with discrete time models the state space form provides a general framework within which estimation and prediction may be carried out. The first sub-section shows how a continuous time transition equation implies a discrete time transition equation at the observation points. The state space treatment for stocks and flows is then set out.

8.1 Transition equations

The continuous time analogue of the time-invariant discrete time transition equation

$$d\boldsymbol{\alpha}(t) = \mathbf{A}\boldsymbol{\alpha}(t)\,dt + \mathbf{R}\mathbf{Q}^{1/2}d\mathbf{W}_\eta(t) \tag{128}$$

where the $\mathbf{A}$ and $\mathbf{R}$ are $m \times m$ and $m \times g$ respectively, and may be functions of hyperparameters, $\mathbf{W}_\eta(t)$ is a standard multivariate Wiener process and $\mathbf{Q}$ is a $g \times g$ psd matrix. The condition for $\boldsymbol{\alpha}(t)$ to be stationary is that the characteristic roots of $\mathbf{A}$ should have negative real parts.

The treatment of continuous time models hinges on the solution to the differential equations in (128). By defining $\boldsymbol{\alpha}_\tau$ as $\boldsymbol{\alpha}(t_\tau)$ for $\tau = 1, ..., T$, we are able to establish the discrete time transition equation,

$$\boldsymbol{\alpha}_\tau = \mathbf{T}_\tau \boldsymbol{\alpha}_{\tau-1} + \boldsymbol{\eta}_\tau \quad \tau = 1, ..., T, \tag{129}$$

where

$$\mathbf{T}_\tau = \exp(\mathbf{A}\boldsymbol{\delta}_\tau) \tag{130}$$

and $\boldsymbol{\eta}_\tau$ is a multivariate white-noise disturbance term with zero and covariance matrix

$$\mathbf{Q}_\tau = \int_0^{\delta_\tau} e^{\mathbf{A}(\delta_\tau - s)} \mathbf{RQR}' e^{\mathbf{A}'(\delta_\tau - s)} ds \tag{131}$$

The evaluation of the transition matrix follows from the definition of a matrix exponential which is

$$\exp(\mathbf{A}) = \mathbf{I} + \mathbf{A} + \frac{1}{2!}\mathbf{A}^2 + \frac{1}{3!}\mathbf{A}^3 + \cdots \tag{132}$$

where $\mathbf{A}$ is an $N \times N$ matrix. The more difficult task is to determine $\mathbf{Q}_\tau$ since this involves the integration of a matrix exponential.

The condition for $\boldsymbol{\alpha}(t)$ to be stationary is that the real parts of the roots of $\mathbf{A}$ should be negative. This translates into the discrete time condition that the roots of $\mathbf{T} = \exp(\mathbf{A})$ should lie outside the unit circle. If $\boldsymbol{\alpha}(t)$ is stationary, the mean of $\boldsymbol{\alpha}(t)$ is zero and the covariance matrix is

$$Var[\boldsymbol{\alpha}(t)] = \int_{-\infty}^{0} e^{-\mathbf{A}s} \mathbf{RQR}' e^{-\mathbf{A}'s} ds \tag{133}$$

The initial conditions for $\boldsymbol{\alpha}(t_0)$ are therefore $\mathbf{a}_{1|0} = \mathbf{0}$ and $\mathbf{P}_{1|0} = Var[\boldsymbol{\alpha}(t)]$.

The main structural components are formulated in continuous time in the following way.

Trend In the local level model, the level component, $\mu(t)$, is defined by $d\mu(t) = \sigma_\eta dW_\eta(t)$, where $W_\eta(t)$ is a standard Wiener process. Thus the increment $d\mu(t)$ has mean zero and variance $\sigma_\eta^2 dt$.

The linear trend component is

$$\begin{bmatrix} d\mu(t) \\ d\beta(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mu(t)\,dt \\ \beta(t)\,dt \end{bmatrix} + \begin{bmatrix} \sigma_\eta dW_\eta(t) \\ \sigma_\zeta dW_\zeta(t) \end{bmatrix} \tag{134}$$

where $W_\eta(t)$ and $W_\zeta(t)$ are mutually independent Wiener processes.

Cycle The continuous cycle is

$$\begin{bmatrix} d\psi(t) \\ d\psi^*(t) \end{bmatrix} = \begin{bmatrix} \log\rho & \lambda_c \\ -\lambda_c & \log\rho \end{bmatrix} \begin{bmatrix} \psi(t) \\ \psi^*(t) \end{bmatrix} + \begin{bmatrix} \sigma_\kappa dW_\kappa(t) \\ \sigma_\kappa dW_\kappa^*(t) \end{bmatrix} \tag{135}$$

where $\kappa(t)$ and $\kappa^*(t)$ are mutually uncorrelated continuous time white-noise processes with the same variance, σ_κ^2, and ρ and λ_c are parameters, the latter being the frequency of the cycle. The characteristic roots of the matrix containing these parameters are $\log\rho \pm i\lambda_c$, the condition for $\psi(t)$ to be a stationary process is $\rho < 1$.

Seasonal The continuous time seasonal model is the sum of a suitable number of trigonometric components, $\gamma_j(t)$, generated by processes of the form (135) with ρ equal to unity and λ_c set equal to the appropriate seasonal frequency λ_j for $j = 1, ..., [s/2]$.

8.2 Stock variables

The discrete state space form for a stock variable generated by a continuous time process consists of the transition equation (129) together with the measurement equation

$$y_\tau = \mathbf{z}'\boldsymbol{\alpha}(t_\tau) + \varepsilon_\tau = \mathbf{z}'\boldsymbol{\alpha}_\tau + \varepsilon_\tau, \quad \tau = 1, ..., T \tag{136}$$

where ε_τ is a white-noise disturbance term with mean zero and variance σ_ε^2 which is uncorrelated with integrals of $\boldsymbol{\eta}(t)$ in all time periods. The Kalman filter can therefore be applied in a standard way. The discrete time model is time-invariant for equally spaced observations, in which case it is usually convenient to set δ_τ equal to unity. In a Gaussian model, estimation can proceed as in discrete time models since, even with irregularly spaced observations, the construction of the likelihood function can proceed via the prediction error decomposition.

8.2.1 Structural time series models

The continuous time components defined earlier can be combined to produce a continuous time structural model. As in the discrete case, the components are usually assumed to be mutually independent. Hence the **A** and **Q** matrices are block diagonal and so the discrete time components can be evaluated separately.

Trend For a stock observed at times t_τ, $\tau = 1, ..., T$, it follows almost immediately that

$$\mu_\tau = \mu_{\tau-1} + \eta_\tau, \qquad Var(\eta_\tau) = \delta_\tau\sigma_\eta^2 \tag{137}$$

since

$$\eta_\tau = \mu(t_\tau) - \mu(t_{\tau-1}) = \sigma_\eta \int_{t_{\tau-1}}^{t_\tau} dW_\eta(t) = \sigma_\eta(W_\eta(t_\tau) - W_\eta(t_{\tau-1})).$$

The discrete model is therefore a random walk for equally spaced observations. If the observation at time τ is made up of $\mu(t_\tau)$ plus a white noise disturbance term, ε_τ, with variance σ_ε^2, the discrete time measurement equation can be written

$$y_\tau = \mu_\tau + \varepsilon_\tau, \quad Var(\varepsilon_\tau) = \sigma_\varepsilon^2, \qquad \tau = 1, ..., T \tag{138}$$

and the set-up corresponds exactly to the familiar random walk plus noise model with signal-noise ratio $q_\delta = \delta\sigma_\eta^2/\sigma_\varepsilon^2 = \delta q$.

For the local linear trend model

$$\begin{bmatrix} \mu_\tau \\ \beta_\tau \end{bmatrix} = \begin{bmatrix} 1 & \delta_\tau \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{\tau-1} \\ \beta_{\tau-1} \end{bmatrix} + \begin{bmatrix} \eta_\tau \\ \zeta_\tau \end{bmatrix} \tag{139}$$

In view of the simple structure of the matrix exponential, the evaluation of the covariance matrix of the discrete time disturbances can be carried out directly, yielding

$$Var \begin{bmatrix} \eta_\tau \\ \zeta_\tau \end{bmatrix} = \delta_\tau \left[\begin{array}{ccc} \sigma_\eta^2 + \frac{1}{3}\delta_\tau^2\sigma_\zeta^2 & \vdots & \frac{1}{2}\delta_\tau\sigma_\zeta^2 \\ \dots\dots\dots\dots\dots & \vdots & \dots\dots\dots \\ \frac{1}{2}\delta_\tau\sigma_\zeta^2 & \vdots & \sigma_\zeta^2 \end{array} \right] \tag{140}$$

When δ_τ is equal to unity, the transition equation is of the same form as the discrete time local linear trend (17). However, (140) shows that independence for the continuous time disturbances implies that the corresponding discrete time disturbances are correlated.

When $\sigma_\eta^2 = 0$, signal extraction with this model yields a cubic spline. Harvey and Koopman (2000) argue that this is a good way of carrying out nonlinear regression. The fact that a model is used means that the problem of making forecasts from a cubic spline is solved.

Cycle For the cycle model, use of the matrix exponential definition together with the power series expansions for the cosine and sine functions gives the discrete time model

$$\begin{bmatrix} \psi_\tau \\ \psi_\tau^* \end{bmatrix} = \rho^\delta \begin{bmatrix} \cos\lambda_c\delta_\tau & \sin\lambda_c\delta_\tau \\ -\sin\lambda_c\delta_\tau & \cos\lambda_c\delta_\tau \end{bmatrix} \begin{bmatrix} \psi_{\tau-1} \\ \psi_{\tau-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_\tau \\ \kappa_\tau^* \end{bmatrix} \tag{141}$$

When δ_τ equals one, the transition matrix corresponds exactly to the transition matrix of the discrete time cyclical component. Specifying that $\kappa(t)$ and $\kappa^*(t)$ be independent of each other with equal variances implies that

$$Var \begin{bmatrix} \kappa_\tau \\ \kappa_\tau^* \end{bmatrix} = \left(\sigma_\kappa^2/\log\rho^{-2}\right)\left(1-\rho^{2\delta_\tau}\right)\mathbf{I}$$

If $\rho = 1$, the covariance matrix is simply $\sigma_\kappa^2\delta_\tau\mathbf{I}$.

8.2.2 Prediction

In the general model of (128), the optimal predictor of the state vector for any positive lead time, l, is given by the forecast function

$$\mathbf{a}(t_T + l \mid T) = e^{\mathbf{A}l}\mathbf{a}_T \tag{142}$$

with associated MSE matrix

$$\mathbf{P}(t_T + l \mid T) = \mathbf{T}_l\mathbf{P}_T\mathbf{T}_l' + \mathbf{R}\mathbf{Q}_l\mathbf{R}', \quad l > 0 \tag{143}$$

where $\mathbf{T}_l$ and $\mathbf{Q}_l$ are, respectively (130) and (131) evaluated with δ_τ set equal to l.

The forecast function for the systematic part of the series,

$$\bar{y}\,(t) = \mathbf{z}'\boldsymbol{\alpha}\,(t) \tag{144}$$

can also be expressed as a continuous function of l, namely

$$\widetilde{\bar{y}}\,(t_T + l \mid T) = \mathbf{z}'e^{\mathbf{A}l}\mathbf{a}_T$$

The forecast of an observation made at time $t_T + l$, is

$$\tilde{y}_{T+1|T} = \widetilde{\bar{y}}\,(t_T + l \mid T) \tag{145}$$

where the observation to be forecast has been classified as the one indexed $\tau = T + 1$; its MSE is

$$MSE\left(\tilde{y}_{T+1|T}\right) = \mathbf{z}'\mathbf{P}\,(t_T + l \mid T)\,\mathbf{z} + \sigma_\varepsilon^2$$

The evaluation of forecast functions for the various structural models is relatively straightforward. In general they take the same form as for the corresponding discrete time models. Thus the local level model has a forecast function

$$\tilde{y}\,(t_T + l \mid T) = m\,(t_T + l \mid T) = m_T$$

and the MSE of the forecast of the $(T + 1)$-th observation, at time $t_T + l$, is

$$MSE\left(\tilde{y}_{T+1|T}\right) = p_T + l\sigma_\eta^2 + \sigma_\varepsilon^2$$

which is exactly the same form as (15).

8.3 Flow variables

For a flow

$$y_\tau = \int_0^{\delta_\tau} \mathbf{z}'\boldsymbol{\alpha}\,(t_{\tau-1} + r) + \sigma_\varepsilon \int_0^{\delta_\tau} dW_\varepsilon(t_{\tau-1} + r), \quad \tau = 1, ..., T \tag{146}$$

where $W_\varepsilon\,(t)$ is independent of the Brownian motion driving the transition equation. Thus the irregular component is cumulated continuously whereas in the stock case it only comes into play when an observation is made.

The key feature in the treatment of flow variables in continuous time is the introduction of a cumulator variable, $y^f\,(t)$, into the state space model. The cumulator variable for the series at time t_τ is equal to the observation, y_τ, for $\tau = 1, ..., T$, that is $y^f\,(t_\tau) = y_\tau$. The result is an augmented state space system

$$\begin{bmatrix} \boldsymbol{\alpha}_\tau \\ y_\tau \end{bmatrix} = \begin{bmatrix} e^{\mathbf{A}\boldsymbol{\delta}} & 0 \\ \mathbf{z}'\mathbf{W}\,(\delta_\tau) & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_{\tau-1} \\ y_{\tau-1} \end{bmatrix} + \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0}' & \mathbf{z}' \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_\tau \\ \boldsymbol{\eta}_\tau^f \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \varepsilon_\tau^f \end{bmatrix} \tag{147}$$

$$y_\tau = [\mathbf{0}' \quad 1]\begin{bmatrix}\boldsymbol{\alpha}_\tau \\ y_\tau\end{bmatrix}, \quad \tau = 1, ..., T$$

with $Var\left(\varepsilon_\tau^f\right) = \delta_\tau \sigma_\varepsilon^2$,

$$\mathbf{W}(r) = \int_0^r e^{\mathbf{A}s} ds \tag{148}$$

and

$$Var\begin{bmatrix}\boldsymbol{\eta}_\tau \\ \boldsymbol{\eta}_\tau^f\end{bmatrix} = \int_0^{\delta_\tau} \begin{bmatrix} e^{\mathbf{A}r}\mathbf{RQR}'e^{\mathbf{A}'r} & \vdots & e^{\mathbf{A}r}\mathbf{RQR}'\mathbf{W}'(r) \\ \cdots\cdots\cdots\cdots\cdots & \vdots & \cdots\cdots\cdots \\ \mathbf{W}(r)\,\mathbf{RQR}'e^{\mathbf{A}'r} & \vdots & \mathbf{W}(r)\,\mathbf{RQR}'\mathbf{W}'(r) \end{bmatrix} = \mathbf{Q}_\tau^\dagger$$

Maximum likelihood estimators of the hyperparameters can be constructed via the prediction error decomposition by running the Kalman filter on (147). No additional starting value problems are caused by bringing the cumulator variable into the state vector as $y^f(t_0) = 0$.

An alternative way of approaching the problem is not to augment the state vector, as such, but to treat the equation

$$y_\tau = \mathbf{z}'\mathbf{W}(\delta_\tau)\,\boldsymbol{\alpha}_{\tau-1} + \mathbf{z}'\boldsymbol{\eta}_\tau^f + \varepsilon_\tau^f \tag{149}$$

as a measurement equation. Redefining $\boldsymbol{\alpha}_{\tau-1}$ as $\boldsymbol{\alpha}_\tau^*$ enables this equation to be written as

$$y_\tau = \mathbf{z}_\tau'\boldsymbol{\alpha}_\tau^* + \varepsilon_\tau, \quad \tau = 1, ..., T \tag{150}$$

where $\mathbf{z}_\tau' = \mathbf{z}'\mathbf{W}(\delta_\tau)$ and $\varepsilon_\tau = \mathbf{z}'\boldsymbol{\eta}_\tau^f + \varepsilon_\tau^f$. The corresponding transition equation is

$$\boldsymbol{\alpha}_{\tau+1}^* = \mathbf{T}_{\tau+1}\boldsymbol{\alpha}_\tau^* + \boldsymbol{\eta}_\tau, \quad \tau = 1, ..., T \tag{151}$$

where $\mathbf{T}_{\tau+1} = \exp(\mathbf{A}\delta_\tau)$. Taken together these two equations are a system of the form (52) and (54) with the measurement equation disturbance, ε_τ, and the transition equation disturbance, $\boldsymbol{\eta}_\tau$, correlated. The covariance matrix of $[\boldsymbol{\eta}_\tau' \quad \varepsilon_\tau]'$ is given by

$$Var\begin{bmatrix}\boldsymbol{\eta}_\tau \\ \varepsilon_\tau\end{bmatrix} = \begin{bmatrix}\mathbf{Q}_\tau & \mathbf{g}_\tau \\ \mathbf{g}_\tau' & h_\tau\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{0}' & \mathbf{z}'\end{bmatrix}\mathbf{Q}_\tau^\dagger\begin{bmatrix}\mathbf{I} & \mathbf{0}' \\ \mathbf{0} & \mathbf{z}\end{bmatrix} + \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \delta_\tau\sigma_\varepsilon^2\end{bmatrix} \tag{152}$$

The modified version of the Kalman filter needed to handle such systems is described in Harvey (1989, sub-section 3.2.4). It is possible to find a SSF in which the measurement error is uncorrelated with the state disturbances, but this is at the price of introducing a moving average into the state disturbances; see Bergstrom (1984) and Chambers and McGarry (2002, p 395).

The various matrix exponential expressions that need to be computed for the flow variable are relatively easy to evaluate for trend and seasonal components in STMs.

8.3.1 Prediction

In making predictions for a flow it is necessary to distinguish between the total accumulated effect from time t_τ to time $t_\tau + l$ and the amount of the flow in a single time period ending at time $t_\tau + l$. The latter concept corresponds to the usual idea of prediction in a discrete model.

Cumulative predictions Let $y^f\left(t_T + l\right)$ denote the cumulative flow from the end of the sample to time $t_T + l$. In terms of the state space model of (147) this quantity is y_{T+1} with δ_{T+1} set equal to l. The optimal predictor, $\tilde{y}^f\left(t_T + l \mid T\right)$, can therefore be obtained directly from the Kalman filter as $\tilde{y}_{T+1|T}$. In fact the resulting expression gives the forecast function which we can write as

$$\tilde{y}^f\left(t_T + l \mid T\right) = \mathbf{z}'\mathbf{W}\left(l\right)\mathbf{a}_T, \quad l \geqslant 0 \tag{153}$$

with

$$MSE\left[\tilde{y}^f\left(t_T + l \mid T\right)\right] = \mathbf{z}'\mathbf{W}\left(l\right)\mathbf{P}_T\mathbf{W}'\left(l\right)\mathbf{z} + \mathbf{z}'Var\left(\boldsymbol{\eta}_\tau^f\right)\mathbf{z} + Var\left(\varepsilon_{T+1}^f\right) \tag{154}$$

For the local linear trend,

$$\tilde{y}^f\left(t_T + l \mid T\right) = lm_T + \frac{1}{2}l^2 b_T, \quad l \geqslant 0$$

$$MSE\left[\tilde{y}^f\left(t_T + l \mid T\right)\right] = l^2 p_T^{(1,1)} + l^3 p_T^{(1,2)} + \frac{1}{4}l^4 p_T^{(2,2)} + \frac{1}{3}l^3\sigma_\eta^2 + \frac{1}{20}l^5\sigma_\zeta^2 + l\sigma_\varepsilon^2 \tag{155}$$

where $p_T^{(i,j)}$ is the ij-th element of $\mathbf{P}_T$. Because the forecasts from a linear trend are being cumulated, the result is a quadratic. Similarly, the forecast for the local level, lm_T, is linear.

Predictions over the unit interval Predictions over the unit interval emerge quite naturally from the state space form, (147), as the predictions of $y_{T+l}, l = 1, 2, ...$ with δ_{T+l} set equal to unity for all l. Thus

$$\tilde{y}_{T+l|T} = \mathbf{z}'\mathbf{W}\left(1\right)\mathbf{a}_{T+l-1|T}, \quad l = 1, 2, ... \tag{156}$$

with

$$\mathbf{a}_{T+l-1|T} = e^{A(l-1)}\mathbf{a}_T, \quad l = 1, 2, ... \tag{157}$$

The forecast function for the state vector is therefore of the same form as in the corresponding stock variable model. The presence of the term $\mathbf{W}\left(1\right)$ in (156) leads to a slight modification when these forecasts are translated into a prediction for the series itself. For STMs, the forecast functions are not too different from the corresponding discrete time forecast functions. However, an interesting feature is that pattern of weighting functions is somewhat more general. For example, for a continuous time local level, the MA parameter in the ARIMA(0,1,1) reduced form can take values up to 0.268 and the smoothing constant in the EWMA used to form the forecasts is in the range 0 to 1.268.

8.3.2 Cumulative predictions over a variable lead time

In some applications, the lead time itself can be regarded as a random variable. This happens, for example, in inventory control problems where an order is put in to meet demand, but the delivery time is uncertain. In such situations it may be useful to determine the unconditional distribution of the flow from the current point in time, that is

$$p\left(y_T^f\right) = \int_0^\infty p\left(y^f(t_T + l \mid T\right) p\left(l\right) dl \tag{158}$$

where $p\left(l\right)$ is the p.d.f. of the lead time and $p\left(y^f\left(t_T + l \mid T\right)\right)$ is the distribution of $y^f\left(t_T + l\right)$ conditional on the information at time T. In a Gaussian model, the mean of $y^f\left(t_T + l\right)$ is given by (153), while its variance is the same as the expression for the MSE of $y^f\left(t_T + l\right)$ given in (154). Although it may be difficult to derive the full unconditional distribution of y_T^f, expressions for the mean and variance of this distribution may be obtained for the principal structural time series models. In the context of inventory control, the unconditional mean might be the demand expected in the period before a new delivery arrives.

The mean of the unconditional distribution of y_T^f is

$$E\left(y_T^f\right) = E[\tilde{y}^f\left(t_T + l \mid T\right)] \tag{159}$$

where the expectation is with respect to the distribution of the lead time. Similarly, the unconditional variance is

$$Var\left(y_T^f\right) = E\left[\tilde{y}^f\left(t_T + l \mid T\right)\right]^2 - \left[E\left(\tilde{y}_T^f\right)\right]^2 \tag{160}$$

where the second raw moment of y_T^f can be obtained as

$$E\left[\tilde{y}^f\left(t_T + l \mid T\right)\right]^2 = MSE\left[\tilde{y}^f\left(t_T + l \mid T\right)\right] + \left[\tilde{y}^f\left(t_T + l \mid T\right)\right]^2$$

The expressions for the mean and variance of y_T^f depend on the moments of the distribution of the lead time. This can be illustrated by the local level model. Let the $j-$th raw moment of this distribution be denoted by μ_j', with the mean abbreviated to μ. Then, by specialising (155),

$$E\left(y_T^f\right) = E\left(lm_T\right) = E\left(l\right) m_T = \mu m_T$$

and

$$Var\left(y_T^f\right) = m_T^2 Var\left(l\right) + \mu\sigma_\varepsilon^2 + \mu_2' p_T + \frac{1}{3}\mu_3'\sigma_\eta^2 \tag{161}$$

The first two terms are the standard formulae found in the operational research literature, corresponding to a situation in which σ_η^2 is zero and the (constant) mean is known. The third term allows for the estimation of the mean, which now may or may not be constant, while the fourth term allows for the movements in the mean that take place beyond the current time period.

The extension to the local linear trend and trigonometric seasonal components is dealt with in Harvey and Snyder (1990). As regards the lead time distribution, it may be possible to estimate moments from past observations. Alternatively, a particular distribution may be assumed. Snyder (1984) argues that the gamma distribution has been found to work well in practice.

9 Nonlinear and Non-Gaussian Models

In the *linear state space* form set out at the beginning of section 6 the system matrices are non-stochastic and the disturbances are all white noise. The system is rather flexible in that the system matrices can vary over time. The additional assumption that the disturbances and initial state vector are normally distributed ensures that we have a *linear model,* that is, one in which the optimal predictions are a linear function of the observations. If there is only one disturbance term, as in an ARIMA model, then serial independence of the disturbances is sufficient for the model to be linear, but with more than one disturbance this is not usually the case. What is required is that the innovations be independent.

Non-linearities can be introduced into state space models in a variety of ways. A completely general formulation is laid out in the first sub-section below, but more tractable classes of models are obtained by focussing on different sources of non-linearity. In the first place, the time-variation in the system matrices may be endogenous. This opens up a wide range of possibilities for modelling with the stochastic system matrices incorporating *feedback* in that they depend on past observations or combinations of observations. The Kalman filter can still be applied when the models are conditionally Gaussian, as described in sub-section 9.2. A second source of nonlinearity arises in an obvious way when the measurement and/or transition equations have a nonlinear functional form. Finally the model may be *non-Gaussian.* The state space may still be linear as for example when the measurement equation has disturbances generated by a $t-$distribution. More fundamentally non-normality may be intrinsic to the data. Thus the observations may be count data in which the number of events occuring in each time period is recorded. If these numbers are small, a normal approximation is unreasonable and in order to be data-admissible the model should explicitly take account of the fact that the observations must be non-negative integers. A more extreme example is when the data are dichotomous and can take one of only two values, zero and one. The structural approach to time series model-building attempts to take such data characteristics into account.

Count data models are usually based on distributions like the Poisson and negative binomial. Thus the non-Gaussianity implies a nonlinear measurement equation that must somehow be combined with a mechanism that allows the mean of the distribution to change over time. Sub-section 9.3.1 sets out a class of models which deal with non-Gaussian distributions for the observations by means of conjugate filters. However, while these filters are analytic, the range of dynamic effects that can be handled is limited. A more general class of models is considered in sub-section 9.3.2. The statistical treatment of such models depends on applying computer intensive methods. Considerable progess has been made in recent years in both a Bayesian and classical framework.

When the state variables are discrete, a whole class of models can be built up based on Markov chains. Thus there is intrinsic non-normality in the transition equations and this may be combined with feedback effects. Analytic filters are possible in some cases such as the autoregressive models introduced by Hamilton (1989).

In setting up nonlinear models, there is often a choice between what Cox calls 'parameter driven' models, based on a latent or unobserved process, and 'observation driven' models in which the starting point is a one-step ahead predictive distribution. As a general rule, the properties of parameter driven models are easier to derive, but observation driven models have the advantage that the likelihood function is immediately available. This survey concentrates on parameter driven models, though it is interesting that some models, such as the conjugate ones of sub-section 9.3.1, belong to both classes.

9.1 General State Space Model

In the general formulation of a state space model, the distribution of the observations is specified conditional on the current state and past observations, that is

$$p(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathbf{Y}_{t-1}) \tag{162}$$

where $\mathbf{Y}_{t-1} = \{\mathbf{y}_{t-1,}\mathbf{y}_{t-2,}....\}$. Similarly the distribution of the current state is specified conditional on the previous state and observations so that

$$p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \mathbf{Y}_{t-1}) \tag{163}$$

The initial distribution of the state, $p(\boldsymbol{\alpha}_0)$ is also specified. In a linear Gaussian model the conditional distributions in (162) and (163) are characterised by their first two moments and so they are specified by the measurement and transition equations.

Filtering The statistical treatment of the general state space model requires the derivation of a recursion for $p(\boldsymbol{\alpha}_t|\mathbf{Y}_t)$, the distribution of the state vector conditional on the information at time t. Suppose this is given at time $t-1$. The distribution of $\boldsymbol{\alpha}_t$ conditional on $\mathbf{Y}_{t-1}$ is

$$p(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1}) = \int_{-\infty}^{\infty} p(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\mathbf{Y}_{t-1})d\boldsymbol{\alpha}_{t-1}$$

but the right-hand side may be rearranged as

$$p(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1}) = \int_{-\infty}^{\infty} p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \mathbf{Y}_{t-1})p(\boldsymbol{\alpha}_{t-1}|\mathbf{Y}_{t-1})d\boldsymbol{\alpha}_{t-1} \tag{164}$$

The conditional distribution $p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \mathbf{Y}_{t-1})$ is given by (163) and so $p(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1})$ may, in principle, be obtained from $p(\boldsymbol{\alpha}_{t-1}|\mathbf{Y}_{t-1})$.

As regards updating,

$$\begin{aligned} p(\boldsymbol{\alpha}_t|\mathbf{Y}_t) &= p(\boldsymbol{\alpha}_t|\mathbf{y}_t, \mathbf{Y}_{t-1}) = p(\boldsymbol{\alpha}_t, \mathbf{y}_t|\mathbf{Y}_{t-1})/p(\mathbf{y}_t|\mathbf{Y}_{t-1}) \\ &= p(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathbf{Y}_{t-1})p(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1})/p(\mathbf{y}_t|\mathbf{Y}_{t-1}) \end{aligned} \tag{165}$$

where

$$p(\mathbf{y}_t|\mathbf{Y}_{t-1}) = \int_{-\infty}^{\infty} p(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathbf{Y}_{t-1})p(\boldsymbol{\alpha}_t|\mathbf{Y}_{t-1})d\boldsymbol{\alpha}_t \tag{166}$$

The likelihood function may be constructed as the product of the predictive distributions, (166), as in (67).

Prediction Prediction is effected by repeated application of (164), starting from $p(\boldsymbol{\alpha}_T|\mathbf{Y}_T)$, to give $p(\boldsymbol{\alpha}_{T+l}|\mathbf{Y}_T)$. The conditional distribution of y_{T+l} is then obtained by evaluating

$$p(\mathbf{y}_{T+l}|\mathbf{Y}_T) = \int_{-\infty}^{\infty} p(\mathbf{y}_{T+l}|\boldsymbol{\alpha}_{T+l}, \mathbf{Y}_T)p(\boldsymbol{\alpha}_{T+l}|\mathbf{Y}_T)d\boldsymbol{\alpha}_{T+l} \tag{167}$$

An alternative route is based on noting that the *predictive* distribution of $\mathbf{y}_{T+l}$ for $l > 1$ is given by

$$p\left(\mathbf{y}_{T+l} \mid \mathbf{Y}_T\right) = \int \cdots \int \prod_{j=1}^{l} p\left(\mathbf{y}_{T+j} \mid \mathbf{Y}_{T+j-1}\right) d\mathbf{y}_{T+j}...d\mathbf{y}_{T+l-1} \tag{168}$$

This expression follows by observing that the joint distribution of the future observations may be written in terms of conditional distributions, that is

$$p\left(\mathbf{y}_{T+l}, \mathbf{y}_{T+l-1}, ..., \mathbf{y}_{T+1} \mid \mathbf{Y}_T\right) = \prod_{j=1}^{l} p\left(\mathbf{y}_{T+j} \mid \mathbf{Y}_{T+j-1}\right)$$

The predictive distribution of y_{T+l} is then obtained as a marginal distribution by integrating out $\mathbf{y}_{T+1}$ to $\mathbf{y}_{T+l-1}$. The usual point forecast is the conditional mean

$$E(\mathbf{y}_{T+l}|\mathbf{Y}_T) = \underset{T}{E}\left(\mathbf{y}_{T+l}\right) = \int_{-\infty}^{\infty} \mathbf{y}_{T+l} p\left(\mathbf{y}_{T+l}|\mathbf{Y}_T\right) d\mathbf{y}_{T+l} \tag{169}$$

as this is the minimum mean square estimate. Other point estimates may be constructed. In particular the maximum *a posteriori* estimate is the mode of the conditional distribution. However, once we move away from normality, there is a case for expressing forecasts in terms of the whole of the predictive distribution.

The general filtering expressions may be difficult to solve analytically. Linear Gaussian models are an obvious exception and tractable solutions are possible in a number of other cases. Of particular importance is the class of conditionally Gaussian models described in the next sub-section and the conjugate filters for count and qualitative observations developed in the sub-section afterwards. Where an analytic solution is not available, Kitagawa (1987) has suggested using numerical methods to evaluate the various densities. The main drawback with this approach is the computational requirement: this can be considerable if a reasonable degree of accuracy is to be achieved.

9.2 Conditionally Gaussian Models

A conditionally Gaussian state space model may be written as

$$\mathbf{y}_t = \mathbf{Z}_t\left(\mathbf{Y}_{t-1}\right)\boldsymbol{\alpha}_t + \mathbf{d}_t\left(\mathbf{Y}_{t-1}\right) + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \mid \mathbf{Y}_{t-1} \sim N\left(\mathbf{0}, \mathbf{H}_t\left(\mathbf{Y}_{t-1}\right)\right) \qquad (170)$$

$$\boldsymbol{\alpha}_t = \mathbf{T}_t\left(\mathbf{Y}_{t-1}\right)\boldsymbol{\alpha}_{t-1} + \mathbf{c}_t\left(\mathbf{Y}_{t-1}\right) + \mathbf{R}_t\left(\mathbf{Y}_{t-1}\right)\boldsymbol{\eta}_t, \ \boldsymbol{\eta}_t \mid \mathbf{Y}_{t-1} \sim N\left(\mathbf{0}, \mathbf{Q}_t\left(\mathbf{Y}_{t-1}\right)\right) \qquad (171)$$

with $\boldsymbol{\alpha}_0 \sim N\left(\mathbf{a}_0, \mathbf{P}_0\right)$. Even though the system matrices may depend on observations up to and including $\mathbf{y}_{t-1}$, they may be regarded as being fixed once we are at time $t-1$. Hence the derivation of the Kalman filter goes through exactly as in the linear model with $\mathbf{a}_{t|t-1}$ and $\mathbf{P}_{t|t-1}$ now interpreted as the mean and covariance matrix of the distribution of $\boldsymbol{\alpha}_t$ conditional on the information at time $t-1$. However, since the conditional mean of $\boldsymbol{\alpha}_t$ will no longer be a linear function of the observations, it will be denoted by $\tilde{\boldsymbol{\alpha}}_{t|t-1}$ rather than by $\mathbf{a}_{t|t-1}$. When $\tilde{\boldsymbol{\alpha}}_{t|t-1}$ is viewed as an estimator of $\boldsymbol{\alpha}_t$, then $\mathbf{P}_{t|t-1}$ can be regarded as its conditional error covariance, or MSE, matrix. Since $\mathbf{P}_{t|t-1}$ will now depend on the particular realisation of observations in the sample, it is no longer an unconditional error covariance matrix as it was in the linear case.

The system matrices will usually contain unknown parameters, $\boldsymbol{\psi}$. However, since the distribution of $\mathbf{y}_t$, conditional on $\mathbf{Y}_{t-1}$, is normal for all $t = 1, ..., T$, the likelihood function can be constructed from the predictive errors, as in (94).

The predictive distribution of $\mathbf{y}_{T+l}$ will not usually be normal for $l > 1$. Furthermore it is not usually possible to determine the form of the distribution. Evaluating conditional moments tends to be easier, though whether it is a feasible proposition depends on the way in which past observations enter into the system matrices. At the least one would hope to be able to use the law of iterated expectations to evaluate the conditional expectations of future observations thereby obtaining their MMSEs.

9.3 Count data and qualitative observations

Count data models are usually based on distributions such as the Poisson or negative binomial. If the means of these distributions are constant, or can be modelled in terms of observable variables, then estimation is relatively easy; see, for example, the book on GLIM models by McCullagh and Nelder (1983). The essence of a time series model, however, is that the mean of a series cannot be modelled in terms of observable variables but so has to be captured by some stochastic mechanism. The structural approach explicitly takes into account the notion that there may be two sources of randomness, one affecting the underlying mean and the other coming from the distribution of the observations around that mean. Thus one can consider setting up a model in which the distribution of an observation conditional on the mean is Poisson or negative binomial, while the mean itself evolves as a stochastic process that is always positive. The same ideas can be used to handle qualitative variables.

9.3.1 Models with conjugate filters

The essence of the conjugate filter approach is to formulate a mechanism that allows the distribution of the underlying level to be updated as new observations become available and at the same time to produce a predictive distribution of the next observation. The solution to the problem rests on the use of natural-conjugate distributions of the type used in Bayesian statistics. This allows the formulation of models for count and qualitative data that are analogous to the random walk plus noise model in that they allow the underlying level of the process to change over time, but in a way that is implicit rather than explicit. By introducing a hyperparameter, ω, into these local level models, past observations are discounted in making forecasts of future observations. Indeed it transpires that in all cases the predictions can be constructed by an EWMA, which is exactly what happens in the random walk plus noise model under the normality assumption. Although the models draw on Bayesian techniques, the approach is can still be seen as classical as the likelihood function can be constructed from the predictive distributions and used as the basis for estimating ω. Furthermore the approach is open to the kind of model-fitting methodology used for linear Gaussian models.

The technique can be illustrated with the model devised for observations drawn from a Poisson distribution. Let

$$p\left(y_t \mid \mu_t\right) = \mu_t^{y_t} e^{-\mu_t}/y_t!, \quad t = 1, ..., T. \tag{172}$$

The conjugate prior for a Poisson distribution is the gamma distribution. Let $p\left(\mu_{t-1} \mid Y_{t-1}\right)$ denote the p.d.f. of μ_{t-1} conditional on the information at time $t-1$. Suppose that this distribution is gamma, that is

$$p\left(\mu; a, b\right) = \frac{e^{-b\mu}\mu^{a-1}}{\Gamma\left(a\right) b^{-a}}, \quad a, b > 0$$

with $\mu = \mu_{t-1}, a = a_{t-1}$ and $b = b_{t-1}$ where a_{t-1} and b_{t-1} are computed from the first $t-1$ observations, Y_{t-1}. In the random walk plus noise with normally distributed observations, $\mu_{t-1} \sim N\left(m_{t-1}, p_{t-1}\right)$ at time $t-1$ implies that $\mu_{t-1} \sim N\left(m_{t-1}, p_{t-1} + \sigma_\eta^2\right)$ at time $t-1$. In other words the mean of $\mu_t \mid Y_{t-1}$ is the same as that of $\mu_{t-1} \mid Y_{t-1}$ but the variance increases. The same effect can be induced in the gamma distribution by multiplying a and b by a factor less than one. We therefore suppose that $p\left(\mu_t \mid Y_{t-1}\right)$ follows a gamma distribution with parameters $a_{t|t-1}$ and $b_{t|t-1}$ such that

$$a_{t|t-1} = \omega a_{t-1} \quad \textit{and} \quad b_{t|t-1} = \omega b_{t-1} \tag{173}$$

and $0 < \omega \leqslant 1$. Then

$$E\left(\mu_t \mid Y_{t-1}\right) = a_{t|t-1}/b_{t|t-1} = a_{t-1}/b_{t-1} = E\left(\mu_{t-1} \mid Y_{t-1}\right)$$

while

$$Var\left(\mu_t \mid Y_{t-1}\right) = a_{t|t-1}/b_{t|t-1}^2 = \omega^{-1} Var\left(\mu_{t-1} \mid Y_{t-1}\right)$$

The stochastic mechanism governing the transition of μ_{t-1} to μ_t is therefore defined implicitly rather than explicitly. However, it is possible to show that it is formally equivalent to a multiplicative transition equation of the form

$$\mu_t = \omega^{-1}\mu_{t-1}\eta_t \tag{174}$$

where η_t has a beta distribution with parameters ωa_{t-1} and $(1-\omega)a_{t-1}$; see the discussion in Smith and Miller (1986).

Once the observation y_t becomes available, the posterior distribution, $p(\mu_t \mid Y_t)$, is obtained by evaluating an expression similar to (165). This yields a gamma distribution with parameters

$$a_t = a_{t|t-1} + y_t \quad and \quad b_t = b_{t|t-1} + 1 \tag{175}$$

The initial prior gamma distribution, that is the distribution of μ_t at time $t = 0$, tends to become diffuse, or non-informative, as $a, b \to 0$, although it is actually degenerate at $a = b = 0$ with $\Pr(\mu = 0) = 1$. However, none of this prevents the recursions for a and b being initialised at $t = 0$ and $a_0 = b_0 = 0$. A proper distribution for μ_t is then obtained at time $t = \tau$ where τ is the index of the first non-zero observation. It follows that, conditional on Y_τ, the joint density of the observations $y_{\tau+1}, ..., y_T$ can be constructed as the product of the predictive distributions. For Poisson observations and a gamma prior, the predictive distribution is a negative binomial distribution, that is

$$p(y_t \mid Y_{t-1}) = \frac{\Gamma\left(a_{t|t-1} + y_t\right)}{\Gamma(y_t+1)\,\Gamma\left(a_{t|t-1}\right)} b_{t|t-1}^{a_{t|t-1}} \left(1 + b_{t|t-1}\right)^{-(a_{t|t-1}+y_t)} \tag{176}$$

Hence the log-likelihood function can easily constructed and then maximised with respect to the unknown hyperparameter ω.

It follows from the properties of the negative binomial that the mean of the predictive distribution of y_{T+1} is

$$E(y_{T+1} \mid Y_T) = a_{T+1|T}/b_{T+1|T} = a_T/b_T = \sum_{j=0}^{T-1} \omega^j y_{T-j} / \sum_{j=0}^{T-1} \omega^j \tag{177}$$

the last equality following from repeated substitution from (173) and (175). In large samples the denominator of (177) is approximately equal to $1/(1-\omega)$ when $\omega < 1$ and the weights decline exponentially, as in (7) with $\lambda = 1 - \omega$. When $\omega = 1$, the right-hand side of (177), is equal to the sample mean; it is reassuring that this is the solution given by setting a_0 and b_0 equal to zero.

The l-step-ahead predictive distribution at time T is given by

$$p(y_{T+l} \mid Y_T) = \int_0^\infty p\left(y_{T+l} \mid \mu_{T+l}\right) p\left(\mu_{T+l} \mid Y_T\right) d\mu_{T+l}$$

It could be argued that the assumption embodied in (173) suggests that $p\left(\mu_{T+l} \mid Y_T\right)$ has a gamma distribution with parameters $\omega^l a_T$ and $\omega^l b_T$.This would mean the predictive distribution for y_{T+l} was negative binomial with a and b given by

$\omega^l a_T$ and $\omega^l b_T$ in the formulae above. Unfortunately the evolution that this implies for μ_t is not consistent with what would occur if observations were made at times $T+1, T+2, ..., T+l-1$. In the latter case, the distribution of y_{T+l} at time T is

$$p\left(y_{T+l} \mid Y_T\right) = \sum_{y_{T+l-1}} \cdots \sum_{y_{T+1}} \prod_{j=1}^{l} p\left(y_{T+j} \mid Y_{T+j-1}\right) \tag{178}$$

This is the analogue of (167) for discrete observations. It is difficult to derive a closed form expression for $p\left(y_{T+l|T}\right)$ from (178) for $l > 1$ but it can, in principle, be evaluated numerically. Note, however, by the law of iterated expectations, $E\left(y_{T+l} \mid Y_T\right) = a_T/b_T$ for $l = 1, 2, 3, ...$, so the mean of the predictive distribution is the same for all lead times, just as in the Gaussian random walk plus noise.

Goals scored by England against Scotland Harvey and Fernandes (1989) modelled the number of goals scored by England in international football matches played against Scotland in Glasgow up 1987. Estimation of the Poisson-gamma model gives $\tilde{\omega} = 0.844$. The forecast is 0.82; the full one-step-ahead predictive distribution is shown in Table 1. (For the record, England won the 1989 match, two-nil).

Table 1 *Predictive probability distribution of goals in next match*

Number of goals					
0	1	2	3	4	>4
0.471	0.326	0.138	0.046	0.013	0.005

Similar filters may be constructed for the binomial distribution, in which case the conjugate prior is the beta distribution and the predictive distribution is the beta-binomial, and the negative binomial for which the conjugate prior is again the beta distribution and the predictive distribution is the beta-Pascal. Exponential distributions fit into the same framework with gamma conjugate distributions and Pareto predictive distributions. In all cases the predicted level is an EWMA.

Boat race The Oxford-Cambridge boat race provides an example of modelling qualitative variables by using the filter for the binomial distribution. Ignoring the dead heat of 1877, there were 130 boat races up to and including 1985. We denote a win for Oxford as one, and a win for Cambridge as zero. The runs test clearly indicates serial correlation and fitting the local Bernoulli model by ML gives an estimate of ω of 0.866. This results in an estimate of the probability of Oxford winning a future race of .833. The high probability is a reflection of the fact that Oxford won all the races over the previous ten years. Updating the data to 2000 gives a dramatic change as Cambridge were dominant in the 1990s. Despite Oxford winning in 2000, the estimate of the probability of Oxford winning future races falls to .42. Further updating can be carried out[10] very easily since the probability of Oxford winning is given by an

[10] Cambridge won in 2001 and 2004, Oxford in 2002 and 2003; see www.theboatrace.org/therace/history

EWMA. Note that because the data are binary, the distribution of the forecasts is just binomial (rather than beta-binomial) and this distribution is the same for any lead time.

A criticism of the above class of forecasting procedures is that when simulated the observations tend to go to zero. Specifically, if $\omega < 1, \mu_t \rightarrow 0$ almost surely, as $t \rightarrow \infty$; see Grunwald, Hamza and Hyndman (1997). Nevertheless for a given data set, fitting such a model gives a sensible weighting pattern-an EWMA - for the mean of the predictive distribution. It was argued in the opening section that this is the purpose of formulating a time series model. The fact that a model may not generate data sets with desirable properties is unfortunate but not fatal.

Explanatory variables can be introduced into these local level models via the kind of link functions that appear in GLIM models. Time trends and seasonal effects can be included as special cases. The framework does not extend to allowing these effects to be stochastic, as is typically the case in linear structural models. This may not be a serious restriction. Even with data on continuous variables, it is not unusual to find that the slope and seasonal effects are close to being deterministic. With count and qualitative data it seems even less likely that the observations will provide enough information to pick up changes in the slope and seasonal effects over time.

9.3.2 Exponential family models with explicit transition equations

The exponential family of distributions contains many of the distributions used for modelling count and quantitative data. For a multivariate series

$$p(\mathbf{y}_t|\boldsymbol{\theta}_t) = exp\{\mathbf{y}_t'\boldsymbol{\theta}_t - b_t(\boldsymbol{\theta}_t) + c(\mathbf{y}_t)\}, \quad t = 1, ..., T$$

where $\boldsymbol{\theta}_t$ is an $N \times 1$ vector of 'signals', $b_t(\boldsymbol{\theta}_t)$ is a twice differentiable function of $\boldsymbol{\theta}_t$ and $c(\mathbf{y}_t)$ is a function of $\mathbf{y}_t$ only. The $\boldsymbol{\theta}_t$ vector is related to the mean of the distribution by a link function, as in GLIM models. For example when the observations are supposed to come from a univariate Poisson distribution with mean λ_t we set $\exp(\theta_t) = \lambda_t$. By letting $\boldsymbol{\theta}_t$ depend on a state vector that changes over time, it is possible to allow the distribution of the observations to depend on stochastic components other than the level. Dependence of $\boldsymbol{\theta}_t$ on past observations may also be countenanced, so that

$$p(\mathbf{y}_t|\boldsymbol{\theta}_t) = p(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathbf{Y}_{t-1})$$

where $\boldsymbol{\alpha}_t$ is a state vector. Explanatory variables could also be included. Unlike the models of the previous sub-section, a transitional distribution is explicitly specified rather than being formed implicitly by the demands of conjugacy. The simplest option is to let $\boldsymbol{\theta}_t = \mathbf{Z}_t\boldsymbol{\alpha}_t$ and have $\boldsymbol{\alpha}_t$ generated by a linear transition equation. The statistical treatment is by simulation methods. Shephard and Pitt (1997) base their approach on Markov chain Monte Carlo (MCMC) while Durbin and Koopman (2001) use importance sampling and antithetic variables. Both techniques can also be applied in a Bayesian framework. A full discussion can be found in Durbin and Koopman (2001).

Van drivers Durbin and Koopman (2001, p 230-3) estimate a Poisson model for monthly data on van drivers killed in road accidents in Great Britain. However, they are able to allow the seasonal component to be stochastic; a stochastic slope could also have been included but the case for employing a slope of any kind is weak. Thus the signal is taken to be

$$\theta_t = \mu_t + \gamma_t + \lambda w_t,$$

where μ_t is a random walk and w_t is the seat belt intervention variable. The estimate of σ_ω^2 is, in fact, zero so the seasonal component turns out to be fixed after all. The estimated reduction in van drivers killed is 24.3% which is not far from the 24.1% obtained by Harvey and Fernandes (1989) using the conjugate filter.

Boat race Durbin and Koopman (2001, p 237) allow the probability of an Oxford win, π_t, to change over time, but remain in the range zero to one by taking the link function for the Bernouilli (binary) distribution to be a logit. Thus they let $\pi_t = \exp(\theta_t)/(1 + \exp(\theta_t))$ and let θ_t follow a random walk.

9.4 Heavy-tailed distributions and robustness

Simulation techniques of the kind alluded to in the previous sub-section, are relatively easy to use when the measurement and transition equations are linear but the disturbances are non-Gaussian. Allowing the disturbances to have heavy-tailed distributions provides a robust method of dealing with outliers and structural breaks. While outliers and breaks can be dealt with *ex post* by dummy variables, only a robust model offers a viable solution to coping with them in the future.

9.4.1 Outliers

Allowing ε_t to have a heavy-tailed distribution, such as Student's t, provides a robust method of dealing with outliers. This is to be contrasted with an approach where the aim is to try to detect outliers and then to remove them by treating them as missing or modeling them by an intervention. An outlier is defined as an observation that is inconsistent with the model. By employing a heavy-tailed distribution, such observations are consistent with the model whereas with a Gaussian distribution they would not be. Treating an outlier as though it were a missing observation effectively says that it contains no useful information. This is rarely the case except, perhaps, when an observation has been recorded incorrectly.

Gas consumption in the UK Estimating a Gaussian BSM for gas consumption produces a rather unappealing wobble in the seasonal component at the time North Sea gas was introduced in 1970. Durbin and Koopman (2001, p 233-5) allow the irregular to follow a t-distribution and estimate its degrees of freedom to be 13. The robust treatment of the atypical observations in 1970 produces a more satisfactory seasonal pattern around that time.

Another example of the application of robust methods is the seasonal adjustment paper of Bruce and Jurke (1996).

In small samples it may prove difficult to estimate the degrees of freedom. A reasonable solution then is to impose a value, such as six, that is able to handle outliers. Other heavy tailed distributions may also be used; Durbin and Koopman (2001, p 184) suggest mixtures of normals and the general error distribution.

9.4.2 Structural breaks

Clements and Hendry (2003, p305) conclude that '..shifts in deterministic terms (intercepts and linear trends) are the major source of forecast failure'. However, unless breaks within the sample are associated with some clearly defined event, such as a new law, dealing with them by dummy variables may not be the best way to proceed. In many situations matters are rarely clear cut in that the researcher does not know the location of breaks or indeed how many there may be. When it comes to forecasting matters are even worse.

The argument for modelling breaks by dummy variables is at its most extreme in the advocacy of piecewise linear trends, that is deterministic trends subject to changes in slope modelled as in sub-section 4.1. This is to be contrasted with a stochastic trend where there are small random breaks at all points in time. Of course, stochastic trends can easily be combined with deterministic structural breaks. However, if the presence and location of potential breaks are not known *a priori* there is a strong argument for using heavy-tailed distributions in the transition equation to accommodate them. Such breaks are not deterministic and their size is a matter of degree rather than kind. From the forecasting point of view this makes much more sense: a future break is virtually never deterministic - indeed the idea that its location and size might be known in advance is extremely optimistic. A robust model, on the other hand, takes account of the possibility of future breaks in its computation of MSEs and in the way it adapts to new observations.

9.5 Switching regimes

The observations in a time series may sometimes be generated by different mechanisms at different points in time. When this happens, the series is subject to *switching regimes*. If the points at which the regime changes can be determined directly from currently available information, the Kalman filter provides the basis for a statistical treatment. The first sub-section below gives simple examples involving endogenously determined changes. If the regime is not directly observable but is known to change according to a Markov process we have *hidden Markov chain* models, as described in the book by MacDonald and Zucchini (1997). Models of this kind are described in later sub-sections.

9.5.1 Observable breaks in structure

If changes in regime are known to take place at particular points in time, the SSF is time-varying but the model is linear. The construction of a likelihood function still proceeds via the prediction error decomposition, the only difference being that there are more parameters to estimate. Changes in the past can easily be allowed for in this way.

The point at which a regime changes may be endogenous to the model, in which case it becomes nonlinear. Thus it is possible to have a finite number of regimes each with a different set of hyperparameters. If the signal as to which regime holds depends on past values of the observations, the model can be set up so as to be conditionally Gaussian. Two possible models spring to mind. The first is a two-regime model in which the regime is determined by the sign of $\triangle y_{t-1}$. The second is a *threshold* model, in which the regime depends on whether or not y_t has crossed a certain threshold value in the previous period. More generally, the switch may depend on the estimate of the state based in information at time $t-1$. Such a model is still conditionally Gaussian and allows a fair degree of flexibility in model formulation.

Business cycles In work on the business cycle, it has often been observed that the downward movement into a recession proceeds at a more rapid rate than the subsequent recovery. This suggests some modification to the cyclical components in structural models formulated for macroeconomic time series. A switch from one frequency to another can be made endogenous to the system by letting

$$\lambda_c = \begin{cases} \lambda_1 & \text{if } \tilde{\psi}_{t|t-1} - \tilde{\psi}_{t-1} > 0 \\ \lambda_2 & \text{if } \tilde{\psi}_{t|t-1} - \tilde{\psi}_{t-1} \leqslant 0 \end{cases}$$

where $\tilde{\psi}_{t|t-1}$ and $\tilde{\psi}_{t-1}$ are the MMSEs of the cyclical component based on the information at time $t-1$. A positive value of $\tilde{\psi}_{t|t-1} - \tilde{\psi}_{t-1}$ indicates that the cycle is in an upswing and hence λ_1 will be set to a smaller value than λ_2. In other words the period in the upswing is larger.

9.5.2 Markov chains

Markov chains can be used to model the dynamics of binary data, that is $y_t = 0$ or 1 for $t = 1, ..., T$. The movement from one state, or *regime*, to another is governed by transition probabilities. In a Markov chain these probabilities depend only on the current state. Thus if $y_{t-1} = 1$, $\Pr(y_t = 1) = \pi_1$ and $\Pr(y_t = 0) = 1 - \pi_1$, while if $y_{t-1} = 0$, $\Pr(y_t = 0) = \pi_0$ and $\Pr(y_t = 1) = 1 - \pi_0$. This provokes an interesting contrast with the EWMA that results from the conjugate filter model.[11]

The above ideas may be extended to situations where there is more than one state. The Markov chain operates as before, with a probability specified for

[11] Having said that it should be noted that the Markov chain transition probabilities may be allowed to evolve over time in the same way as a single probability can be allowed to change in a conjugate binomial model ; see Harvey (1989, p 355).

moving from any of the states at time $t-1$ to any other state at time t.

9.5.3 Markov chain switching models

A general state space model was set up at the beginning of this section by specifying a distribution for each observation conditional on the state vector, $\boldsymbol{\alpha}_t$, together with a distribution of $\boldsymbol{\alpha}_t$ conditional on $\boldsymbol{\alpha}_{t-1}$. The filter and smoother were written down for continuous state variables. The concern here is with a single state variable that is discrete. The filter presented below is the same as the filter for a continuous state, except that integration is replaced by summation. The series is assumed to be univariate.

The state variable takes the values 1,2,...,m, and these values represent each of m different regimes. (In the previous sub-section, the term 'state' was used where here we use regime; the use of 'state' for the value of the state variable could be confusing here.) The transition mechanism is a Markov process which specifies $\Pr(\alpha_t = i \mid \alpha_{t-1} = j)$ for $i, j = 1, ..., m$. Given probabilities of being in each of the regimes at time $t-1$, the corresponding probabilities in the next time period are

$$\Pr(\alpha_t = i \mid Y_{t-1}) = \sum_{j=1}^{m} \Pr(\alpha_t = i \mid \alpha_{t-1} = j) \Pr(\alpha_{t-1} = j \mid Y_{t-1}), \quad i = 1, 2, ..., m,$$

and the conditional PDF of y_t is a mixture of distributions given by

$$p(y_t \mid Y_{t-1}) = \sum_{j=1}^{m} p(y_t \mid \alpha_t = j) \Pr(\alpha_t = j \mid Y_{t-1}) \tag{179}$$

where $p(y_t \mid \alpha_t = j)$ is the distribution of y_t in regime j. As regards updating

$$\Pr(\alpha_t = i \mid Y_t) = \frac{p(y_t \mid \alpha_t = i) \cdot \Pr(\alpha_t = i \mid Y_{t-1})}{p(y_t \mid Y_{t-1})}, \quad i = 1, 2, ..., m$$

Given initial conditions for the probability that α_t is equal to each of its m values at time zero, the filter can be run to produce the probability of being in a given regime at the end of the sample. Predictions of future observations can then be made. If $\mathbf{M}$ denotes the transition matrix with ijth element equal to $\Pr(\alpha_t = i \mid \alpha_{t-1} = j)$ and $\mathbf{p}_{t|t-k}$ is the $m \times 1$ vector with ith element $\Pr(\alpha_t = i \mid Y_{t-k})$, $k = 0, 1, 2, ...$, then

$$\mathbf{p}_{T+l|T} = \mathbf{M}^l \mathbf{p}_{T|T}, \qquad l = 1, 2, ...$$

and so

$$p(y_{T+l} \mid Y_T) = \sum_{j=1}^{m} p(y_{T+l} \mid \alpha_{T+l} = j) \Pr(\alpha_{T+l} = j \mid Y_T) \tag{180}$$

The likelihood function can be constructed from the one-step predictive distributions (179). The unknown parameters consist of the transition probabilities in the matrix $\mathbf{M}$ and the parameters in the measurement equation distributions, $p(y_t \mid \alpha_t = j)$, $j = 1, ..., m$.

The above state space form may be extended by allowing the distribution of y_t to be conditional on past observations as well as on the current state. It may also depend on past regimes, so the current state becomes a vector containing the state variables in previous time periods. This may be expressed by writing the state vector at time t as $\boldsymbol{\alpha}_t = (s_t, s_{t-1}, ..., s_{t-p})'$, where s_t is the state variable at time t.

In the model of Hamilton (1989), the observations are generated by an $AR(p)$ process of the form

$$y_t - \mu(s_t) = \phi_1 [y_{t-1} - \mu(s_{t-1})] + ... + \phi_p [y_{t-p} - \mu(s_{t-p})] + \varepsilon_t \tag{181}$$

where $\varepsilon_t \sim NID(0, \sigma^2)$. Thus the expected value of y_t, denoted $\mu(s_t)$, varies according to the regime, and it is the value appropriate to the corresponding lag on y_t that enters into the equation. Hence the distribution of y_t is conditional on s_t and s_{t-1} to s_{t-p} as well as on y_{t-1} to y_{t-p}. The filter of the previous sub-section can still be applied although the summation must now be over all values of the $p+1$ state variables in α_t. An exact filter is possible here because the time series model in (181) is an autoregression. The is no such analytic solution for an ARMA or structural time series model. As a result simulation methods have to be used as in Kim and Nelson (1999) and Luginbuhl and de Vos (1999).

10 Stochastic Volatility

It is now well established that while financial variables such as stock returns are serially uncorrelated over time, their squares are not. The most common way of modelling this serial correlation in volatility is by means of the GARCH class in which it is assumed that the conditional variance of the observations is an exact function of the squares of past observations and previous variances. An alternative approach is to model volatility as an unobserved component in the variance. This leads to the class of *stochastic volatility* (SV) models. Taylor (1994) and Ghysels et al (1996) review the literature.

The stochastic volatility model has two attractions. The first is that it is the natural discrete time analogue (though it is only an approximation) of the continuous time model used in work on option pricing; see Hull and White (1987) and the review by Hang (1998). The second is that its statistical properties are relatively easy to determine. The disadvantage with respect to the conditional variance models of the GARCH class is that whereas GARCH can be estimated by maximum likelihood, the full treatment of an SV model requires the use of computer intensive methods such as MCMC and importance sampling. However, these methods are now quite rapid and it would be wrong to rule out SV models on the grounds that they make unreasonably heavy computational demands.

10.1 Basic specification and properties

The basic discrete time SV model for a demeaned series of returns, y_t, may be written as

$$= \sigma_t \varepsilon_t = \sigma \varepsilon_t e^{0.5h_t}, \qquad \varepsilon_t \sim IID(0,1), \quad t = 1,...,T, \tag{182}$$

where σ is a scale parameter and h_t is a stationary first-order autoregressive process, that is

$$h_{t+1} = \phi h_t + \eta_t, \qquad \eta_t \sim IID(0, \sigma_\eta^2) \tag{183}$$

where η_t is a disturbance term which may or may not be correlated with ε_t. If ε_t and η_t are allowed to be correlated with each other, the model can pick up the kind of asymmetric behaviour which is often found in stock prices.

The following properties of the SV model hold even if ε_t and η_t are contemporaneously correlated. Firstly y_t is a martingale difference. Secondly, stationarity of h_t implies stationarity of y_t. Thirdly, if η_t is normally distributed, terms involving exponents of h_t may be evaluated using properties of the lognormal distribution. Thus, the variance of y_t can be found and its kurtosis shown to be $\kappa_\varepsilon \exp(\ \sigma_h^2) > \kappa_\varepsilon$ where κ_ε is the kurtosis of ε_t. Similarly, the autocorrelations of powers of the absolute value of y_t, and its logarithm, can be derived; see Ghysels et al (1996).

10.2 Estimation

Squaring the observations in (182) and taking logarithms gives

$$\log y_t^2 = \omega + h_t + \xi_t, \tag{184}$$

where $\xi_t = \log \varepsilon_t^2 - E \log \varepsilon_t^2$ and $\omega = \log \sigma^2 + E \log \varepsilon_t^2$,so that ξ_t has zero mean by construction. If ε_t has a $t_\nu-$distribution, it can be shown that the moments of ξ_t exist even if the distribution of ε_t is Cauchy, that is $\nu = 1$. In fact in this case ξ_t is symmetric with excess kurtosis two, compared with excess kurtosis four and a highly skewed distribution when ε_t is Gaussian.

The transformed observations, the $\log y_t^{2\prime} s$, can be used to construct a linear state space model. The measurement equation is (184) while (183) is the transition equation. The quasi maximum likelihood (QML) estimators of the parameters ϕ, σ_η^2 and the variance of ξ_t, σ_ξ^2, are obtained by treating ξ_t and η_t as though they were normal in the linear SSF and maximizing the prediction error decomposition form of the likelihood obtained via the Kalman filter; see Harvey, Ruiz and Shephard (1994). Harvey and Shephard (1996) show how the linear state space form can be modified so as to deal with an asymmetric model. The QML method is relatively easy to apply and, even though it is not efficient, it provides a reasonable alternative if the sample size is not too small.

Simulation based methods of estimation are proposed in Jacquier, Polson and Rossi (1994, p 416), Kim, Shephard and Chib (1998), Watanabe (1999) and Durbin and Koopman (2000). Further discussion can be found in the chapter by Bollerslev et al.

10.3 Comparison with GARCH

The GARCH(1,1) model has been applied extensively to financial time series. The variance in $y_t = \sigma_t \varepsilon_t$ is assumed to depend on the variance and squared observation in the previous time period. Thus

$$\sigma_t^2 = \gamma + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2, \qquad t = 1, ..., T. \tag{185}$$

The GARCH(1,1) model displays similar properties to the SV model, particularly if ϕ is close to one (in which case $\alpha + \beta$ is also close to one). Jacquier et al (1994, p373) present a graph of the correlogram of the squared weekly returns of a portfolio on the New York Stock Exchange together with the ACFs implied by fitting SV and GARCH(1,1) models. The main difference in the ACFs seems to show up most at lag one with the ACF implied by the SV model being closer to the sample values.

The Gaussian SV model displays excess kurtosis even if ϕ is zero since y_t is a mixture of distributions. The σ_η^2 parameter governs the degree of mixing independently of the degree of smoothness of the variance evolution. This is not the case with a GARCH model where the degree of kurtosis is tied to the roots of the variance equation, α and β in the case of GARCH(1,1). Hence, it is very often necessary to use a non-Gaussian distribution for ε_t to capture the high kurtosis typically found in a financial time series. Kim, Shephard and Chib (1998) present strong evidence against the use of the Gaussian GARCH, but find GARCH$-t$ and Gaussian SV to be similar. In the exchange rate data they conclude on p 384 that the two models '...fit the data more or less equally well.' Further evidence on kurtosis is in Carnero, Pena and Ruiz (2004).

Fleming and Kirby (2003) compare the forecasting performance of GARCH and SV models. They conclude that '.. GARCH models produce less precise forecasts ....', but go on to observe that '... in the simulations, it is not clear that the performance differences are large enough to be economically meaningful.'

10.4 Extensions of the Model

Seasonality Other nonstationary components can easily be brought into h_t. For example, a seasonal or intra-daily component can be included; the specification is exactly as in the corresponding levels models. The state space formulation follows along the lines of the corresponding structural time series models for levels. Handling such effects is not so easy within the GARCH framework.

Long memory Breidt, Crato and de Lima (1998) and Harvey (1998) propose a long memory SV model in which h_t is generated by fractional noise.

SV with Markov Switching So, Lam and Li (1998) propose a model that enables the intercept coefficient in the volatility equation to switch between states so as to capture the effects associated with important events. They apply the model to S and P weekly returns with 3 states. The high volatility state is closely related to sudden abnormal fluctuations such as the 1987 crash.

Unobserved components with SV errors Bos and Koopman (2004) and Bos and Shephard (2003)

Multivariate Models The multivariate model corresponding to (182) assumes that each series is generated by a model of the form

$$y_{it} = \sigma_i \varepsilon_{it} e^{0.5h_{it}}, t = 1, ..., T, \tag{186}$$

with the covariance (correlation) matrix of the vector $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, ..., \varepsilon_{Nt})'$ being denoted by $\boldsymbol{\Sigma}_\varepsilon$. The vector of volatilities, $\mathbf{h}_t$, follows a VAR(1) process, that is

$$\mathbf{h}_{t+1}= \boldsymbol{\Phi}\mathbf{h}_t+\boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t\sim IID(\mathbf{0}, \boldsymbol{\Sigma}_\eta),$$

This specification allows the movements in volatility to be correlated across different series via $\boldsymbol{\Sigma}_\eta$. Interactions can be picked up by the off-diagonal elements of $\boldsymbol{\Phi}$. A simple nonstationary model is obtained by assuming that the volatilities follow a multivariate random walk, that is $\boldsymbol{\Phi} = \mathbf{I}$. If $\boldsymbol{\Sigma}_\eta$ is singular, of rank $K < N$, there are only K components in volatility, that is each h_{it} in (186) is a linear combination of $K < N$ common trends. Harvey, Ruiz and Shephard (1994) apply the nonstationary model to four exchange rates and find just two common factors driving volatility.

11 Conclusions

The principal STMs can be regarded as regression models in which the explanatory variables are functions of time and the parameters are time-varying. As such they provide a model based method of forecasting with an implicit weighting scheme that takes account of the properties of the time series and its salient features. The simplest procedures coincide with *ad hoc* methods that typically do well in forecasting competitions. For example the EWMA is rationalised as a random walk plus noise, though once non-Gaussian models are brought into the picture, the EWMA can also be obtained for distributions such as the Poisson and binomial.

Because of the interpretation in terms of components of interest, model selection of STMs does not rely on correlograms and related statistical devices. This is important, since it means that the models chosen are typically more robust to changes in structure as well as being less susceptible to the distortions caused by sampling error. Furthermore plausible models can be selected in situations where the observations are subject to data irregularities. Once a model has been chosen, problems like missing observations are easily handled within the state space framework Indeed, even irregularly spaced observations are easily dealt with as the principal STMs can be set up in continuous time and the implied discrete time SSF derived.

The STM framework can be adapted to produce forecasts - and ‘nowcasts’ - for a target series taking account of the information in an auxiliary series - possibly at a different sampling interval. Again the freedom from the model selection procedures needed for ARIMA models and the flexibility afforded by the SSF is of crucial importance.

As well as drawing attention to some of the attractions of STMs, the article has also set out some basic results for the SSF and derived some formulae linking models that can be put in the SSF with ARIMA and autoregressive representations. In a multivariate context, the VECM representation of a common trends STM is obtained.

Finally, it is pointed out how recent advances in computer intensive methods have opened up the way to dealing with non-Gaussian and nonlinear models. Such models may be motivated in a variety of ways: for example by the need to fit heavy tailed distributions in order to handle outliers and structural breaks in a robust fashion or by a complex nonlinear functional form suggested by economic theory.

Acknowledgements

I would like to thank Fabio Busetti, David Dickey, Siem Koopman, Ralph Snyder, Thomas Trimbur and participants at the San Diego conference in April 2004 for helpful comments on earlier drafts.

References

Anderson, B.D.O., and J.B. Moore (1979) *Optimal Filtering.* Englewood Cliffs: Prentice-Hall.

Andrews, R.C. (1994) "Forecasting Performance of Structural Time Series Models", *Journal of Business and Economic Statistics,* 12**,** 237-52.

Assimakopoulos, V. and K. Nikolopoulos, (2000) "The Theta Model: a Decomposition Approach to Forecasting", *International Journal of Forecasting, 16, 521-530.*

Bazen, S. and V. Marimoutou (2002) "Looking for a Needle in a Haystack? A Re-examination of the Time Series Relationship between Teenage Employment and Minimum Wages in the United States", *Oxford Bulletin of Economics and Statistics,* 64, 699-725.

Bergstrom, A.R. (1984) "Continuous Time Stochastic Models and Issues of Aggregation over Time", in Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, pp. 1145-1212. Amsterdam: North Holland.

Bos, C. and S.J.Koopman (2004) "Time Series Models with a Common Stochastic Variance for Analysing Economic Time Series", *Journal of Business and Economic Statistics***,** 22**,** 346-57

Bos, C. and N. Shephard (2003) "Inference for Adaptive Time Series Models: Stochastic Volatility and Conditionally Gaussian State Space Form", Mimeo

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

Box, G.E.P., D.A. Pierce and P. Newbold (1987). Estimating Trend and Growth Rates in Seasonal Time Series. *Journal of the American Statistical Association* 82: 276-82.

Breidt, F. J., Crato, N. and P. de Lima (1998). The Detection and Estimation of Long Memory in Stochastic Volatility. *Journal of Econometrics, 83: 325-48.*

Brown R.G. (1963). *Smoothing, Forecasting and Prediction.* Englewood Cliffs: Prentice Hall.

Bruce, A. G., and S. R. Jurke (1996). Non-Gaussian Seasonal Adjustment: X-12-ARIMA versus Robust Structural Models. *Journal of Forecasting* 15: 305-28.

Burridge, P. and K.F.Wallis (1988), Prediction Theory for Autoregressive-Moving Average Processes. *Econometric Reviews,* 7, 65-9.

Busetti, F.and A.C. Harvey (2003), Seasonality tests. *Journal of Business and Economic Statistics, 21, 420-36.*

Canova, F and E Ghysels (1994). Changes in Seasonal Patterns. Are they cyclical? *Journal of Economic Dynamics and Control,* 18, 1143-1172

Canova, F., and B.E. Hansen (1995), Are Seasonal Patterns Constant over Time? A Test for Seasonal Stability. *Journal of Business and Economic Statistics* 13**:** 237-52.

Carnero, M A, Pena, D and E Ruiz (2004) Persistence and Kurtosis in GARCH and Stochastic Volatility Models *Journal of Financial Econometrics*,, 2, 319-342.

Carter, C. K., and R. Kohn (1996). Markov Chain Monte Carlo in Conditionally Gaussian State Space Models. *Biometrika* 83: 589-601.

Carvalho, V.M and A.C. Harvey (2002). Growth, cycles and convergence in US regional time series. DAE Working paper 0221, University of Cambridge.

Chambers, M J and J McGarry (2002). Modeling Cyclical Behaviour with Differential-Difference Equations in an Unobserved Components framework. *Econometric Theory* 18: 387-419.

Chatfield, C., Koehler, A.B., Ord, J.K, and R.D. Snyder (2001). A New Look at Models for Exponential Smoothing. *The Statistician*, 50, 147-59.

Chow, G.C. (1984). Random and Changing Coefficient Models. In Z. Griliches and M. Intriligator (eds.) *Handbook of Econometrics*, vol2, pp. 1213-45. Amsterdam: North Holland.

Clements, M.P. and D.F. Hendry (1998). Forecasting Economic Time Series. Cambridge: Cambridge University Press.

Clements, M.P. and D.F. Hendry (2003). Economic Forecasting: Some Lessons from Recent Research. *Economic Modelling 20, 301-329.*

Dagum, E.B., Quenneville, B. and Sutradhar, B. (1992) Trading-day Multiple Regression Models with Random Parameters. *International Statistical Review,* 60, 57-73.

Davidson, J., D.F. Hendry, F. Srba and S. Yeo (1978). Econometric Modelling of the Aggregate Time-Series Relationship between Consumers' Expenditure and Income in the United Kingdom. *Economic Journal* 88: 661-92.

de Jong, P. and N.Shephard (1995). The Simulation Smoother for Time Series Models. *Biometrika* 82: 339-50.

Durbin, J. and B. Quenneville (1997). Benchmarking by State Space Models. *International Statistical Review* 65: 23-48.

Durbin, J., and S.J. Koopman (2000). Time Series Analysis of Non-Gaussian Observations based on State-Space Models from Both Classical and Bayesian Perspectives (with discussion). *Journal of Royal Statistical Society, Series* B 62: 3-56.

Durbin, J., and S.J. Koopman (2001). *Time series analysis by state space methods.* Oxford University Press, Oxford.

Durbin, J., and S.J. Koopman (2002). A simple and efficient simulation smoother for state space time series models. *Biometrika* 89: 603–16.

Engle, R. and S. Kozicki (1993) Testing for Common Features. *Journal of Business and Economic Statistics* 11, 369-80.

Fleming, J and C Kirby (2003) A Closer Look at the Relation between GARCH and Stochastic Autoregressive Volatility. *Journal of Financial Econometrics*, 1, 365-419.

Franses, P. H. and R. Papp (2004). *Periodic time series models.* Oxford University Press, Oxford.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis 15:* 183-202.

Frühwirth-Schnatter, S. (2004). Efficient Bayesian parameter estimation, in *State Space and Unobserved Component Models,* ed Harvey, A.C. *et al.*, 123-51. Cambridge University Press, Cambridge.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*, 2nd edition. John Wiley and Sons, New York.

Ghysels, E., A.C.Harvey and E.Renault (1996). Stochastic volatility, in G.S.Maddala and C.R.Rao (eds). *Handbook of Statistics,* vol 14, 119-192.

Godolphin, E., and M. Stone (1980). On the Structural Representation for Polynomial Predictor Models. *Journal of the Royal Statistical Society, Series B* 42: 35-45.

Grunwald, G. K., K Hamza and R. J. Hyndman (1997). Some Properties and Generalizations of Non-negative Bayesian Time Series Models. *Journal of the Royal Statistical Society, Series B* 59: 615-626.

Hamilton, J.D. (1989), A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357-84.

Hang, J.J. (1998). Stochastic Volatility and Option Pricing, in J.Knight and S.Satchell, eds. *Forecasting Volatility, 47-96.* Oxford: Butterworth-Heinemann.

Hannan, E.J., R.D. Terrell and N. Tuckwell (1970). The seasonal adjustment of economic time series. *International Economic Review* 11: 24-52.

Harrison, P.J., and C.F. Stevens (1976). Bayesian Forecasting. *Journal of the Royal Statistical Society, Series B* 38: 205-47.

Harvey, A.C. (1984). A unified view of statistical forecasting procedures [with discussion]. *Journal of Forecasting* 3: 245-83.

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

Harvey A.C. (1998). Long memory in stochastic volatility, in J.Knight and S.Satchell, eds. *Forecasting Volatility, 307-20.* Oxford: Butterworth-Heinemann.

Harvey A.C. (2001). Testing in unobserved components models, *Journal of Forecasting,* 20**,** 1-19.

Harvey, A.C., and C-H. Chung (2000). Estimating the Underlying Change in Unemployment in the UK (with discussion), *Journal of the Royal Statistical Society, Series A,* 163: 303-39.

Harvey A.C. and C. Fernandes (1989). Time Series Models for Count Data or Qualitative Observations. *Journal of Business and Economic Statistics,* 7, 409-422.

Harvey, A.C., and A. Jaeger (1993). Detrending, Stylised Facts and the Business Cycle. *Journal of Applied Econometrics* 8: 231-47.

Harvey, A.C., and S.J.Koopman (1992). Diagnostic Checking of Unobserved Components Time Series Models. *Journal of Business and Economic Statistics 10**:** 377-89..*

Harvey, A. C., and S. J. Koopman (1993). Forecasting Hourly Electricity Demand Using Time-Varying Splines. *Journal of American Statistical Association* 88: 1228-36.

Harvey, A.C., and S.J Koopman (2000). Signal Extraction and the Formulation of Unobserved Components Models, *Econometrics Journal,* 3**,** 84-107.

Harvey, A.C., S. J. Koopman, and M. Riani (1997). The Modeling and Seasonal Adjustment of Weekly Observations. *Journal of Business and Economic Statistics* 15: 354-68.

Harvey A.C., E. Ruiz and N. Shephard (1994). Multivariate Stochastic Variance Models. *Review of Economic Studies* 61: 247-64.

Harvey, A.C., and A. Scott (1994). Seasonality in Dynamic Regression Models. *Economic Journal* 104: 1324-45

Harvey A. C., and N. Shephard, (1996). Estimation of an asymmetric stochastic volatility model for asset returns. *Journal of Business and Economic Statistics,* 14: 429-34.

Harvey A.C., and R.D.Snyder (1990). Structural Time Series Models in Inventory Control. *International Journal of Forecasting, 6,187-98.*

Harvey A. C., and P.H.J. Todd (1983). Forecasting economic time series with structural and Box-Jenkins models [with discussion]. *Journal of Business and Economic Statistics* 1: 299-315.

Harvey, A.C. and T. Trimbur (2003). General model-based filters for extracting cycles and trends in economic time series. *Review of Economics and Statistics, 85, 244-55.*

Harvey, A.C., T. Trimbur and H. van Dijk (2003). Cyclical components in economic time series: a Bayesian approach. DAE discussion paper 0302, Cambridge.

Hillmer, S.C. (1982). Forecasting Time Series with Trading Day Variation. *Journal of Forecasting* 1: 385-95.

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.

Hipel, R. W., and McLeod, A. I. (1994). *Time Series Modelling of Water Resources and Environmental Systems.* Developments in Water Science, 45, Elsevier, Amsterdam.

Holt, C.C. (1957). Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages. ONR Research Memorandum 52, Carnegie Institute of Technology, Pittsburgh, Pennsylvania.

Hull, J. and A. White, (1987). The Pricing of Options on Assets with Stochastic Volatilities. *Journal of Finance.* 42: 281-300.

Hyndman R. J., and Billah B. (2003). Unmasking the Theta method. *International Journal of Forecasting 19, 287-290.*

Ionescu, V., Oara, C., and M.Weiss, (1997), General Matrix Pencil Techniques for the Solution of Algebraic Riccati Equations: A Unified Approach. *IEEE Transactions in Automatic Control,* 42, 1085-97.

Jacquier, E., Polson, N.G., and P.E. Rossi, (1994). Bayesian analysis of stochastic volatility models (with discussion). *Journal of Business and Economic Statistics* 12**:** 371-417.

Johnston, F.R., and P.J. Harrison (1986). The Variance of Lead Time Demand. *Journal of the Operational Research Society* 37: 303-8.

Jones, R.H.(1993). *Longitudinal Data with Serial Correlation: A State Space Approach.* London: Chapman and Hall.

Kalman, R.E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering, Transactions ASME. Series D* 82: 35-45.

Kim, C. J. and C. Nelson (1999). *State-Space Models with Regime-Switching.* Cambridge MA: MIT Press.

Kim, S., Shephard, N.S. and S.Chib (1998). Stochastic Volatility: Likelihood Inference and Comparison with ARCH models. *Review of Economic Studies* 65: 361-93.

Kitagawa, G. (1987). Non-Gaussian State Space Modeling of Nonstationary Time Series [with discussion]. *Journal of the American Statistical Association* 82: 1032-63.

Kitagawa, G., and W Gersch (1996). *Smoothness Priors Analysis of Time Series.* Berlin: Springer-Verlag.

Koop, G. and van Dijk H.K.(2000), Testing for Integration using Evolving Trend and Seasonals Models: A Bayesian Approach, *Journal of Econometrics* 97, 261-91.

Koopman, S.J. and Harvey, A.C. (2003) Computing Observation Weights for Signal Extraction and Filtering. *Journal of Economic Dynamics and Control*, 27, 1317-33.

Koopman, S.J., A.C. Harvey, J.A. Doornik and N. Shephard (2000). *STAMP 6.0 Structural Time Series Analyser, Modeller and Predictor*, London: Timberlake Consultants Ltd.

Kozicki, S. (1999). Multivariate detrending under common trend restrictions: Implications for business cycle research. *Journal of Economic Dynamics and Control*, 23, 997-1028.

Krane, S. and W. Wascher (1999). The cyclical sensitivity of seasonality in U.S. employment. *Journal of Monetary Economics,* 44, 523-53.

Kuttner, K.N. (1994). Estimating potential output as a latent variable. *Journal of Business and Economic Statistics,* 12, 361-68.

Lenten, L.J.A., and I.A. Moosa. (1999) Modelling the Trend and Seasonality in the Consumption of Alcoholic Beverages in the United Kingdom. *Applied Economics*, 1999, 31, 795-804.

Luginbuhl, R. and A. de Vos (1999). Bayesian analysis of an unobserved components time series model of GDP with Markov-switching and time-varying growths. *Journal of Business and Economic Statistics* 17, 456-65.

MacDonald, I. L., and W. Zucchini (1997). *Hidden Markov Chains and Other Models for Discrete-Valued Time Series,* London: Chapman and Hall.

Makridakis, S., and Hibon, M. (2000). The M3-Competitions: Results, Conclusions and Implications. *International Journal of Forecasting, 16, 451-476.*

Maravall A.(1985). On structural time series models and the characterization of components. *Journal of Business and Economic Statistics* 3: 350-5.

Moosa, I.A., and P. Kennedy. (1998) Modelling Seasonality in the Australian Consumption Function. *Australian Economics Paper*, 37, 88-102.

Morley, J.C., C.R. Nelson, and Zivot, E. (2003). Why are Beveridge-Nelson and unobserved components decompositions of GDP so different? *Review of Economic and Statistics, 85, 235-24.*

Muth, J.F. (1960). Optimal Properties of Exponentially Weighted Forecasts. *Journal of the American Statistical Association* 55: 299-305.

Nerlove, M. and S. Wage (1964). On the Optimality of Adaptive Forecasting. *Management Science* 10: 207-29.

Nerlove, M., D.M. Grether and J.L. Carvalho (1979). *Analysis of Economic Time Series*. New York: Academic Press.

Nicholls, D.F., and A.R. Pagan (1985). Varying Coefficient Regression. In E.J. Hannan, P.R. Krishnaiah and M.M. Rao (eds.) *Handbook of Statistics,* vol. 5, pp. 413-50. Amsterdam: North Holland.

Ord, J.K, A.B. Koehler and R.D. Snyder (1997). Estimation and prediction for a class of dynamic nonlinear statistical model, *Journal of the American Statistical Association*, 92 :1621-1629.

Osborn, D. R and J. R. Smith (1989). The performance of periodic autoregressive models in forecasting U.K consumption. *Journal of Business and Economic Statistics*, 7, 117-27.

Patterson, K.D. (1995). An Integrated Model of the Date Measurement and Data Generation Processes with an Application to Consumers' Expenditure. *Economic Journal*, 105, 54-76.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics* 9: 163-75.

Planas, C. and A. Rossi (2004). Can inflation data improve the real-time reliability of output gap estimates ? *Journal of Applied Econometrics, 19, 121-33.*

Proietti, T. (1998). Seasonal Heteroscedasticity and Trends, *Journal of Forecasting* 17: 1-17.

Proietti, T. (2000). Comparing Seasonal Components for Structural Time Series Models, *International Journal of Forecasting,* 16: 247-60.

Quenneville, B. and Singh, A.C. (2000) Bayesian Prediction MSE for State Space Models with Estimated Parameters. *Journal of Time Series Analysis,* **21,** 219-36.

Rosenberg, B. (1973). Random Coefficient Models: the Analysis of a Cross-Section of Time Series by Stochastically Convergent Parameter Regression. *Annals of Economic and Social Measurement* 2: 399-428.

Schweppe, F. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Transactions on Information Theory* 11: 61-70.

Smith, R.L., and J.E. Miller (1986). A non-Gaussian state space model and application to prediction of records. *Journal of the Royal Statistical Society, Series B* 48: 79-88.

Snyder, R.D. (1984). Inventory Control with the Gamma Probability Distribution. *European Journal of Operational Research* 17: 373-81.

So, M., Lam, K. and W. K. Li (1998) A Stochastic Volatility Model with Markov Switching. *Journal of Business and Economic Statistics* 15**:** 244-53.

Stoffer, D. and K. Wall (2004). Resampling in State Space Models, in *State Space and Unobserved Component Models,* ed Harvey, A.C. *et al.*, 171-202. Cambridge University Press, Cambridge.

Taylor, S. J. (1994). Modelling stochastic volatility. *Mathematical Finance* 4: 183-204.

Trimbur, T (2004). Properties of higher order stochastic cycles. *Journal of Time Series Analysis.* (to appear)

Watanabe, T. (1999). A Non-Linear Filtering Approach to Stochastic Volatility Models with an Application to Daily Stock Returns. *Journal of Applied Econometrics,* 14: 101-21.

West, M. and P.J.Harrison (1989). *Bayesian Forecasting and Dynamic Models.* New York: Springer-Verlag.

Winters, P.R. (1960). Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science* 6: 324-42.

Young, P. (1984). *Recursive Estimation and Time-Series Analysis.* Berlin: Springer-Verlag.