Journal of Forecasting, Vol. 9, 89–108 (1990)

Estimation Procedures for Structural Time Series Models

A. C. HARVEY
London School of Economics, UK

S. PETERS
Monash University, Australia

ABSTRACT

A univariate structural time series model based on the traditional decomposition into trend, seasonal and irregular components is defined. A number of methods of computing maximum likelihood estimators are then considered. These include direct maximization of various time domain likelihood function. The asymptotic properties of the estimators are given and a comparison between the various methods in terms of computational efficiency and accuracy is made. The methods are then extended to models with explanatory variables.

KEY WORDS Structural time series model Forecasting Kalman filter Stochastic trend Unobserved components model EM algorithm

INTRODUCTION

A univariate economic time series model can be formulated directly in terms of the traditional components of trend, seasonal, cycle and irregular. A model of this kind is called a structural time series model. The attraction of such a formulation is that it has an immediate interpretation and so is a natural vehicle for making forecasts. Harvey and Todd (1983) compare the forecasts made by a basic form of the structural model with the forecasts made by ARIMA models, and conclude that there may be strong arguments in favour of using structural models in practice. In another paper, Harvey (1985) shows how structural models can be used to model cycles in macroeconomic time series. Other studies include Kitagawa and Gersch (1984).

The forecasts obtained from a particular structural model depend on certain variance parameters. These parameters play a similar role to the autoregressive and, more particularly, the moving-average parameters in an ARIMA model. The aim of the present paper is to set out various methods for computing the maximum likelihood estimators of these parameters. The methods are then compared in terms of sample properties and computational tractability. Note that Harrison and Stevens (1976) make structural models the basis of their Bayesian approach, but tend to fix the variance parameters *a priori* rather than to estimate them.

0277–6693/90/020089–20$10.00

Received June 1988
Revised March 1989

The class of structural models is introduced briefly in the next section. It is shown how the state space form can be used to handle the model and to make predictions of future observations. The third section considers ML estimation in the time domain, with attention being focused on direct maximization of the prediction error decomposition form of the likelihood function and on indirect maximization via the EM algorithm. The role of initial conditions for the Kalman filter is explored in some detail. Maximum likelihood estimation in the frequency domain is discussed in the fourth section, while the asymptotic properties of the estimators are set out in the fifth. The sixth section reports practical experience with the estimators. The estimation procedures are extended to models with explanatory variables in the seventh section and new developments outlined in the final section.

STRUCTURAL TIME SERIES MODELS

The essence of a structural model is that it is formulated in terms of independent components which have a direct interpretation in terms of quantities of interest. One of the most important models for economic time series is the basic structural model: this consists of a trend, a seasonal and an irregular component. Our computational results are centred on this model, though they have clear implications for other models within the structural class.

The basic structural model is

$$y_t = \mu_t + \gamma_t + \varepsilon_t \qquad t = 1, \ldots, T \tag{1}$$

where μ_t, γ_t and ε_t are the trend, seasonal and irregular components, respectively.

The process generating the trend can be regarded as a local approximation to a linear trend, i.e.

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t \tag{2a}$$

$$\beta_t = \beta_{t-1} + \zeta_t \qquad t = 1, \ldots, T \tag{2b}$$

where η_t and ζ_t are distributed independently of each other and over time with mean zero and variances σ_η^2 and σ_ζ^2, respectively. The process generating the seasonal component is

$$\gamma_t = -\sum_{j=1}^{s-1} \gamma_{t-j} + \omega_t \qquad t = 1, \ldots, T \tag{3}$$

where ω_t is an independently distributed disturbance term with mean zero and variance σ_ω^2 and s is the number of 'seasons' in the year. The seasonal pattern is therefore slowly changing, but by a mechanism which ensures that the sum of the seasonal components over any s consecutive time periods has an expected value of zero and a variance which remains constant over time. Writing out equation (3) in terms of the lag operator, L, gives

$$(1 + L + \cdots + L^{s-1})\gamma_t = S(L)\gamma_t = \omega_t \qquad t = 1, \ldots, T \tag{4}$$

However, since

$$\Delta_s = 1 - L^s = (1 + L + \cdots + L^{s-1})(1 - L) = S(L)\Delta \tag{5}$$

the model can also be expressed in terms of the seasonal difference operator as

$$\Delta_s \gamma_t = (1 - L)\omega_t \tag{6}$$

Unlike the trend and seasonal components, the irregular component is assumed to be

stationary. In the basic model it is a white-noise disturbance term with mean zero and variance, σ_ε^2.

Multiplying equation (1) through by the first and seasonal difference operators renders the model stationary, i.e.

$$\Delta\Delta_s y_t = \Delta_s \eta_t + S(L)\zeta_{t-1} + \Delta^2\omega_t + \Delta\Delta_s\varepsilon_t \qquad t = s+2,\ldots,T \tag{7}$$

The right-hand side is an MA($s+1$) process, but with a number of non-linear restrictions on the parameters. The ARIMA model corresponding to a particular structural model is called the reduced or canonical form; see Engle (1978) and Nerlove *et al.* (1979).

State space, prediction and smoothing

The state space form of a univariate time series model consists of a transition equation

$$\boldsymbol{\alpha}_t = \mathbf{T}_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t \qquad t = 1,\ldots,T \tag{8a}$$

and a measurement equation

$$y_t = \mathbf{z}_t'\boldsymbol{\alpha}_t + \varepsilon_t \qquad t = 1,\ldots,T \tag{8b}$$

in which $\boldsymbol{\alpha}_t$ is an $m \times 1$ state vector, $\mathbf{z}_t$ is an $m \times 1$ fixed vector, $\mathbf{T}_t$ is a fixed matrix of order $m \times m$ and ε_t and $\boldsymbol{\eta}_t$ are, respectively, a scalar disturbance term and an $m \times 1$ vector of disturbances which are distributed independently of each other. It is assumed that ε_t is white noise with mean zero and variance, h_t, and $\boldsymbol{\eta}_t$ is multivariate white noise with mean vector zero and covariance matrix $\mathbf{Q}_t$. In the models considered here, ε_t and $\boldsymbol{\eta}_t$ will also be assumed to be normally distributed.

Let $\mathbf{a}_{t-1}$ be the minimum mean square estimator (MMSE) of $\boldsymbol{\alpha}_{t-1}$ at time $t-1$ and let $\mathbf{P}_{t-1}$ be the covariance matrix of the estimation error, $\mathbf{a}_{t-1} - \boldsymbol{\alpha}_{t-1}$. When y_t becomes available, $\mathbf{a}_{t-1}$ and $\mathbf{P}_{t-1}$ can be updated by the Kalman filter. If $\boldsymbol{\alpha}_0 \sim N(\mathbf{a}_0, \mathbf{P}_0)$ and $\mathbf{a}_0$ and $\mathbf{P}_0$ are known, the Kalman filter produces a set of T one-step-ahead prediction errors v_t, $t = 1,\ldots,T$, together with their variances, f_t. Estimates of future observations, together with their MSEs, can be made using the Kalman filter, while MMSEs of the elements of each $\boldsymbol{\alpha}_t$ based on all the observations can be computed by a smoothing algorithm; see Anderson and Moore (1979) and Harvey (1981a, Chapter 4).

The basic structural model can be put in state space form very easily. Suppose, for simplicity, that $s = 4$. The transition equation is then

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \mu_t \\ \beta_t \\ \hline \gamma_t \\ \gamma_{t-1} \\ \gamma_{t-2} \end{bmatrix} = \left[\begin{array}{cc:ccc} 1 & 1 & & 0 & \\ 0 & 1 & & & \\ \hdashline & & -1 & -1 & -1 \\ & 0 & 1 & 0 & 0 \\ & & 0 & 1 & 0 \end{array}\right] \begin{bmatrix} \mu_{t-1} \\ \beta_{t-1} \\ \hline \gamma_{t-1} \\ \gamma_{t-2} \\ \gamma_{t-3} \end{bmatrix} + \begin{bmatrix} \eta_t \\ \zeta_t \\ \omega_t \\ 0 \\ 0 \end{bmatrix} \tag{9a}$$

while the measurement equation is, from equation (2),

$$y_t = (1\,0\,1\,0\,0)\boldsymbol{\alpha}_t + \varepsilon_t \qquad t = 1,\ldots,T \tag{9b}$$

Both h_t and $\mathbf{Q}_t$ are time invariant with $h_t = \sigma_\varepsilon^2$ and $\mathbf{Q}_t = \text{diag.}\,\{\sigma_\eta^2, \sigma_\zeta^2, \sigma_\omega^2, 0, 0\}$. In order to run the Kalman filter it is necessary that these parameters be known, and it is the question of how they should be estimated which is taken up in the next section onwards. Another important question concerns starting values for the Kalman filter. Since the components of the state vector are non-stationary, $\mathbf{P}_0$ cannot simply be taken to be the covariance matrix of a stationary vector

AR(1) process as in Gardner *et al.* (1980). There are several solutions and these are discussed in the next section. However, one possibility is to let α_0 have a diffuse prior, i.e. $\mathbf{P}_0 = \kappa\mathbf{I}$ where $\kappa = \infty$. In practice, κ may be set equal to a large but finite number.

The state space model in equation (9) is time invariant since $\mathbf{T}_t$, $\mathbf{z}_t$, $\mathbf{Q}_t$ and h_t do not change over time. When a time invariant state space model is completely detectable and completely stabilizable, it can be shown that the Kalman filter converges to a steady state in the sense that $\mathbf{P}_t$ becomes a time-invariant positive definite matrix $\bar{\mathbf{P}}$; see Chan *et al.* (1984).

MAXIMUM LIKELIHOOD ESTIMATION IN THE TIME DOMAIN

If the disturbances ε_t and η_t in equation (8) are normally distributed, the likelihood function for the observations may be obtained from the Kalman filter *via* the prediction error decomposition. This may then be maximized with respect to any unknown parameters contained in $\mathbf{T}_t$, $\mathbf{z}_t$, $\mathbf{Q}_t$ or h_t using a suitable numerical optimization procedure. However, within the present context there are a number of different ways of defining the likelihood function, depending on the assumptions made about initial conditions. All are equivalent asymptotically, although they may have very different small sample properties.

The first three methods described below are all based on direct maximization of the likelihood function. As regards initial conditions, the first assumes a diffuse prior for α_0, the second assumes α_0 is fixed (but unknown) while the third assumes a steady-state Kalman filter from the outset. After a brief subsection on estimation *via* the reduced form ARIMA model, the relationship between the various methods is discussed and a parallel is drawn with ML procedures for ARIMA models, i.e. exact ML, conditional sum of squares (CSS) and so on. The last time domain ML estimation procedure described is the EM algorithm. This provides an alternative approach to computing ML estimates in which the likelihood function is maximized indirectly as a result of a stepwise procedure.

The basic structural model contains four unknown variance parameters, σ_η^2, σ_ξ^2, σ_ω^2 and σ_ε^2. However, non-linear optimization need only be carried out with respect to three of these since the fourth may be concentrated out of the likelihood function. This is done by redefining h_t and $\mathbf{Q}_t$ and hence $\mathbf{P}_t$. The Kalman filter can then be run as a function of the relative values of the three parameters, and the fourth estimated from the prediction errors at the end. This may be done whatever estimation procedure is adopted.

In describing the first three methods it will be assumed that the state vector contains k non-stationary elements. For the basic structural model $k = s + 1$.

Prediction error decomposition I: diffuse prior

If α_0 is taken to have a diffuse prior, starting values can be constructed from the first k observations, i.e. $\mathbf{a}_k$ and $\mathbf{P}_k$. The likelihood function for the observations $y_{k+1}, \ldots, y_T$ can then be defined conditional on $y_1, \ldots, y_k$. In terms of the prediction error decomposition the likelihood is

$$\log L = \frac{-(T-k)}{2}\log 2\pi - \frac{1}{2}\sum_{t=k+1}^{T}\log f_t - \frac{1}{2}\sum_{t=k+1}^{T} v_t^2/f_t \qquad (10)$$

where v_t is the one-step-ahead prediction error and f_t is its variance. Both quantities are given by the Kalman filter. Further justification for setting up the likelihood function in this way can be found in de Jong (1988).

There are several ways of computing equation (10). The easiest, both conceptually and

computationally, is to initialize the Kalman filter at $t = 0$ with $\mathbf{a}_0 = \mathbf{0}$ and $\mathbf{P}_0 = \kappa\mathbf{I}$, where κ is a large but finite number. Close approximations to $\mathbf{a}_k$ and $\mathbf{P}_k$ are then obtained after k iterations of the Kalman filter. Alternatively, $\mathbf{a}_k$ and $\mathbf{P}_k$ may be computed explicitly by writing out equations for the first k observations in terms of $\boldsymbol{\alpha}_k$. Computing $\mathbf{a}_k$ and $\mathbf{P}_k$ is then an exercise in generalized least squares (GLS); see Harvey (1982) for further details. More general ways of avoiding the 'large κ' approximation include the methods due to Ansley and Kohn (1985) and de Jong (1988).

The above methods can all be generalized to cover cases where the state vector contains a subvector of elements generated by a stationary process with a known mean and a covariance matrix $\mathbf{V}$. The 'large κ' method, for example, has

$$\mathbf{P}_0 = \begin{bmatrix} \kappa\mathbf{I} & 0 \\ 0 & \mathbf{V} \end{bmatrix} \tag{11}$$

Finally, if the Kalman filter is monitored, it is possible to switch to steady-state recursions once a steady state is reached. A device of this kind was used with some success in Gardner *et al.* (1980).

Prediction error decomposition II: fixed initial state vector

If the initial state vector, $\boldsymbol{\alpha}$ is taken to be fixed, the observations in equation (8) may be expressed in terms of $\boldsymbol{\alpha}_0$ by repeated substitution. Thus, with $\mathbf{T}_t$ assumed time invariant for simplicity

$$y_t = \mathbf{x}_t'\boldsymbol{\alpha}_0 + w_t \qquad t = 1, \ldots, T \tag{12}$$

where $\mathbf{x}_t = \mathbf{T}'\mathbf{z}_t$ and

$$w_t = \mathbf{z}_t' \sum_{j=1}^{t} \mathbf{T}^{t-j}\boldsymbol{\eta}_j + \varepsilon_t \qquad t = 1, \ldots, T \tag{13}$$

The disturbance term, w_t, has mean zero and a covariance matrix, $\boldsymbol{\Omega}_0$, which in the basic structural model depends on σ_η^2, α_ξ^2, σ_ω^2 and σ_ε^2; compare Franzini and Harvey (1983). The ML estimator of $\boldsymbol{\alpha}_0$, $\hat{\boldsymbol{\alpha}}_0$, conditional on the other parameters in the model can be computed by GLS. It can therefore be concentrated out of the likelihood function so that

$$\log L = -\frac{T}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Omega}_0| - \frac{1}{2}(\mathbf{y} - \mathbf{x}_0\hat{\boldsymbol{\alpha}}_0)'\boldsymbol{\Omega}_0^{-1}(\mathbf{y} - \mathbf{x}_0\hat{\boldsymbol{\alpha}}_0) \tag{14}$$

where

$$\mathbf{y} = \mathbf{x}_0\boldsymbol{\alpha}_0 + \mathbf{w}, \qquad \mathbf{V}(\mathbf{w}) = \boldsymbol{\Omega}_0 \tag{15}$$

denotes equation (12) written in matrix terms, with $\mathbf{y}$ a $T \times 1$ vector and so on; compare Sarris (1973, 505–8). Fortunately, the repeated construction and inversion of the $T \times T$ matrix, $\boldsymbol{\Omega}_0$, can be avoided by using the methods of Rosenberg (1973) or Wecker and Ansley (1983). Both methods are based on the state space form.

Prediction error decomposition III: steady-state Kalman filter

A third possibility for a time-invariant model is to compute $\mathbf{a}_k$ from the first k observations as in method I, but to set the covariance matrix of the estimation error, $\mathbf{P}_k$, equal to the steady-state covariance matrix $\bar{\mathbf{P}}$. The Kalman filter can then be run without the recursions for the $\mathbf{P}_t$ matrix, and the likelihood function can be expressed in the prediction error decomposition form (10). However, since f_t is time invariant, concentrating a parameter, say σ_ε^2, out of the

likelihood in the basic structural model leads to ML estimates being computed by minimizing the sum of squares function

$$S = \sum_{t=k+1}^{T} v_t^2 \tag{16}$$

In order to run the steady-state Kalman filter it is, however, necessary to first solve the Riccati equations to compute $\bar{\mathbf{P}}$. A number of ways of doing this are given in Anderson and Moore (1979, 156–8).

Estimation via the reduced form

As noted below equation (7), for the basic structural model, $\Delta\Delta_s y_t$ will follow an ARMA process in which the parameters are subject to a number of non-linear constraints. The likelihood function of this ARMA process can be computed in a number of ways, but whichever is adopted it must be maximized with respect to the original parameters in the model. Nerlove *et al.* (1979, 125–31) discuss methods for computing the ARMA parameters from the original parameters and vice versa. These appear to be relatively complex and time consuming, in general, although some simplification is possible for certain structural time series models. In such cases exact ML of the reduced form ARMA model via an efficient algorithm, such as that of Mélard (1984), may yield a competitive estimation procedure. Approximate ML estimation based on minimizing the conditional sum of squares (CSS) function of the ARMA model may also have some attraction: note that the minimizing CSS function is equivalent to direct estimation via method III above.

Relationship between different methods

Prediction error decomposition I corresponds to the usual definition of the exact likelihood function for an ARIMA model. Method II gives the exact likelihood function if $\boldsymbol{\alpha}_0$ is regarded as being fixed, but there seems to be no theoretical reason for preferring this assumption to the assumption implicit in method I. It is interesting to note, however, that the numerical values of the two likelihood functions (10) and (14) only differ because of their determinantal terms. It is shown in Appendix 2 that the sum of squares terms are identical.

The three ML estimators defined above all have the same asymptotic properties. However, as is well known from studies of ARMA models, the small sample properties of, say, CSS and exact ML procedures, may differ considerably; see *inter alia*, Harvey (1981a, Chapter 6) and Ansley and Newbold (1980).

Finally, note that the method used by Kitagawa (1981) contains elements of all three methods I–III. He first computes an estimator of $\boldsymbol{\alpha}_0$ by smoothing, using a diffuse prior starting value and an information filter. He then uses this estimator together with the steady-state covariance matrix $\bar{\mathbf{P}}$ as starting values for a second, steady-state, Kalman (information) filter which gives a likelihood function in terms of prediction errors from $t = 1$ to T.

EM algorithm

Watson and Engle (1983) have used the EM algorithm to estimate the unknown parameters in an unobserved components model, and the algorithm for the basic structural model is essentially a special case of the one they use. The application of the EM algorithm to a state space model requires the MMSEs of the state vector based on all the observations up to time T. These estimates will be denoted by $\mathbf{a}_{t|T}$, $t = 1, \ldots, T$, while the covariance matrix of the corresponding estimation errors will be denoted by $\mathbf{P}_{t|T}$, $t = 1, \ldots, T$. Both quantities can be computed by applying a fixed interval smoothing algorithm. This requires a forward and a

backward pass through the data. The forward pass is simply the Kalman filter. As it is applied, $\mathbf{a}_t$ and $\mathbf{P}_t$ are stored and used when the backward, smoothing, pass is made. An alternative way of carrying out these calculations is by means of an information filter as in Kitagawa (1981).

Consider the state space model with a time invariant covariance matrix $\mathbf{Q}$, and suppose, initially, that $\mathbf{Q}$ is unrestricted and that $\boldsymbol{\alpha}_0 \sim N(\mathbf{a}_0, \mathbf{P}_0)$ with $\mathbf{a}_0$ and $\mathbf{P}_0$ known. If the elements in the state vector for $t = 0, \ldots, T$ were observed, the log-likelihood function for the y_t's and $\boldsymbol{\alpha}_t$'s would be

$$\log L(\mathbf{y}, \boldsymbol{\alpha}) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^{T}(y_t - \mathbf{z}_t'\boldsymbol{\alpha}_t)^2$$

$$-\frac{Tn}{2}\log 2\pi - \frac{T}{2}\log|\mathbf{Q}| - \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{\alpha}_t - \mathbf{T}_t\boldsymbol{\alpha}_{t-1})'\mathbf{Q}^{-1}(\boldsymbol{\alpha}_t - \mathbf{T}_t\boldsymbol{\alpha}_{t-1})$$

$$-\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{P}_0| - \frac{1}{2}(\boldsymbol{\alpha}_0 - \mathbf{a}_0)'\mathbf{P}_0^{-1}(\boldsymbol{\alpha}_0 - \mathbf{a}_0) \quad (17)$$

Let the unknown parameters in the model be denoted by the vector $\boldsymbol{\psi}$. The EM algorithm proceeds iteratively by evaluating

$$E\left\{\frac{\partial \log L(\mathbf{y})}{\partial \boldsymbol{\psi}} \,\middle|\, \mathbf{y}\right\} \quad (18)$$

conditional on the latest estimate of $\boldsymbol{\psi}$. The expression is then set equal to a vector of zeros and solved to yield a new set of estimates of $\boldsymbol{\psi}$. The procedure is repeated until convergence. It can be shown that, under suitable conditions, the likelihood will remain the same or increase at each iteration and it will converge to a local maximum; see Dempster *et al.* (1977), Wu (1983) and Boyles (1983).

Applying equation (18) to the likelihood function in equation (17) gives the following expression for the estimator[1] of σ_ε^2,

$$\hat{\sigma}_\varepsilon^2 = T^{-1}\sum_{t=1}^{T}\{e_{t|T}^2 + \mathbf{z}_t'\mathbf{P}_{t|T}\mathbf{z}_t\} \quad (19)$$

where

$$e_{t|T} = y_t - \mathbf{z}_t'\mathbf{a}_{t|T} \qquad t = 1, \ldots, T \quad (20)$$

Note that since $\mathbf{z}_t$ is time invariant its effect on equation (19) is simply to pick out certain elements from $\sum \mathbf{P}_{t|T}$. The estimator of $\mathbf{Q}$ is

$$\hat{\mathbf{Q}} = T^{-1}\sum_{t=1}^{T}[\mathbf{n}_{t|T}\mathbf{n}_{t|T}' + \mathbf{P}_{t|T} + \mathbf{T}_t\mathbf{P}_{t-1|T}\mathbf{T}_t' - \mathbf{T}_t\mathbf{P}_{t,t-1|T} - \mathbf{P}_{t,t-1|T}\mathbf{T}_t'] \quad (21)$$

where

$$\mathbf{n}_{t|T} = \mathbf{a}_{t|T} - \mathbf{T}_t\mathbf{a}_{t-1|T} \qquad t = 1, \ldots, T \quad (22)$$

and

$$\mathbf{P}_{t,t-1|T} = E[(\mathbf{a}_{t|T} - \boldsymbol{\alpha}_t)(\mathbf{a}_{t-1|T} - \boldsymbol{\alpha}_{t-1})'] \qquad t = 1, \ldots, T \quad (23)$$

As Watson and Engle (1983) point out, a relatively straightforward way of computing $\mathbf{P}_{t,t-1|T}$ is to augment the state vector by the vector of lagged values, $\boldsymbol{\alpha}_{t-1}$. The matrix $\mathbf{P}_{t,t-1|T}$ then appears as the off-diagonal block of the $\mathbf{P}_{t|T}$ matrix of the augmented state vector.

For the specification of the initial state vector, $\mathbf{a}_0$ and $\mathbf{P}_0$ will, in general, be unknown. The simplest way to proceed is to set $\mathbf{a}_0$ equal to some arbitrary vector, say $\mathbf{a}_0 = \mathbf{0}$, and to let $\mathbf{P}_0 = \varkappa\mathbf{I}$, where $\varkappa$ is a large but finite number. If an information filter is used to compute the smoothed estimates of the $\boldsymbol{\alpha}_t$'s the approximation involved in the use of the large $\varkappa$ device is avoided by the use of $\mathbf{P}_0^{-1} = \mathbf{0}$.

The application of the EM algorithm to the basic structural model requires some modifications. All but the first three equations in the transition equation are identities, and the only unknown elements in $\mathbf{Q}$ are σ_η^2, σ_ζ^2 and σ_ω^2. Thus, in equation (3.17), n is three and $\mathbf{Q}$ is diagonal. Furthermore, because the only elements in $\boldsymbol{\alpha}_{t-1}$ which do not appear in $\boldsymbol{\alpha}_t$ are μ_{t-1}, β_{t-1} and γ_{t-s}, it is only necessary to augment $\boldsymbol{\alpha}_t$ by these three elements. Thus, the augmented state vector is an $(s+4)\times 1$ vector defined by $\boldsymbol{\alpha}_t^\dagger = (\boldsymbol{\alpha}_t', \gamma_{t-s}, \mu_{t-1}, \beta_{t-1})'$. Let $n_{i,t|T}$ denote the ith element of $\mathbf{n}_{t|T}$ and define three vectors, $\mathbf{d}_1$, $\mathbf{d}_2$ and $\mathbf{d}_3$ such that the first element of $\mathbf{d}_1$ is unity, the last two are both minus one and the remainder are zero, the second element of $\mathbf{d}_2$ is one, the last is minus one and the remainder are zero, and all the elements in $\mathbf{d}_3$ are unity apart from the first two and the last two, that is, $\mathbf{d}_1 = (1\,0'\,-1\,-1)$, $\mathbf{d}_2 = (0\,1\,0'\,-1)$ and $\mathbf{d}_3 = (0\,0\,1\ldots 1\,0\,0)'$. Then

$$\eta_{i,t|T} = \mathbf{d}_i'\mathbf{a}_{t|T}^\dagger \qquad i = 1,2,3, \qquad t = 1,\ldots,T \tag{24}$$

and

$$\hat{\sigma}_i^2 = T^{-1}\left\{\sum_{t=1}^{t} n_{i,t|T}^2 + \mathbf{d}_i'\left[\sum_{t=1}^{T}\mathbf{P}_{t|T}^\dagger\right]\mathbf{d}_i\right\} \qquad i = 1,2,3 \tag{25}$$

where $\sigma_1^2 = \sigma_\eta^2$, $\sigma_2^2 = \sigma_\zeta^2$ and $\sigma_3^2 = \sigma_\omega^2$.

Three minor points may be made concerning the implementation of the algorithm. First, since $\mathbf{n}_{1|T} = 0$, equation (3.25) can be written with a divisor of $T-1$ and a summation running from $t = 2$ to T. Second, if the Kalman filter reaches a steady state, considerable computational savings can be effected in the smoothing algorithm, because $\mathbf{P}_t$ is effectively time invariant. Third, if the variances of η_t, ζ_t and ω_t are expressed relative to σ_ϵ^2, the latter may be concentrated out of the likelihood function, as in the prediction error decomposition methods.

Finally, note that the form of equations (3.19) and (3.25) is such that the estimates of the variances will always satisfy the non-negativity constraints. This is a reflection of the self-consistency property of the EM algorithm; see Efron (1982, Remark H).

MAXIMUM LIKELIHOOD ESTIMATION IN THE FREQUENCY DOMAIN

Consider the stationary form of the basic structural model, equation (7), and let $T^* = T - s - 1$. Let λ_j, $j = 0,\ldots,T^*-1$, be equally spaced frequencies over the interval $[-\pi,\pi]$, let $f(\lambda_j)$ denote the spectral density of $\Delta\Delta_s y_t$ at frequency λ_j and let $I(\lambda_j)$ denote the corresponding periodogram (sample spectral density) ordinate. By making appropriate circularity assumptions for $\Delta\Delta_s y_t$ it can be shown that the likelihood function can be written in the form

$$\log L = -T^*\log 2\pi - \frac{1}{2}\sum_{j=0}^{T^*-1}\log f(\lambda_j) - \frac{1}{2}\sum_{j=0}^{T^*-1}\frac{I(\lambda_j)}{f(\lambda_j)} \tag{26}$$

Since the circularity assumption cannot really be taken seriously in this context, equation (26) is best regarded as an approximation to one of the exact likelihood functions defined earlier.

The likelihood in equation (26) can be written in terms of the autocovariance-generating function, $g(L)$, by noting that setting $L = \exp(i\lambda_j)$ and dividing by 2π yields the spectral density

at λ_j, i.e.

$$f(\lambda_j) = (1/2\pi)g_j \tag{27}$$

where g_j denotes $g(\exp(i\lambda_j))$. Substituting in equation (26) then gives

$$\log L = -\frac{T^*}{2}\log 2\pi - \frac{1}{2}\sum_{j=0}^{T^*-1}\log g_j - \pi\sum_{j=0}^{T^*-1}\frac{I(\lambda_j)}{g_j} \tag{28}$$

The autocovariance-generating function for the basic structural model is

$$g(L) = (1-L^s)(1-L^{-s})\sigma_\eta^2 + \frac{(1-L^s)(1-L^{-s})}{(1-L)(1-L^{-1})}\sigma_\xi^2 + (1-L)^2(1-L^{-1})^2\sigma_\omega^2 + (1-L)(1-L^{-1})(1-L^s)(1-L^{-s})\sigma_\varepsilon^2 \tag{29}$$

and setting $L = \exp(i\lambda_j)$ yields

$$g_j = 2(1-\cos\lambda_j s)\sigma_\eta^2 + \frac{(1-\cos\lambda_j s)}{1-\cos\lambda_j}\sigma_\xi^2 + (6 - 8\cos\lambda_j + 2\cos 2\lambda_j)\sigma_\omega^2 + 4(1-\cos\lambda_j)(1-\cos\lambda_j s)\sigma_\varepsilon^2 \tag{30}$$

Note that since $\Delta_s/\Delta = S(L)$, the term involving σ_ξ^2 can also be expressed as

$$(1-\cos\lambda_j s)/(1-\cos\lambda_j) = s + 2\sum_{h=1}^{s-1}(s-h)\cos\lambda_j h \tag{31}$$

The right-hand side of equation (31) can be used to evaluate the left-hand side when $\lambda_j = 0$.

As in the time domain, one of the parameters, usually σ_ε^2 or σ_η^2, may be concentrated out of the likelihood function. If this parameter is, say, σ_ε^2, then $g(L) = \sigma^2\bar{g}_\varepsilon(L)$, and following Nerlove *et al.* (1979, 135) we have

$$\tilde{\sigma}_\varepsilon^2 = (2\pi/T^*)\sum_j I(\lambda_j)/\bar{g}_j \tag{32}$$

Expression for the first and second derivatives of the frequency domain likelihood are derived in Appendix 1. A large sample approximation to the information matrix has as its ihth element

$$\frac{1}{2}\sum_j \frac{1}{g_j^2}\frac{\partial g_j}{\partial\psi_i}\frac{\partial g_j}{\partial\psi_h} \tag{33}$$

where ψ_i denotes a typical element in the parameter vector ψ. (In the basic structural model ψ consists of σ_η^2, σ_ξ^2, σ_ω^2, and σ_ε^2.) Since equation (33) depends only on first derivatives a scoring algorithm[2] is an attractive proposition. However, in the results reported in the next section we used the same general Gill–Murray–Pitfield algorithm as was used for computing the time domain ML estimators.

It is important to note that each periodogram ordinate, $I(\lambda_j)$ need be calculated only once, rather than at each iteration of a numerical optimization procedure. Furthermore, the same values can be used in the estimation of different models. In calculating the periodogram ordinates it pays to make use of various trigonometric identities. This was done in the calculations reported below. For moderate and large sample sizes considerable further saving can be made by breaking down the computation of the periodogram ordinates into several parts; see Priestley (1981, 575–7). Since our computations were primarily for short time series this option was not used.

Nerlove *et al.* (1979) made considerable use of frequency domain ML estimation in

estimating unobserved components ARIMA models. They found it to be much more efficient and reliable than estimating the models in the time domain via a reduced-form ARMA model. However, they did not consider time domain methods based on the Kalman filter.

ASYMPTOTIC THEORY

The asymptotic properties of the ML estimators of structural models can be determined most easily in the frequency domain. The relevant results are given by Walker (1964) and are summarized in Hannan (1970, 395–8). They relate to a stationary Gaussian process with a continuous spectral density, $f(\lambda)$ in $[-\pi, \pi]$ which is nowhere zero in that range. The spectral density is assumed to depend on a set of n unknown parameters which are contained in a vector ψ. It is shown that:

(1) $\tilde{\psi}$, the ML estimator, converges in probability to ψ;

(2) Under the assumptions of continuity for the derivatives up to order three w.r.t. ψ of the function $1/g(L)$ in the neighbourhood of the true parameter value, ψ_0, and assuming that the parameters θ_τ in the MA representation of a stationary process

$$w_t = \sum_{\tau=0}^{\infty} \theta_\tau \xi_{t-\tau}, \qquad \xi_t \sim \text{NID}(0, \sigma_\xi^2) \tag{34}$$

are such that

$$\sum_{\tau=0}^{\infty} \tau |\theta_\tau| < \infty \tag{35}$$

the vector $\sqrt{T}(\tilde{\psi} - \psi_0)$ has a limiting distribution with zero mean vector and covariance matrix $\mathbf{IA}^{-1}(\psi)$, where the ijth element of $\mathbf{IA}(\psi)$ is:

$$\frac{1}{4\pi} \int_{-\pi}^{\pi} \left[\frac{\partial \log g(e^{i\lambda})}{\partial \psi_i}\right]\left[\frac{\partial \log g(e^{i\lambda})}{\partial \psi_j}\right] d\lambda \tag{36}$$

evaluated at ψ_0. (It is assumed that $\mathbf{IA}$ is non-singular, that is, the model is identifiable.)

In the case of the basic structural model, $f(\lambda)$ is proportional to the expression given in equation (30). The conditions for $f(\lambda)$ to be everywhere non-zero in the range $[-\pi, \pi]$ are that σ_ξ^2 and σ_ω^2 should be strictly positive. This can be seen by first considering $\lambda_j = 0$. In this case all the components of equation (30) are zero except for the term involving σ_ξ^2. Similarly, for the seasonal frequencies, $\lambda_j = 2\pi j/s$, $j = 1, \ldots, s/2$, all the terms disappear except for the one involving σ_ω^2. The conditions $\sigma_\omega^2 > 0$ and $\sigma_\xi^2 > 0$ are also sufficient to ensure the continuity of the derivatives of $1/g(L)$.

The following points should be noted:

(1) The frequency domain likelihood function for the basic structural model is unbounded if either $\sigma_\omega^2 = 0$ or $\sigma_\xi^2 = 0$. The exact time domain likelihood is not unbounded at this point, and therefore estimates of σ_ω^2 and σ_ξ^2 equal to zero can arise in practice. In fact, the result in the next section show that it is not unusual for this to happen. The situation parallels that which occurs in MA models. For example with an MA(1) model,

$$y_t = \xi_t + \theta \xi_{t-1}, \qquad \xi_t \sim \text{NID}(0, \sigma_\xi^2) \tag{37}$$

the usual asymptotic theory is not valid when the model is strictly non-invertible, i.e. when $|\theta| = 1$. Furthermore, when $|\theta| < 1$, there is a finite probability that the ML estimator of

θ will be *exactly* equal to plus or minus one; see Sargan and Bhargava (1983) and Cryer and Ledolter (1981). In the very simple random walk plus noise structural model

$$y_t = \eta_t + \varepsilon_t, \qquad \varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2) \tag{38a}$$

$$\mu_t = \mu_{t-1} + \eta_t, \qquad \eta_t \sim \text{NID}(0, \sigma_\eta^2) \tag{38b}$$

the condition for $f(\lambda)$ to be everywhere non-zero is that $\sigma_\eta^2 > 0$. The reduced form of equation (38) is an ARIMA (0, 1, 1) model with $-1 \leqslant \theta \leqslant 0$, and $\theta = -1$ when $\sigma_\eta^2 = 0$.

(2) The information matrix can be approximated by equation (33) in finite samples.

(3) Likelihood ratio, Lagrange multiplier and Wald tests of a null hypothesis in which either, or both, of the parameters σ_ξ^2 and σ_ω^2 are set equal to zero cannot be based on the usual asymptotic theory. Tests can, however, be constructed using the principle of a most powerful invariant test; see Franzini and Harvey (1983).

(4) The situation with regard to testing hypothesis in which either σ_η^2 or σ_ε^2 is equal to zero is somewhat different. A likelihood ratio test can be employed here provided that account is taken of the fact that values of σ_η^2 or σ_ε^2 equal to zero lie on the boundary of the parameter; space see Chernoff (1954). Thus, if, for example, $\sigma_\eta^2 = 0$ under the null hypothesis, the distribution of the likelihood ratio statistic is a mixture of a χ_1^2 distribution and a value of zero, both with a probability of one-half. The Wald statistic has a similar distribution, while the Lagrange multiplier statistic is χ^2 as usual. However, a more powerful one-sided LM test based on an asymptotic normal distribution can also be constructed. The formulation of LM tests in the frequency domain is described in Harvey and Hotta (1982).

(5) For the basic structural model, having σ_ω^2 and σ_ξ^2 strictly positive is necessary and sufficient for the model to be detectable and stabilizable; compare the conditions needed for the asymptotic theory to go through in the time domain as discussed in Pagan (1980).

(6) The identifiability of the basic structural model can be demonstrated by deriving the autocovariance function for the stationary process, $\Delta\Delta_s y_t$, of equation (7); see Harvey and Todd[3] (1983, 201) for $s = 4$. In general, there are $s + 2$ equations in four unknowns and consistent estimators of the four unknown parameters can be obtained by solving four of these. (There is more than one way of selecting the four equations). See also de Jong (1984).

ESTIMATION RESULTS WITH REAL AND SIMULATED DATA

This section reports the results of applying the various ML estimation procedures to various data sets. The aim was to obtain some insight into the properties, both with regard to the values of the estimates produced and the time required to compute them. Table 1 shows the results. The first series is the well-known airline passenger series—see, for example, Box and Jenkins (1976, 531)—aggregated to give 48 quarterly observations. The first 40 were used to estimate the model and the last eight were retained for post-sample prediction. The remaining six series are the quarterly UK macroeconomic time series used in the study by Harvey and Todd (1983). These each consist of 37 observations from 1957/3 to 1966/3, with the next eight observations used for post-sample predictive testing.

Estimation procedures

Estimates of the unknown parameters in the basic structural model were computed by five ML procedures.

(1) TD—The time domain prediction error decomposition form of the likelihood function, equation (10), maximizing using the variable metric Gill–Murray–Pitfield algorithm, EO4JBF in the NAG library.
(2) EM—The EM algorithm, modified by incorporating a line search procedure into it. This was found to be essential, since otherwise the EM algorithm was very slow indeed. The stopping criterion used was $L^* - L^\dagger < 10^{-7}$, where $L^\dagger$ is the 'likelihood' function obtained using the parameters when the EM is called and L^* is the corresponding function obtained from the first line search.
(3) EM*—As (2) but with a stopping value criterion based on the estimate of the one-step-ahead prediction error variance, $\tilde{\sigma}^2$ (see the sub-section below). The algorithm stopped when $(\tilde{\sigma}^2 - \tilde{\sigma}^{*2})/\tilde{\sigma}^{*2} < 10^{-4}$, where $\tilde{\sigma}^2$ is the prediction error variance from the EM algorithm and $\tilde{\sigma}^{*2}$ is the corresponding quantity from the first line search.
(4) FD—The frequency domain likelihood, equation (28), maximized using the Gill–Murray – Pitfield algorithm.

The computations were carried out on the University of London CDC 7600 and Cray machines. There were some differences between the two machines with respect to timings but the general picture was the same in both cases.

For all estimation procedures, σ_ϵ^2 was concentrated out of the likelihood function and the iterations were started off with the relative variances of σ_η^2, σ_ξ^2, and σ_ω^2 all set at unity.

Prediction error variance and post-sample predictions

The prediction error variance, $\tilde{\sigma}^2$, can be estimated from the Kalman filter as $\tilde{\sigma}^2 = f_T$, or $\tilde{\sigma}_\epsilon^2 f_T$ if σ_ϵ^2 is concentrated out of the likelihood function;[4] see also Harvey and Todd (1983). Of course, σ^2 is the variance of the disturbances in the reduced form ARIMA model; see the discussion in Harvey (1981b). Hence $\tilde{\sigma}^2$ corresponds to a standard measure of goodness of fit.

The prediction error variance can also be estimated in the frequency domain. If σ_ϵ^2 is concentrated out of the likelihood function, the appropriate formula is

$$\hat{\sigma}^2 = \tilde{\sigma}_\epsilon^2 \exp\left[\frac{1}{T^*}\sum_{j=0}^{T^*-1} \log g_j\right] \tag{39}$$

where $\tilde{\sigma}_\epsilon^2$ is defined by equation (32). Expression (39) can be obtained as a simple transformation of the maximized frequency domain likelihood function.

For all the estimation procedures considered, the likelihood surfaces are relatively flat for small samples. Hence differences between estimates of the unknown parameters are likely to be the exception rather than the rule. In order to assess the practical importance of these differences it is helpful to look at the predictions obtained in a post-sample period. We therefore calculated the sum of squares of one-step-ahead post-sample prediction errors and the sum of squares of the prediction errors from the unconditional predictions. Both quantities are presented in Table 1 divided by the number of post-sample observations.

Results

(1) The EM* algorithm gives similar results to the EM algorithm in most cases, but it usually required less iterations. In some cases this difference was quite significant, with EM taking ten or twenty times as long as EM*. In other cases the time was the same.
(2) There are some differences in the estimates obtained by the TD and FD procedures. In particular, the FD procedure always yields positive estimates σ_ξ^2. As observed, this is because the FD likelihood function becomes unbounded if σ_ξ^2 is zero. Similarly, $\tilde{\sigma}_\omega^2$ is always positive for FD.

Table 1. Quarterly airline data and UK macroeconomic time series[a]

Series	Estimation procedure	Parameter values					Average post-sample SS	
		σ_η^2	σ_ξ^2	σ_ω^2	σ_ε^2	σ^2	One-step	Unconditional
Airline (all estimates × 10^5)	TD	66	0.39	13	0	185	46	176
	EM	103	0.00	10	0	179	44	73
	EM*	48	8.43	0	0	191	60	111
	FD	83	0.30	9	0	188	99	154
Consumer durables	TD	408	0	181	0	1349	8963	5777
	EM	413	0	177	6	1350	8937	5741
	EM*	427	0	168	21	1353	8646	5654
	FD	310	4.76	188	77	1439	9943	5746
Other expenditure	TD	306	0	30	182	924	2526	2757
	EM	305	0	30	182	921	2542	3132
	EM*	299	0	30	190	920	2540	2778
	FD	243	0.19	75	264	965	2752	3182
Investment	TD	1392	0	1	112	1823	1706	944
	EM	1389	0	0	118	1817	1781	953
	EM*	1393	0	0	118	1815	1779	953
	FD	1240	4.19	20	273	1975	1363	854
Inventory investment	TD	1204	0	169	371	3274	7835	9616
	EM	1204	0	169	372	3274	7858	9617
	EM*	1191	0	165	397	3270	7815	9625
	FD	1149	1.40	89	655	3281	7382	9724
Imports	TD	880	0	0	268	1532	7179	4279
	EM	880	0	0	268	1533	7266	4279
	EM*	882	0	0	268	1533	7282	4279
	FD	885	0.62	11	320	1609	6645	4276
GDP	TD	3375	0	600	0	7663	5977	7011
	EM	3593	0	444	249	7709	5550	7049
	EM*	3594	0	444	249	7708	5549	7049
	FD	2966	2.35	146	1432	7605	4563	7694

[a] A zero entry for σ_ξ^2 indicates a value less than 0.005. A zero entry for any other parameter indicates a value less than 0.5.

For the macroeconomic series, the estimates of the slope variance, σ_ξ^2, are typically very close to zero, and in the case of TD estimation are often exactly equal to zero. A constant slope appears to be a fairly general feature for macroeconomic series of real, as opposed to monetary, variables. Much of the difference between TD and FD estimates arises because the FD estimates of σ_ξ^2 must be positive. Had the slope been constrained to be zero, the estimates of the remaining parameters would have been much closer.

(3) Computing times for the TD and FD procedures have not been given because of the different performance likely to be encountered on different machines. However, on the machines we used, the FD method was generally considerably faster than the TD method, with one function evaluation of TD taking anything between five and ten times as long as an FD function evaluation for a sample of size 50. The time taken to compute the

periodogram was typically about one-tenth of that required to maximize the TD likelihood functions. Since there was no indication of TD requiring fewer iterations than FD, the computational advantage appears to rest firmly with FD.

(4) The estimates obtained by the two EM algorithms are fairly close to the TD estimates in most cases.

(5) One function evaluation for EM or EM* takes roughly two or three times as long as for TD. Furthermore, the smoothing algorithm has certain storage requirements. There is therefore only a good case for using the EM algorithm if it can be shown to converge in a relatively small number of iterations. Our results seem to indicate that the EM* algorithm does converge faster than TD, but the convergence is not rapid enough to yield an algorithm which is faster than TD by an order of magnitude. Furthermore, it is quite clear that FD is faster than EM* overall, being from two to ten times as quick.

We had hoped, following the experience of Watson and Engle (1983), that the EM algorithm would yield estimates in the vicinity of the ML estimates in a relatively small number of iterations. Such estimates would be useful as starting values for direct maximization of the likelihood function. However, for our model this did not appear to be the case.

(6) For the series examined here, the estimates of σ_η^2 tend to dominate those of σ_ε^2. Furthermore, the estimates of σ_ε^2 are sometimes close to zero, as, for example, in the airline data. As a general rule, therefore, it may be preferable to concentrate σ_η^2, rather than σ_ε^2, out the likelihood function.

(7) Asymptotic standard errors for the estimates of the variance parameters were computed from the inverse of equation (33). Since the sample sizes are rather small, the standard errors are relatively large and it is not clear to what extent they are an accurate reflection of the true standard errors. In order to give some idea of the order of magnitude of the estimated standard errors we present the values we computed for inventory investment and imports in Table 2.

If σ_ξ^2 is zero, the asymptotic theory does not apply. When the estimate of σ_ξ^2 was zero the standard errors for the other parameters were computed by setting σ_ξ^2 equal to 10^{-10} in expression (33). The standard errors computed in this way were found to be close to those computed by numerically evaluating the Hessian of the time domain likelihood under the assumption that σ_ξ^2 was fixed. These numerical standard errors are shown in parentheses in Table 2 under the estimated standard errors for the time domain (TD) estimates.

Table 2. Estimated asymptotic standard errors for inventory investment and imports

Series	Estimation procedure	Parameters			
		σ_ε^2	σ_η^2	σ_ξ^2	σ_ω^2
Inventory investment	TD	514	673	—	123
		(492)	(559)		(155)
	FD	619	1005	65	70
Imports	TD	264	472	—	—
		(248)	(446)		
	FD	345	686	47	9

Conclusions

The estimates obtained by the TD and FD methods are comparable, and even when the actual values are not very close the forecasting performance is similar. The differences which do arise may be a reflection of the properties of exact as opposed to approximate ML estimators rather than of the difference between time and frequency domain estimators. Thus the use of a steady-state Kalman filter in the time domain may give estimates closer to those obtained in the frequency domain; compare the situation for ARMA models with exact ML, conditional sum of squares and frequency domain estimators.

Despite the necessary caveats concerning program structure and machine dependency it does appear that the FD method is computationally faster than TD. Furthermore, our results are based on quarterly observations, and with monthly observations the balance may swing even further to the frequency domain method. Against this there is an argument that the Kalman filter could be speeded up considerably by switching over to steady-state recursions at a suitable point. Of course, one possibility is to compute estimates by the FD method and then use these as starting values in TD.

The EM algorithm was not as fast as we had hoped, even given the modifications such as line searches and prediction error variance stopping rules which we introduced into it. It is slower than FD, and although EM* appears to be faster than TD it is not significantly so. This might seem to contradict the findings of Watson and Engle (1983), but we note that their model was a multivariate one with a relatively large number of parameters. It may well be that the EM algorithm becomes relatively more attractive in such cases.

EXPLANATORY VARIABLES

The basic structural model may be extended by the introduction of explanatory variables. For a single explanatory variable this yields

$$y_t = \mu_t + \gamma_t + \delta(L)x_t + \varepsilon_t \tag{40}$$

where $\delta(L)$ is the polynomial in the lag operator,

$$\delta(L) = \delta_0 + \delta_1 L + \cdots + \delta_m L^m \tag{41}$$

Estimation in the time domain can be carried out by letting the lag coefficients be part of the state vector as in Harvey and Phillips (1979). Alternatively, the GLS method of Wecker and Ansley (1983) may be used. The attraction of the first method is that it can be readily adapted to handle the case when the δ_h's are time varying. Note that in both cases the δ's are effectively concentrated out of the likelihood function when they are time invariant. Thus numerical optimisation is only carried out with respect to the variance parameters of the basic structural model.

In the frequency domain, the likelihood function for equation (40) is of the form (28), but with the periodogram defined by

$$I(\lambda_j) = \frac{1}{2\pi T^*}\left| \sum_{t=s+2}^{T} \Delta\Delta_s y_t \exp(-i\lambda_j t) - \delta(\exp(-i\lambda_j)) \sum_j \Delta\Delta_s x_t \exp(-i\lambda_j t)\right|^2 \tag{42}$$

Let $I_y(\lambda_j)$, $I_x(\lambda_j)$ and $I_{xy}(\lambda_j)$ denote the periodograms of $\Delta\Delta_s y_t$ and $\Delta\Delta_s x_t$ and the cross periodogram, respectively. Then

$$I(\lambda_j) = I_y(\lambda_j) - 2\boldsymbol{\delta}' \mathrm{Re}\{\mathbf{e}_j I_{xy}(\lambda_j)\} + \boldsymbol{\delta}' \mathbf{e}_j \bar{\mathbf{e}}_j' \boldsymbol{\delta} I_x(\lambda_j) \tag{43}$$

where $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_m)'$ and $\mathbf{e}_j$ is an $(m+1) \times 1$ vector with $\exp(-\lambda_j(h-1))$ in the hth position. For given values of the parameters in $g(.)$, differentiating equation (43) and setting to zero gives

$$\hat{\boldsymbol{\delta}} = \left[\sum_j \mathbf{e}_j \bar{\mathbf{e}}_j' \frac{I_x(\lambda_j)}{g_j}\right]^{-1} \sum_j \frac{1}{g_j} \mathrm{Re}\{\mathbf{e}_j I_{xy}(\lambda_j)\} \tag{44}$$

compare Fishman (1969, 154). Substituting $\hat{\boldsymbol{\delta}}$ in equation (43) therefore enables $\boldsymbol{\delta}$ to be concentrated out of the likelihood function. Note that the periodograms and cross-periodograms need be calculated only once, even if the specification of the lag length is changed.

All the above methods can be applied when there are several explanatory variables or the lag structure is modelled by an Almon polynomial distributed lag.

Lagged dependent variables can also be brought into the model. Thus

$$\sum_{l=0}^{r} \varphi_l y_{t-l} = \mu_t + \gamma_t + \sum_{h=0}^{m} \delta_h x_{t-h} + \varepsilon_t \tag{45}$$

with $\varphi_0 = 1$. The time domain method can again be used with $\varphi_1, \ldots, \varphi_r$ added to the state vector. The φ_l's can also be allowed to be time varying even though stochastic time variation makes the model non-linear; see Anderson and Moore (1979; 43–4) and Liptser and Shiryayev (1978, 62). When the φ_l's and δ_h's are time invariant the frequency domain method can be applied using methods similar to those described above. An application involving explanatory variables may be found in Harvey *et al.* (1986).

NEW DEVELOPMENTS

The results reported above were computed several years ago on a University of London mainframe computer. The programs for computing time domain and frequency domain ML estimators have now been successfully adapted for use on an IBM personal computer, and a menu-driven program for estimating models for univariate time series, with and without explanatory variables, has been written. Frequency domain estimates may be computed by the method of scoring, which is generally very rapid. The TD estimates are computed by a quasi-Newton algorithm, with starting values normally obtained from the FD scoring algorithm. Once estimated, the model may be subjected to diagnostic checking, and smoothed estimates and predictions may be made and graphed.

The PC program is known as STAMP, that is, *S*tructural *T*ime series *A*nalyser, *M*odeller and *P*redictor. Further details may be obtained by writing to the first author.

APPENDIX 1: DERIVATIVES FOR THE FREQUENCY DOMAIN ESTIMATOR

Let ψ_i be a typical parameter in a structural model. Differentiating equation (28) with respect to ψ_i gives

$$\frac{\partial \log \mathrm{L}}{\partial \psi_i} = -\frac{1}{2}\sum \frac{1}{g}\frac{\partial g}{\partial \psi} + \pi \sum \frac{I}{g^2}\frac{\partial g}{\partial \psi_i} = \frac{1}{2}\sum \left[2\pi \cdot \frac{\mathrm{I}}{g} - 1\right]\frac{1}{g}\frac{\partial g}{\partial \psi_i} \tag{A1.1}$$

where $g = g_j$ and the summations run from $j = 0$ to $T^* - 1$. Differentiating with respect to a second parameter ψ_h gives

$$\frac{\partial^2 \log L}{\partial \psi_i \partial \psi_h} = \sum\left[2\pi \frac{\mathrm{I}}{g} - 1\right]\left[\frac{1}{2g}\right]\frac{\partial^2 g}{\partial \psi_i \partial \psi_h} - 2\sum\left[4\pi \frac{\mathrm{I}}{g} - 1\right]\left[\frac{1}{2g}\right]^2 \frac{\partial g}{\partial \psi_i}\frac{\partial g}{\partial \psi_h} \tag{A1.2}$$

The derivatives of g will normally be quite easy to obtain. For example, in the case of equation (30) they can be written down immediately.

Second derivatives of g are more complicated. However, if expectations are taken in equation (A1.2) the first term effectively disappears if the sample size is reasonably large as

$$\lim_{T\to\infty} E\{I(\lambda_j)\} = g_j/2\pi \tag{A1.3}$$

This leaves equation (33) as an estimate of the ith element of the information matrix, cf. Whittle (1954, p. 214).

APPENDIX 2: COMPARISON OF LIKELIHOOD FUNCTIONS

Consider a state space model of the form (8), with $k \leqslant$ m non-stationary elements in the state vector. Let $\boldsymbol{\gamma}_t$ denote the $k \times 1$ vector containing these non-stationary elements. The model can be written conditional on $\boldsymbol{\gamma}_t$ at any particular point in time, $t = \tau$. This can include values of t outside the range 1 to T. Thus

$$\mathbf{y} = \bar{\mathbf{x}}_\tau \boldsymbol{\gamma}_\tau + \mathbf{w}_\tau \tag{A2.1}$$

where $\mathbf{y}$ is the $T \times 1$ vector $(y_1, \ldots, y_T)'$, $\bar{\mathbf{x}}_\tau$ is a $T \times k$ matrix of fixed values and $\mathbf{w}_\tau$ is a $T \times 1$ vector of disturbances with mean zero and covariance matrix $\boldsymbol{\Omega}_\tau$; cf. equation (15). The subscripts on $\bar{\mathbf{x}}_\tau$, $\boldsymbol{\Omega}_\tau$ and $\mathbf{w}_\tau$ stress their dependence on the chosen value of τ.

Let $\tilde{\boldsymbol{\gamma}}_\tau$ be the GLS estimator of $\boldsymbol{\gamma}_\tau$, conditional on the other parameters in the model. The general result to be proved is that, for all τ,

$$(\mathbf{y} - \bar{\mathbf{x}}_\tau \hat{\boldsymbol{\gamma}}_\tau)' \boldsymbol{\Omega}_\tau^{-1} (\mathbf{y} - \bar{\mathbf{x}}_\tau \hat{\boldsymbol{\gamma}}_\tau) = \sum_{t=k+1}^{T} v_t^2/f_t \tag{A2.2}$$

where v_t is the one-step-ahead prediction error obtained by applying the Kalman filter to equation (8) with starting values constructed from the first k observations (diffuse prior) and f_t is its variance; compare the final term in equation (10).

Given that $\boldsymbol{\Omega}_\tau$ is a p.d. matrix there exists a lower triangular matrix, $\mathbf{L}_\tau$, with positive diagonal elements such that

$$\mathbf{L}_\tau' \mathbf{L}_\tau = \boldsymbol{\Omega}_\tau^{-1} \tag{A2.3}$$

Multiplying equation (A2.1) through by $\mathbf{L}_\tau$ yields a transformed regression model in which the disturbances have a scalar covariance matrix. Estimating $\boldsymbol{\gamma}_\tau$ in this model by recursive least squares yields a set of 'recursive residuals', denoted by $\tilde{v}_\tau$, $t = k+1, \ldots, T$ and it may be shown that

$$\sum_{t=k+1}^{T} \tilde{v}_t^2 = (\mathbf{L}_\tau \mathbf{y} - \mathbf{L}_\tau \bar{\mathbf{x}}_\tau \hat{\boldsymbol{\gamma}}_\tau)' (\mathbf{L}_\tau \mathbf{y} - \mathbf{L}_\tau \bar{\mathbf{x}}_\tau \hat{\boldsymbol{\gamma}}_\tau) = (\mathbf{y} - \bar{\mathbf{x}}_\tau \hat{\boldsymbol{\gamma}}_\tau)' \boldsymbol{\Omega}_\tau^{-1} (\mathbf{y} - \bar{\mathbf{x}}_\tau \hat{\boldsymbol{\gamma}}_\tau) \tag{A2.4}$$

cf. Brown *et al.* (1975)

The tth recursive residual in the transformed model will be equal to a fixed quantity multiplied by the prediction error

$$\tilde{v}_t = \bar{y}_t - \hat{\bar{y}}_{t|t-1} \tag{A2.5}$$

where $\bar{y}_t$ and $\bar{x}_t$ are the tth set of transformed variables and $\hat{\bar{y}}_{t|t-1}$ is the MMSE of $\bar{y}_t$ based on the first $t-1$ transformed observations. However, in view of the lower triangular nature of $\mathbf{L}_\tau$, $\hat{\bar{y}}_{t|t-1}$ depends only on the first $t-1$ untransformed observations while $\bar{y}_t$ is equal to y_t

multiplied by the tth diagonal element of $\mathbf{L}_\tau$ which is fixed, plus a linear combination of $(y_{t-1}, \ldots, y_1)$. Therefore the prediction error for the original observation y_t, given $(y_{t-1}, \ldots, y_1)$, is

$$\tilde{v}_t^* = y_t - g_t(y_{t-1}, \ldots, y_1) \qquad t = k+1, \ldots, T \tag{A2.6}$$

where $g_t(y_{t-1}, \ldots, y_1)$ is a linear function of $(y_{t-1}, \ldots, y_1)$. Now $g_t(y_{t-1}, \ldots, y_1)$ must be the MMSE of y_t based on $(y_{t-1}, \ldots, y_1)$ since if it were not it would be possible to construct a better estimator of $\tilde{y}_t$ than the one implied by equation (A2.5). Since $g_t(y_{t-1}, \ldots, y_1)$ is the MMSE of y_t given $(y_{t-1}, \ldots, y_1)$, its values must be independent of the choice of τ. Therefore the recursive residuals in the transformed system must be identical for all values of τ, and so it follows from equation (A2.4) that the generalized residual sum of squares does not depend on τ.

Consider the prediction errors

$$v_t = y_t - \tilde{y}_{t\,t-1} \qquad t = k+1, \ldots, T \tag{A2.7}$$

obtained from the Kalman filter with starting values constructed from the first k observations as in method (1a). The quantity $\tilde{y}_{t\,t-1}$ is the MMSE of y_t and so it must be identical to $g_t(y_{t-1}, \ldots, y_1)$. Therefore $v_t = \tilde{v}_t^*$ is equal to the tth recursive residual multiplied by a fixed quantity, it follows that when the v_t's are standardized they will be identically equal to the recursive residuals in the transformed system, i.e.

$$\tilde{v}_t = v_t / f_t^{1/2} \qquad t = k+1, \ldots, T \tag{A2.8}$$

Taking equations (A2.1) and (A2.8) together proves the result in equation (A2.2)

ACKNOWLEDGEMENTS

This research was supported by an ESRC grant to the DEMEIC Programme at the LSE, where the second author was a research assistant over the period 1982–5. We are grateful to J. Durbin, L. Hotta, A. Pagan, P. Pereira, T. Pukkila and members of the DEMEIC workshop for helpful comments on our original 1984 LSE discussion paper. We would like to thank Peter Young for encouraging us to revise this discussion paper for publication and a referee for his comments.

NOTES

1. It is interesting to note that if $\mathbf{Q} = \mathbf{0}$, $\sum \mathbf{P}_{t|T} = \sigma_\epsilon^2 (\sum \mathbf{z}_t \mathbf{z}_t')^{-1}$. Since $\sum [\mathbf{z}_t' (\sum \mathbf{z}_t \mathbf{z}_t') \mathbf{z}_t] = n$, it follows that if $\mathbf{Q} = \mathbf{0}$, the ML estimator of σ_ϵ^2 is

$$\tilde{\sigma}_\epsilon^2 = (T-n)^{-1} \sum e_t^2$$

where e_t is the tth OLS residual from a regression of y_t on $\mathbf{z}_t$. The reason for obtaining the unbiased estimator of σ_ϵ^2, rather than the usual biased estimator, is that $\boldsymbol{\alpha}_0$ is regarded as being a random variable.
2. De Jong (1984) has recently shown that an asymptotically efficient two-step estimator may be constructed by a two-step frequency domain regression. This estimator is equivalent to carrying out one iteration of the method of scoring starting from consistent estimates.
3. However, note that the coefficient of σ_ω^2 in the expression for $\gamma(0)$ should be 6, not 4.
4. Note that f_T will be greater than or equal to the steady-state value $\bar{f}$. However, unless the sample size is very small, f_T and $\bar{f}$ are very close.

REFERENCES

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

Ansley, C. F. and Newbold, P., 'Finite sample properties of estimators for auto-regressive-moving average processes', *Journal of Econometrics*, **13** (1980), 159–84.

Ansley, C. F. and Kohn, R., 'Estimation, filtering and smoothing in state space models with incompletely specified initial conditions', *Annals of Statistics*, **13** (1985), 1286–1316.

Box, G. E. P., Hillmer, S. C. and Tia, G. C., 'Analysis and modelling of seasonal time series', in A. Zellner (ed.), *Seasonal Analysis of Economic Time Series*, 309–34, Washington, DC: Bureau of the Census, (1978).

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

Boyles, R. A., 'On the convergence of the EM Algorithm', *Journal of the Royal Statistical Society*, B, **45** (1983), 47–50.

Brown, R. L., Durbin, J. and Evans, J. M., 'Techniques for testing the constancy of regression relationships over time', *Journal of the Royal Statistical Society*, B, **37** (1975), 141–92.

Chan, S. W., Goodwin, G. C. and Sin, K. S., 'Convergence properties of the Riccati difference equation in optimal filtering of nonstabilizable systems', *IEEE Transactions on Automatic Control*, **AC-29** (1984), 110–18.

Chernoff, H., 'On the distribution of the likelihood ratio', *Annals of Mathematical Statistics*, **25** (1954), 573–8.

Cryer, J. D. and Ledolter, J., 'Small sample properties of the parameters of the maximum likelihood estimator in the first order moving average model', *Biometrika*, **68** (1981), 691–4.

De Jong, P., 'An asymptotically efficient estimator for variance components time series models', *Methods of Operations Research*, **50** (1985), 275–86.

De Jong, P., 'The likelihood for a state space model', *Biometrika*, **75**, 165–9.

Dempster, A. P., Laird, N. M. and Rubin, D. B., 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society*, B, **39** (1977), 1–38.

Efron, B., 'Maximum likelihood and decision theory', *Annals of Statistics*, **11** (1982), 95–103.

Engle, R. F., 'Estimating structural models of seasonality', in A. Zellner (ed.), *Seasonal Analysis of Economic Time Series*, 281–308, Washington, DC: Bureau of the Census (1978).

Fishman, G. S., *Spectral Methods in Econometrics*, Cambridge, Mass.: Harvard University Press (1969).

Franzini, L. and Harvey, A. C., 'Testing for deterministic trend and seasonal components in time series models', *Biometrika*, **70** (1983), 673–82.

Gardner, G., Harvey, A. C. and Phillips, G. D. A., 'An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering', *Applied Statistics*, **29** (1980), 311–22.

Hannan, E. J., *Multiple Time Series*, New York: John Wiley (1970).

Harrison, P. J. and Stevens, C. F., 'Bayesian forecasting', *Journal of the Royal Statistical Society*, B, **38** (1976), 205–47.

Harvey, A. C., *Time Series Models*, Oxford: Phillip Allan and Atlantic Highlands, NJ: Humanities Press (1981a).

Harvey, A. C., 'Finite sample prediction and overdifferencing', *Journal of Time Series Analysis*, **2** (1981b), 221–32.

Harvey, A. C., 'Estimation procedures for a class of univariate time series models', LSE Econometrics Programme Discussion Paper, No. A28 (1982).

Harvey, A. C., 'Trends and cycles in macroeconomic time series', *Journal of Business and Economic Statistics*, **3** (1985), 216–227.

Harvey, A. C. and Phillips, G. D. A., 'The estimation of regression models with autoregressive-moving average disturbances', *Biometrika*, **66** (1979), 49–58.

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

Harvey, A. C., Henry, B., Peters, S. and Wren-Lewis, S., 'Stochastic trends in dynamic regression models: an application to the employment-output equation', *Economic Journal*, **96** (1986), 975–85

Kitagawa, G., 'A non-stationary time series model and its fitting by a recursive filter', *Journal of Time Series Analysis*, **2** (1981), 103–16.

Kitagawa, G. and Gersch, W., 'A smoothness prior—state space modelling of time series with trend and seasonality', *Journal of the American Statistical Association*, **79** (1984), 378–89.
Liptser, R. S. and Shiryayev, A. N., *Statistics of Random Processes, II, Applications*, (A. B. Aries, trans.) New York: Springer-Verlag.
Mélard, G., 'A fast algorithm for the exact likelihood of autoregressive moving average models', *Applied Statistics*, **33** (1984), 104–14.
Nerlove, M., Grether, D. M. and Carvalho, J. L., *Analysis of Economic Time Series*, Academic Press: New York (1979).
Pagan, A. R., 'Some identification and estimation results for regression models with stochastically varying parameters', *Journal of Econometrics*, **13** (1980), 341–63.
Priestley, M., *Spectral Analysis and Time Series*, New York: Academic Press (1981).
Pierce, D. A. 'Signal extraction error in nonstationary time series', *Annals of Statistics*, **7** (1979), 1303–20.
Rosenberg, B., 'Random coefficient models: the analysis of a cross section time series by stochastically convergent parameters regression', *Annals of Economic and Social Measurement*, **2** (1973), 399–428.
Sargan, J. D. and Bhargava, A., 'Maximum likelihood estimation of regression models with first-order moving average errors when the root lies on the unit circle', *Econometrica*, **51** (1983), 799–820.
Sarris, A. H., 'Kalman filter models: a Bayesian approach to estimation of time-varying regression coefficients', *Annals of Economic and Social Measurement*, **2** (1973), 501–23.
Walker, A. M., 'Asymptotic properties of least squares estimates of parameters of the spectrum of a stationary non-deterministic time series', *Journal of the Australian Mathematical Society*, **4** (1964), 363–84.
Watson, M. W. and Engle, R. F., 'Alternative algorithms for the estimation of dynamic factor, MIMIC and varying coefficient regression', *Journal of Econometrics*, **23** (1983), 385–400.
Wecker, W. E. and Ansley, C. F., 'The signal extraction approach to nonlinear regression and spline smoothing', *Journal of the American Statistical Association*, **78** (1983), 81–9.
Whittle, P., 'Some recent contributions to the theory of stationary processes', Appendix 2 of H. Wold, *A Study in the Analysis of Time Series*, 2nd edn, Stockholm: Almqvist and Wiksell (1954).
Wu, C. F. J., 'On the convergence of the EM algorithm', *Annals of Statistics*, **11** (1983), 95–103.

Authors' biographies:

Andrew Harvey is Professor of Econometrics in the Department of Statistics at the London School of Economics. He has made both theoretical and applied contributions in econometrics and time series analysis, and published articles in a wide range of journals. He is the author of two widely used text books: *The Econometric Analysis of Time Series* and *Time Series Models*. In addition, he has just published a new book which presents a unified treatment of much of the recent work on time series modelling. This book is entitled *Forecasting, Structural Time Series Models and the Kalman Filter*.

Simon Peters is currently a Lecturer in the Department of Econometrics, Monash University. He has worked as a research assistant at the London School of Economics and the University of Bristol. He was primarily responsible for the design and programming of the structural time series modelling package, STAMP.

Authors' addresses:

Andrew Harvey, Department of Statistics, London School of Economics, Houghton St., London WC2A 2AE.

Simon Peters, Department of Econometrics, Monash University, Clayton, Victoria 3168, Australia.