Chapter 50

STATE-SPACE MODELS*

JAMES D. HAMILTON

University of California, San Diego

Contents

*I am grateful to Gongpil Choi, Robert Engle and an anonymous referee for helpful comments, and to the NSF for support under grant SES-8920752. Data and software used in this chapter can be obtained at no charge by writing James D. Hamilton, Department of Economics 0508, UCSD, La Jolla, CA 92093-0508, USA. Alternatively, data and software can be obtained by writing ICPSR, Institute for Social Research, P.O. Box 1248, Ann Arbor, MI 48106, USA.

Handbook of Econometrics, Volume IV, Edited by R.F. Engle and D.L. McFadden

Abstract

This chapter reviews the usefulness of the Kalman filter for parameter estimation and inference about unobserved variables in linear dynamic systems. Applications include exact maximum likelihood estimation of regressions with ARMA disturbances, time-varying parameters, missing observations, forming an inference about the public's expectations about inflation, and specification of business cycle dynamics. The chapter also reviews models of changes in regime and develops the parallel between such models and linear state-space models. The chapter concludes with a brief discussion of alternative approaches to nonlinear filtering.

1. The state-space representation of a linear dynamic system

Many dynamic models can usefully be written in what is known as a *state-space* form. The value of writing a model in this form can be appreciated by considering a first-order autoregression

$$y_{t+1} = \phi y_t + \varepsilon_{t+1}, \tag{1.1}$$

with $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. Future values of y for this process depend on $(y_t, y_{t-1}, \ldots)$ only through the current value y_t. This makes it extremely simple to analyze the dynamics of the process, make forecasts or evaluate the likelihood function. For example, equation (1.1) is easy to solve by recursive substitution,

$$\begin{aligned} y_{t+m} &= \phi^m y_t + \phi^{m-1}\varepsilon_{t+1} + \phi^{m-2}\varepsilon_{t+2} + \cdots \\ &\quad + \phi^1 \varepsilon_{t+m-1} + \varepsilon_{t+m} \quad \text{for} \quad m = 1, 2, \ldots, \end{aligned} \tag{1.2}$$

from which the optimal m-period-ahead forecast is seen to be

$$E(y_{t+m} | y_t, y_{t-1}, \ldots) = \phi^m y_t. \tag{1.3}$$

The process is stable if $|\phi| < 1$.

The idea behind a state-space representation of a more complicated linear system is to capture the dynamics of an observed $(n \times 1)$ vector y_t in terms of a possibly unobserved $(r \times 1)$ vector ξ_t known as the *state vector* for the system. The dynamics of the state vector are taken to be a vector generalization of (1.1):

$$\xi_{t+1} = F\xi_t + v_{t+1}. \tag{1.4}$$

Here $\boldsymbol{F}$ denotes an $(r \times r)$ matrix and the $(r \times 1)$ vector $\boldsymbol{v}_t$ is taken to be i.i.d. $N(\boldsymbol{0}, \boldsymbol{Q})$. Result (1.2) generalizes to

$$\begin{aligned}\boldsymbol{\xi}_{t+m} &= \boldsymbol{F}^m\boldsymbol{\xi}_t + \boldsymbol{F}^{m-1}\boldsymbol{v}_{t+1} + \boldsymbol{F}^{m-2}\boldsymbol{v}_{t+2} + \cdots \\ &\quad + \boldsymbol{F}^1\boldsymbol{v}_{t+m-1} + \boldsymbol{v}_{t+m} \quad \text{for} \quad m = 1, 2, \ldots, \end{aligned} \tag{1.5}$$

where $\boldsymbol{F}^m$ denotes the matrix $\boldsymbol{F}$ multiplied by itself m times. Hence

$$E(\boldsymbol{\xi}_{t+m} | \boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \ldots) = \boldsymbol{F}^m\boldsymbol{\xi}_t.$$

Future values of the state vector depend on $(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \ldots)$ only through the current value $\boldsymbol{\xi}_t$. The system is stable provided that the eigenvalues of $\boldsymbol{F}$ all lie inside the unit circle.

The observed variables are presumed to be related to the state vector through the *observation equation* of the system,

$$\boldsymbol{y}_t = \boldsymbol{A}'\boldsymbol{x}_t + \boldsymbol{H}'\boldsymbol{\xi}_t + \boldsymbol{w}_t. \tag{1.6}$$

Here $\boldsymbol{y}_t$ is an $(n \times 1)$ vector of variables that are observed at date t, $\boldsymbol{H}'$ is an $(n \times r)$ matrix of coefficients, and $\boldsymbol{w}_t$ is an $(n \times 1)$ vector that could be described as measurement error; $\boldsymbol{w}_t$ is assumed to be i.i.d. $N(\boldsymbol{0}, \boldsymbol{R})$ and independent of $\boldsymbol{\xi}_1$ and $\boldsymbol{v}_\tau$ for $\tau = 1, 2, \ldots$. Equation (1.6) also includes $\boldsymbol{x}_t$, a $(k \times 1)$ vector of observed variables that are exogenous or predetermined and which enter (1.6) through the $(n \times k)$ matrix of coefficients $\boldsymbol{A}'$. There is a choice as to whether a variable is defined to be in the state vector $\boldsymbol{\xi}_t$ or in the exogenous vector $\boldsymbol{x}_t$, and there are advantages if all dynamic variables are included in the state vector so that $\boldsymbol{x}_t$ is deterministic. However, many of the results below are also valid for nondeterministic $\boldsymbol{x}_t$, as long as $\boldsymbol{x}_t$ contains no information about $\boldsymbol{\xi}_{t+m}$ or $\boldsymbol{w}_{t+m}$ for $m = 0, 1, 2, \ldots$ beyond that contained in $\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \ldots, \boldsymbol{y}_1$. For example, $\boldsymbol{x}_t$ could include lagged values of $\boldsymbol{y}$ or variables that are independent of $\boldsymbol{\xi}_\tau$ and $\boldsymbol{w}_\tau$ for all τ.

The state equation (1.4) and observation equation (1.6) constitute a *linear state-space repesentation* for the dynamic behavior of $\boldsymbol{y}$. The framework can be further generalized to allow for time-varying coefficient matrices, non-normal disturbances and nonlinear dynamics, as will be discussed later in this chapter. For now, however, we just focus on a system characterized by (1.4) and (1.6).

Note that when $\boldsymbol{x}_t$ is deterministic, the state vector $\boldsymbol{\xi}_t$ summarizes everything in the past that is relevant for determining future values of $\boldsymbol{y}$,

$$\begin{aligned}&E(\boldsymbol{y}_{t+m} | \boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \ldots, \boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots) \\ &\quad = E[(\boldsymbol{A}'\boldsymbol{x}_{t+m} + \boldsymbol{H}'\boldsymbol{\xi}_{t+m} + \boldsymbol{w}_{t+m}) | \boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \ldots, \boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots] \\ &\quad = \boldsymbol{A}'\boldsymbol{x}_{t+m} + \boldsymbol{H}'E(\boldsymbol{\xi}_{t+m} | \boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}, \ldots, \boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots) \\ &\quad = \boldsymbol{A}'\boldsymbol{x}_{t+m} + \boldsymbol{H}'\boldsymbol{F}^m\boldsymbol{\xi}_t. \end{aligned} \tag{1.7}$$

As a simple example of a system that can be written in state-space form, consider a pth-order autoregression

$$(y_{t+1} - \mu) = \phi_1(y_t - \mu) + \phi_2(y_{t-1} - \mu) + \cdots + \phi_p(y_{t-p+1} - \mu) + \varepsilon_{t+1}, \tag{1.8}$$
$$\varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2).$$

Note that (1.8) can equivalently be written as

$$\begin{bmatrix} y_{t+1} - \mu \\ y_t - \mu \\ \vdots \\ y_{t-p+2} - \mu \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{bmatrix} + \begin{bmatrix} \varepsilon_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{1.9}$$

The first row of (1.9) simply reproduces (1.8) and other rows assert the identity $y_{t-j} - \mu \equiv y_{t-j} - \mu$ for $j = 0, 1, \ldots, p-2$. Equation (1.9) is of the form of (1.4) with $r = p$ and

$$\boldsymbol{\xi}_t = (y_t - \mu, y_{t-1} - \mu, \ldots, y_{t-p+1} - \mu)', \tag{1.10}$$

$$\boldsymbol{v}_{t+1} = (\varepsilon_{t+1}, 0, \ldots, 0)', \tag{1.11}$$

$$\boldsymbol{F} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{1.12}$$

The observation equation is

$$\boldsymbol{y}_t = \mu + \boldsymbol{H}'\boldsymbol{\xi}_t, \tag{1.13}$$

where $\boldsymbol{H}'$ is the first row of the $(p \times p)$ identity matrix. The eigenvalues of $\boldsymbol{F}$ can be shown to satisfy

$$\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \cdots - \phi_{p-1}\lambda - \phi_p = 0; \tag{1.14}$$

thus stability of a pth-order autoregression requires that any value λ satisfying (1.14) lies inside the unit circle.

Let us now ask what kind of dynamic system would be described if $\boldsymbol{H}'$ in (1.13)

is replaced with a general $(1 \times p)$ vector,

$$y_t = \mu + [1 \;\; \theta_1 \;\; \theta_2 \;\; \cdots \;\; \theta_{p-1}]\xi_t, \tag{1.15}$$

where the θ's represent arbitrary coefficients. Suppose that ξ_t continues to evolve in the manner specified for the state vector of an AR(p) process. Letting ξ_{jt} denote the jth element of ξ_t, this would mean

$$\begin{bmatrix} \xi_{1,t+1} \\ \xi_{2,t+1} \\ \vdots \\ \xi_{p,t+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \xi_{1t} \\ \xi_{2t} \\ \vdots \\ \xi_{pt} \end{bmatrix} + \begin{bmatrix} \varepsilon_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{1.16}$$

The jth row of this system for $j = 2, 3, \ldots, p$ states that $\xi_{j,t+1} = \xi_{j-1,t}$, implying

$$\xi_{jt} = L^j \xi_{1,t+1} \qquad \text{for} \quad j = 1, 2, \ldots, p, \tag{1.17}$$

for L the lag operator. The first row of (1.16) thus implies that the first element of ξ_t can be viewed as an AR(p) process driven by the innovations sequence $\{\varepsilon_t\}$:

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)\xi_{1,t+1} = \varepsilon_{t+1}. \tag{1.18}$$

Equations (1.15) and (1.17) then imply

$$y_t = \mu + (1 + \theta_1 L^1 + \theta_2 L^2 + \cdots + \theta_{p-1} L^{p-1})\xi_{1t}. \tag{1.19}$$

If we subtract μ from both sides of (1.19) and operate on both sides with $(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)$, the result is

$$\begin{aligned}(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)(y_t - \mu) &= (1 + \theta_1 L^1 + \theta_2 L^2 + \cdots + \theta_{p-1} L^{p-1}) \\ &\quad \times (1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)\xi_{1t} \\ &= (1 + \theta_1 L^1 + \theta_2 L^2 + \cdots + \theta_{p-1} L^{p-1})\varepsilon_t\end{aligned} \tag{1.20}$$

by virtue of (1.18). Thus equations (1.15) and (1.16) constitute a state-space representation for an ARMA($p, p-1$) process.

The state-space framework can also be used in its own right as a parsimonious time-series description of an observed vector of variables. The usefulness of forecasts emerging from this approach has been demonstrated by Harvey and Todd (1983), Aoki (1987), and Harvey (1989).

The state-space form is particularly convenient for thinking about sums of stochastic processes or the consequences of measurement error. For example, suppose we postulate the existence of an underlying "true" variable, ξ_t, that follows an AR(1) process

$$\xi_t = \phi \xi_{t-1} + v_t, \tag{1.21}$$

with v_t white noise. Suppose that ξ_t is not observed directly. Instead, the econometrician has available data y_t that differ from ξ_t by measurement error w_t:

$$y_t = \xi_t + w_t. \tag{1.22}$$

If the measurement error is white noise that is uncorrelated with v_t, then (1.21) and (1.22) can immediately be viewed as the state equation and observation equation of a state-space system, with $r = n = 1$. Fama and Gibbons (1982) used just such a model to describe the ex ante real interest rate (the nominal interest rate i_t minus the expected inflation rate π_t^e). The ex ante real rate is presumed to follow an AR(1) process, but is unobserved by the econometrician because people's expectation π_t^e is unobserved. The state vector for this application is then $\xi_t = i_t - \pi_t^e - \mu$ where μ is the average ex ante real interest rate. The observed ex post real rate $(y_t = i_t - \pi_t)$ differs from the ex ante real rate by the error people make in forecasting inflation,

$$i_t - \pi_t = \mu + (i_t - \pi_t^e - \mu) + (\pi_t^e - \pi_t),$$

which is an observation equation of the form of (1.6) with $\boldsymbol{H}' = 1$ and $w_t = (\pi_t^e - \pi_t)$. If people do not make systematic errors in forecasting inflation, then w_t might reasonably be assumed to be white noise.

In many economic models, the public's expectations of the future have important consequences. These expectations are not observed directly, but if they are formed rationally there are certain implications for the time-series behavior of observed series. Thus the rational-expectations hypothesis lends itself quite naturally to a state-space representation; sample applications include Wall (1980), Burmeister and Wall (1982), Watson (1989), and Imrohoroglu (1993).

In another interesting econometric application of a state-space representation, Stock and Watson (1991) postulated that the common dynamic behavior of an $(n \times 1)$ vector of macroeconomic variables y_t could be explained in terms of an unobserved scalar c_t, which is viewed as the state of the business cycle. In addition, each series y_{it} is presumed to have an idiosyncratic component (denoted a_{it}) that is unrelated to movements in y_{jt} for $i \neq j$. If each of the component processes could be described by an AR(1) process, then the $[(n+1) \times 1]$ state vector would be

$$\xi_t = (c_t, a_{1t}, a_{2t}, \ldots, a_{nt})' \tag{1.23}$$

with state equation

$$\begin{bmatrix} c_{t+1} \\ a_{1,t+1} \\ a_{2,t+1} \\ \vdots \\ a_{n,t+1} \end{bmatrix} = \begin{bmatrix} \phi_c & 0 & 0 & \cdots & 0 \\ 0 & \phi_1 & 0 & \cdots & 0 \\ 0 & 0 & \phi_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \phi_n \end{bmatrix} \begin{bmatrix} c_t \\ a_{1t} \\ a_{2t} \\ \vdots \\ a_{nt} \end{bmatrix} + \begin{bmatrix} v_{c,t+1} \\ v_{1,t+1} \\ v_{2,t+1} \\ \vdots \\ v_{n,t+1} \end{bmatrix} \tag{1.24}$$

and observation equation

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} + \begin{bmatrix} \gamma_1 & 1 & 0 & \cdots & 0 \\ \gamma_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma_n & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} c_t \\ a_{1t} \\ a_{2t} \\ \vdots \\ a_{nt} \end{bmatrix}. \tag{1.25}$$

Thus γ_i is a parameter measuring the sensitivity of the ith series to the business cycle. To allow for pth-order dynamics, Stock and Watson replaced c_t and a_{it} in (1.23) with the $(1 \times p)$ vectors $(c_t, c_{t-1}, \ldots, c_{t-p+1})$ and $(a_{it}, a_{i,t-1}, \ldots, a_{i,t-p+1})$ so that ξ_t is an $[(n+1)p \times 1]$ vector. The scalars ϕ_i in (1.24) are then replaced by $(p \times p)$ matrices F_i with the structure of (1.12), and blocks of zeros are added in between the columns of H' in the observation equation (1.25). A related theoretical model was explored by Sargent (1989).

State-space models have seen many other applications in economics. For partial surveys see Engle and Watson (1987), Harvey (1987), and Aoki (1987).

2. The Kalman filter

For convenience, the general form of a constant-parameter linear state-space model is reproduced here as equations (2.1) and (2.2).

State equation

$$\underset{(r\times 1)}{\xi_{t+1}} = \underset{(r\times r)(r\times 1)}{F\xi_t} + \underset{(r\times 1)}{v_{t+1}} \tag{2.1}$$

$$E(v_{t+1}v_{t+1}') = \underset{(r\times r)}{Q}$$

Observation equation

$$\underset{(n\times 1)}{y_t} = \underset{(n\times k)(k\times 1)}{A'x_t} + \underset{(n\times r)(r\times 1)}{H'\xi_t} + \underset{(n\times 1)}{w_t} \tag{2.2}$$

$$E(w_t w_t') = \underset{(n\times n)}{R}.$$

Writing a model in state-space form means imposing certain values (such as zero or one) on some of the elements of F, Q, A, H and R, and interpreting the other elements as particular parameters of interest. Typically we will not know the values of these other elements, but need to estimate them on the basis of observation of $\{y_1, y_2, \ldots, y_T\}$ and $\{x_1, x_2, \ldots, x_T\}$.

2.1. *Overview of the Kalman filter*

Before discussing estimation of parameters, it will be helpful first to assume that the values of all of the elements of F, Q, A, H and R are known with certainty; the question of estimation is postponed until Section 3. The filter named for the contributions of Kalman (1960, 1963) can be described as an algorithm for calculating an optimal forecast of the value of ξ_t on the basis of information observed through date $t-1$, assuming that the values of F, Q, A, H and R are all known.

This optimal forecast is derived from a well-known result for normal variables; [see, for example, DeGroot (1970, p. 55)]. Let z_1 and z_2 denote $(n_1 \times 1)$ and $(n_2 \times 1)$ vectors respectively that have a joint normal distribution:

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} \right).$$

Then the distribution of z_2 conditional on z_1 is $N(m, \Sigma)$ where

$$m = \mu_2 + \Omega_{21}\Omega_{11}^{-1}(z_1 - \mu_1), \tag{2.3}$$

$$\Sigma = \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}. \tag{2.4}$$

Thus the optimal forecast of z_2 conditional on having observed z_1 is given by

$$E(z_2|z_1) = \mu_2 + \Omega_{21}\Omega_{11}^{-1}(z_1 - \mu_1), \tag{2.5}$$

with Σ characterizing the mean squared error of this forecast:

$$E[(z_2 - m)(z_2 - m)'|z_1] = \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}. \tag{2.6}$$

To apply this result, suppose that the initial value of the state vector (ξ_1) of a state-space model is drawn from a normal distribution and that the disturbances v_t and w_t are normal. Let the observed data obtained through date $t-1$ be summarized by the vector

$$\zeta_{t-1} \equiv (y_{t-1}', y_{t-2}', \ldots, y_1', x_{t-1}', x_{t-2}', \ldots, x_1')'.$$

Then the distribution of ξ_t conditional on ζ_{t-1} turns out to be normal for $t = 2, 3, \ldots, T$. The mean of this conditional distribution is represented by the $(r \times 1)$ vector $\hat{\xi}_{t|t-1}$ and the variance of this conditional distribution is represented by the $(r \times r)$ matrix $\boldsymbol{P}_{t|t-1}$. The Kalman filter is simply the result of applying (2.5) and (2.6) to each observation in the sample in succession. The input for step t of the iteration is the mean $\hat{\xi}_{t|t-1}$ and variance $\boldsymbol{P}_{t|t-1}$ that characterize the distribution of ξ_t conditional on ζ_{t-1}. The output for step t is the mean $\hat{\xi}_{t+1|t}$ and variance $\boldsymbol{P}_{t+1|t}$ of ξ_{t+1} conditional on ζ_t. Thus the output for step t is used as the input for step $t+1$.

2.2. *Derivation of the Kalman filter*

The iteration is started by assuming that the initial value of the state vector ξ_1 is drawn from a normal distribution with mean denoted $\hat{\xi}_{1|0}$ and variance denoted $\boldsymbol{P}_{1|0}$. If the eigenvalues of $\boldsymbol{F}$ are all inside the unit circle, then the vector process defined by (2.1) is stationary, and $\hat{\xi}_{1|0}$ would be the unconditional mean of this process,

$$\hat{\xi}_{1|0} = \mathbf{0}, \tag{2.7}$$

while $\boldsymbol{P}_{1|0}$ would be the unconditional variance

$$\boldsymbol{P}_{1|0} = E(\xi_t \xi_t').$$

This unconditional variance can be calculated from[1]

$$\operatorname{vec}(\boldsymbol{P}_{1|0}) = [\boldsymbol{I}_{r^2} - (\boldsymbol{F} \otimes \boldsymbol{F})]^{-1} \cdot \operatorname{vec}(\boldsymbol{Q}). \tag{2.8}$$

Here $\boldsymbol{I}_{r^2}$ is the $(r^2 \times r^2)$ identity matrix, "$\otimes$" denotes the Kronecker product and

[1] The unconditional variance of ξ can be found by postmultiplying (2.1) by its transpose and taking expectations:

$$\begin{aligned} E(\xi_{t+1}\xi_{t+1}') &= E(\boldsymbol{F}\xi_t + v_{t+1})(\xi_t'\boldsymbol{F}' + v_{t+1}') \\ &= \boldsymbol{F} \cdot E(\xi_t \xi_t')\boldsymbol{F}' + E(v_{t+1}v_{t+1}'). \end{aligned}$$

If ξ_t is stationary, then $E(\xi_{t+1}\xi_{t+1}') = E(\xi_t\xi_t') = \boldsymbol{P}_{1|0}$, and the above equation becomes

$$\boldsymbol{P}_{1|0} = \boldsymbol{F}\boldsymbol{P}_{1|0}\boldsymbol{F}' + \boldsymbol{Q}.$$

Applying the vec operator to this equation and recalling [e.g. Magnus and Neudecker (1988, p. 30)] that $\operatorname{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}' \otimes \boldsymbol{A}) \cdot \operatorname{vec}(\boldsymbol{B})$ produces

$$\operatorname{vec}(\boldsymbol{P}_{1|0}) = (\boldsymbol{F} \otimes \boldsymbol{F}) \cdot \operatorname{vec}(\boldsymbol{P}_{1|0}) + \operatorname{vec}(\boldsymbol{Q}).$$

$\operatorname{vec}(\boldsymbol{P}_{1|0})$ is the $(r^2 \times 1)$ vector formed by stacking the columns of $\boldsymbol{P}_{1|0}$, one on top of the other, ordered from left to right.

For time-variant or nonstationary systems, $\hat{\xi}_{1|0}$ could represent a guess as to the value of ξ_1 based on prior information, while $\boldsymbol{P}_{1|0}$ measures the uncertainty associated with this guess – the greater our prior uncertainty, the larger the diagonal elements of $\boldsymbol{P}_{1|0}$.[2] This prior cannot be based on the data, since it is assumed in the derivations to follow that v_{t+1} and w_t are independent of ξ_1 for $t = 1, 2, \ldots, T$. The algorithm described below can also be adapted for the case of a completely diffuse prior (the limiting case when $\boldsymbol{P}_{1|0}$ becomes infinite); as described by Ansley and Kohn (1985), Kohn and Ansley (1986) and De Jong (1988, 1989, 1991).

At this point we have described the values of $\hat{\xi}_{t|t-1}$ and $\boldsymbol{P}_{t|t-1}$ that characterize the distribution of ξ_t conditional on ζ_{t-1} for $t = 1$. Since a similar set of calculations will be used for each date t in the sample, it is helpful to describe the next step using notation appropriate for an arbitrary date t. Thus let us assume that the values of $\hat{\xi}_{t|t-1}$ and $\boldsymbol{P}_{t|t-1}$ have been calculated for some t, and undertake the task of using these to evaluate $\hat{\xi}_{t+1|t}$ and $\boldsymbol{P}_{t+1|t}$. If the distribution of ξ_t conditional on ζ_{t-1} is $N(\hat{\xi}_{t|t-1}, \boldsymbol{P}_{t|t-1})$, then under the assumptions about x_t, this is the same as the distribution of ξ_t conditional on ζ_{t-1} and x_t. Since w_t is independent of x_t and ζ_{t-1}, the forecast of y_t conditional on ζ_{t-1} and x_t can be inferred immediately from (2.2):

$$E(y_t | x_t, \zeta_{t-1}) = \boldsymbol{A}'x_t + \boldsymbol{H}'\hat{\xi}_{t|t-1}. \tag{2.9}$$

From (2.2) and (2.9) the forecast error can be written

$$\begin{aligned} y_t - E(y_t|x_t, \zeta_{t-1}) &= (\boldsymbol{A}'x_t + \boldsymbol{H}'\xi_t + w_t) - (\boldsymbol{A}'x_t + \boldsymbol{H}'\hat{\xi}_{t|t-1}) \\ &= \boldsymbol{H}'(\xi_t - \hat{\xi}_{t|t-1}) + w_t. \end{aligned} \tag{2.10}$$

Since $\hat{\xi}_{t|t-1}$ is a function of ζ_{t-1}, the term w_t is independent of both ξ_t and $\hat{\xi}_{t|t-1}$. Thus the conditional variance of (2.10) is

$$\begin{aligned} &E\{[y_t - E(y_t|x_t, \zeta_{t-1})][y_t - E(y_t|x_t, \zeta_{t-1})]'|x_t, \zeta_{t-1}\} \\ &\quad = \boldsymbol{H}' \cdot E\{[\xi_t - \hat{\xi}_{t|t-1}][\xi_t - \hat{\xi}_{t|t-1}]'|\zeta_{t-1}\}\boldsymbol{H} + E(w_t w_t') \\ &\quad = \boldsymbol{H}'\boldsymbol{P}_{t|t-1}\boldsymbol{H} + \boldsymbol{R}. \end{aligned}$$

Similarly, the conditional covariance between (2.10) and the error in forecasting

[2] Meinhold and Singpurwalla (1983) gave a nice description of the Kalman filter from a Bayesian perspective.

the state vector is

$$E\{[y_t - E(y_t|x_t, \zeta_{t-1})][\xi_t - E(\xi_t|x_t, \zeta_{t-1})]'|x_t, \zeta_{t-1}\}$$
$$= H' \cdot E\{[\xi_t - \hat{\xi}_{t|t-1}][\xi_t - \hat{\xi}_{t|t-1}]'|x_t, \zeta_{t-1}\}$$
$$= H' P_{t|t-1}.$$

Thus the distribution of the vector $(y_t', \xi_t')'$ conditional on x_t and ζ_{t-1} is

$$\begin{bmatrix} y_t|x_t, \zeta_{t-1} \\ \xi_t|x_t, \zeta_{t-1} \end{bmatrix} \sim N\left(\begin{bmatrix} A'x_t + H'\hat{\xi}_{t|t-1} \\ \hat{\xi}_{t|t-1} \end{bmatrix}, \begin{bmatrix} (H'P_{t|t-1}H + R) & H'P_{t|t-1} \\ P_{t|t-1}H & P_{t|t-1} \end{bmatrix}\right). \quad (2.11)$$

It then follows from (2.3) and (2.4) that $\xi_t|\zeta_t = \xi_t|x_t, y_t, \zeta_{t-1}$ is distributed $N(\hat{\xi}_{t|t}, P_{t|t})$ where

$$\hat{\xi}_{t|t} = \hat{\xi}_{t|t-1} + P_{t|t-1}H(H'P_{t|t-1}H + R)^{-1}(y_t - A'x_t - H'\hat{\xi}_{t|t-1}), \quad (2.12)$$

$$P_{t|t} = P_{t|t-1} - P_{t|t-1}H(H'P_{t|t-1}H + R)^{-1}H'P_{t|t-1}. \quad (2.13)$$

The final step is to calculate a forecast of ξ_{t+1} conditional on ζ_t. It is not hard to see from (2.1) that $\xi_{t+1}|\zeta_t \sim N(\hat{\xi}_{t+1|t}, P_{t+1|t})$ where

$$\hat{\xi}_{t+1|t} = F\hat{\xi}_{t|t}, \quad (2.14)$$

$$P_{t+1|t} = FP_{t|t}F' + Q. \quad (2.15)$$

Substituting (2.12) into (2.14) and (2.13) into (2.15), we have

$$\hat{\xi}_{t+1|t} = F\hat{\xi}_{t|t-1} + FP_{t|t-1}H(H'P_{t|t-1}H + R)^{-1}(y_t - A'x_t - H'\hat{\xi}_{t|t-1}), \quad (2.16)$$

$$P_{t+1|t} = FP_{t|t-1}F' - FP_{t|t-1}H(H'P_{t|t-1}H + R)^{-1}H'P_{t|t-1}F' + Q. \quad (2.17)$$

To summarize, the Kalman filter is an algorithm for calculating the sequence $\{\hat{\xi}_{t+1|t}\}_{t=1}^{T}$ and $\{P_{t+1|t}\}_{t=1}^{T}$, where $\hat{\xi}_{t+1|t}$ denotes the optimal forecast of ξ_{t+1} based on observation of $(y_t, y_{t-1}, \ldots, y_1, x_t, x_{t-1}, \ldots, x_1)$ and $P_{t+1|t}$ denotes the mean squared error of this forecast. The filter is implemented by iterating on (2.16) and (2.17) for $t = 1, 2, \ldots, T$. If the eigenvalues of F are all inside the unit circle and there is no prior information about the initial value of the state vector, this iteration is started using equations (2.7) and (2.8).

Note that the sequence $\{P_{t+1|t}\}_{t=1}^{T}$ is not a function of the data and can be evaluated without calculating the forecasts $\{\hat{\xi}_{t+1|t}\}_{t=1}^{T}$. Because $P_{t+1|t}$ is not a function of the data, the conditional expectation of the squared forecast error is

the same as its unconditional expectation,

$$P_{t+1|t} = E\{(\xi_{t+1} - \hat{\xi}_{t+1|t})(\xi_{t+1} - \hat{\xi}_{t+1|t})'|\zeta_t\}$$
$$= E\{(\xi_{t+1} - \hat{\xi}_{t+1|t})(\xi_{t+1} - \hat{\xi}_{t+1|t})'\}.$$

This equivalence is a consequence of having assumed normal distributions with constant variances for v_t and w_t.

2.3. *Forecasting with the Kalman filter*

An m-period-ahead forecast of the state vector can be calculated from (1.5):

$$\hat{\xi}_{t+m|t} = E(\xi_{t+m}|y_t, y_{t-1}, \ldots, y_1, x_t, x_{t-1}, \ldots, x_1) = F^m\hat{\xi}_{t|t}. \quad (2.18)$$

The error of this forecast can be found by subtracting (2.18) from (1.5),

$$\xi_{t+m} - \hat{\xi}_{t+m|t} = F^m(\xi_t - \hat{\xi}_{t|t}) + F^{m-1}v_{t+1} + F^{m-2}v_{t+2} + \cdots + F^1 v_{t+m-1} + v_{t+m},$$

from which it follows that the mean squared error of the forecast (2.18) is

$$P_{t+m|t} = E[(\xi_{t+m} - \hat{\xi}_{t+m|t})(\xi_{t+m} - \hat{\xi}_{t+m|t})']$$
$$= F^m P_{t|t}(F^m)' + F^{m-1}Q(F^{m-1})' + F^{m-2}Q(F^{m-2})' + \cdots + FQF' + Q. \quad (2.19)$$

These results can also be used to describe m-period-ahead forecasts of the observed vector y_{t+m}, provided that $\{x_t\}$ is deterministic. Applying the law of iterated expectations to (1.7) results in

$$\hat{y}_{t+m|t} = E(y_{t+m}|y_t, y_{t-1}, \ldots, y_1) = A'x_{t+m} + H'F^m\hat{\xi}_{t|t}. \quad (2.20)$$

The error of this forecast is

$$y_{t+m} - \hat{y}_{t+m|t} = (A'x_{t+m} + H'\xi_{t+m} + w_{t+m}) - (A'x_{t+m} + H'F^m\hat{\xi}_{t|t})$$
$$= H'(\xi_{t+m} - \hat{\xi}_{t+m|t}) + w_{t+m}$$

with mean squared error

$$E[(y_{t+m} - \hat{y}_{t+m|t})(y_{t+m} - \hat{y}_{t+m|t})'] = H'P_{t+m|t}H + R. \quad (2.21)$$

2.4. *Smoothed inference*

Up to this point we have been concerned with a forecast of the value of the state vector at date t based on information available at date $t-1$, denoted $\hat{\xi}_{t|t-1}$, or

with an inference about the value of the state vector at date t based on currently available information, denoted $\hat{\xi}_{t|t}$. In some applications the value of the state vector is of interest in its own right. In the example of Fama and Gibbons, the state vector tells us about the public's expectations of inflation, while in the example of Stock and Watson, it tells us about the overall condition of the economy. In such cases it is desirable to use information through the end of the sample (date T) to help improve the inference about the historical value that the state vector took on at any particular date t in the middle of the sample. Such an inference is known as a *smoothed* estimate, denoted $\hat{\xi}_{t|T} \equiv E(\xi_t|\zeta_T)$. The mean squared error of this estimate is denoted $P_{t|T} = E(\xi_t - \hat{\xi}_{t|T})(\xi_t - \hat{\xi}_{t|T})'$.

The smoothed estimates can be calculated as follows. First we run the data through the Kalman filter, storing the sequences $\{P_{t|t}\}_{t=1}^{T}$ and $\{P_{t|t-1}\}_{t=1}^{T}$ as calculated from (2.13) and (2.15), and storing the sequences $\{\hat{\xi}_{t|t}\}_{t=1}^{T}$ and $\{\hat{\xi}_{t|t-1}\}_{t=1}^{T}$ as calculated from (2.12) and (2.14). The terminal value for $\{\hat{\xi}_{t|t}\}_{t=1}^{T}$ then gives the smoothed estimate for the last date in the sample, $\hat{\xi}_{T|T}$, and $P_{T|T}$ is its mean squared error.

The sequence of smoothed estimates $\{\hat{\xi}_{t|T}\}_{t=1}^{T}$ is then calculated in reverse order by iterating on

$$\hat{\xi}_{t|T} = \hat{\xi}_{t|t} + J_t(\hat{\xi}_{t+1|T} - \hat{\xi}_{t+1|t}) \tag{2.22}$$

for $t = T-1, T-2, \ldots, 1$, where $J_t = P_{t|t}F'P_{t+1|t}^{-1}$. The corresponding mean squared errors are similarly found by iterating on

$$P_{t|T} = P_{t|t} + J_t(P_{t+1|T} - P_{t+1|t})J_t' \tag{2.23}$$

in reverse order for $t = T-1, T-2, \ldots, 1$; see for example Hamilton (1994, Section 13.6).

2.5. *Interpretation of the Kalman filter with non-normal disturbances*

In motivating the Kalman filter, the assumption was made that v_t and w_t were normal. Under this assumption, $\hat{\xi}_{t|t-1}$ is the function of ζ_{t-1} that minimizes

$$E[(\xi_t - \hat{\xi}_{t|t-1})(\xi_t - \hat{\xi}_{t|t-1})'], \tag{2.24}$$

in the sense that any other forecast has a mean squared error matrix that differs from that of $\hat{\xi}_{t|t-1}$ by a positive semidefinite matrix. This optimal forecast turned out to be a constant plus a linear function of ζ_{t-1}. The minimum value achieved for (2.24) was denoted $P_{t|t-1}$.

If v_t and w_t are not normal, one can pose a related problem of choosing $\hat{\xi}_{t|t-1}$ to be a constant plus a linear function of ζ_{t-1} that minimizes (2.24). The solution

to this problem turns out to be given by the Kalman filter iteration (2.16), and its unconditional mean squared error is still given by (2.17). Similarly, when the disturbances are not normal, expression (2.20) can be interpreted as the linear projection of y_{t+m} on ζ_t and a constant, with (2.21) its unconditional mean squared error. Thus, while the Kalman filter forecasts need no longer be optimal for systems that are not normal, no other forecast based on a linear function of ζ_t will have a smaller mean squared error [see Anderson and Moore (1979, pp. 92–98) or Hamilton (1994, Section 13.2)]. These results parallel the Gauss–Markov theorem for ordinary least squares regression.

2.6. *Time-varying coefficient models*

The analysis above treated the coefficients of the matrices F, Q, A, H and R as known constants. An interesting generalization obtains if these are known functions of x_t:

$$\xi_{t+1} = F(x_t)\xi_t + v_{t+1}, \tag{2.25}$$
$$E(v_{t+1}v_{t+1}'|\zeta_t) = Q(x_t),$$

$$y_t = a(x_t) + [H(x_t)]'\xi_t + w_t, \tag{2.26}$$
$$E(w_t w_t'|x_t, \zeta_{t-1}) = R(x_t).$$

Here $F(\cdot)$, $Q(\cdot)$, $H(\cdot)$ and $R(\cdot)$ denote matrix-valued functions of x_t and $a(\cdot)$ is an $(n \times 1)$ vector-valued function of x_t. As before, we assume that, apart from the possible conditional heteroskedasticity allowed in (2.26), x_t provides no information about ξ_τ or w_τ for any τ beyond that contained in ζ_{t-1}.

Even if v_t and w_t are normal, with x_t stochastic the unconditional distributions of ξ_t and y_t are no longer normal. However, the system is conditionally normal in the following sense.[3] Suppose that the distribution of ξ_t conditional on ζ_{t-1} is taken to be $N(\hat{\xi}_{t|t-1}, P_{t|t-1})$. Then ξ_t conditional on x_t and ζ_{t-1} has the same distribution. Moreover, conditional on x_t, all of the matrices can be treated as deterministic. Hence the derivation of the Kalman filter goes through essentially as before, with the recursions (2.16) and (2.17) replaced with

$$\hat{\xi}_{t+1|t} = F(x_t)\hat{\xi}_{t|t-1} + F(x_t)P_{t|t-1}H(x_t)\{[H(x_t)]'P_{t|t-1}H(x_t) + R(x_t)\}^{-1} \times \{y_t - a(x_t) - [H(x_t)]'\hat{\xi}_{t|t-1}\}, \tag{2.27}$$

$$P_{t+1|t} = F(x_t)P_{t|t-1}F(x_t)' - \{F(x_t)P_{t|t-1}H(x_t)[[H(x_t)]'P_{t|t-1}H(x_t) + R(x_t)]^{-1} \times [H(x_t)]'P_{t|t-1}[F(x_t)]'\} + Q(x_t). \tag{2.28}$$

[3] See Theorem 6.1 in Tjøstheim (1986) for further discussion.

It is worth noting three elements of the earlier discussion that change with time-varying parameter matrices. First, the distribution calculated for the initial state in (2.7) and (2.8) is only valid if $\boldsymbol{F}$ and $\boldsymbol{Q}$ are fixed matrices. Second, m-period-ahead forecasts of $\boldsymbol{y}_{t+m}$ or $\boldsymbol{\xi}_{t+m}$ for $m>1$ are no longer simple to calculate when $\boldsymbol{F}$, $\boldsymbol{H}$ or $\boldsymbol{A}$ vary stochastically; Doan et al. (1984) suggested approximating $E(\boldsymbol{y}_{t+2}|\boldsymbol{y}_t,\boldsymbol{y}_{t-1},\ldots,\boldsymbol{y}_1)$ with $E(\boldsymbol{y}_{t+2}|\boldsymbol{y}_{t+1},\boldsymbol{y}_t,\ldots,\boldsymbol{y}_1)$ evaluated at $\boldsymbol{y}_{t+1}=E(\boldsymbol{y}_{t+1}|\boldsymbol{y}_t,\boldsymbol{y}_{t-1},\ldots,\boldsymbol{y}_1)$. Finally, if $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ are not normal, then the one-period-ahead forecasts $\hat{\boldsymbol{\xi}}_{t+1|t}$ and $\hat{\boldsymbol{y}}_{t+1|t}$ no longer have the interpretation as linear projections, since (2.27) is nonlinear in $\boldsymbol{x}_t$.

An important application of a state-space representation with data-dependent parameter matrices is the time-varying coefficient regression model

$$\boldsymbol{y}_t=\boldsymbol{x}_t'\boldsymbol{\beta}_t+\boldsymbol{w}_t. \tag{2.29}$$

Here $\boldsymbol{\beta}_t$ is a vector of regression coefficients that is assumed to evolve over time according to

$$(\boldsymbol{\beta}_{t+1}-\bar{\boldsymbol{\beta}})=\boldsymbol{F}(\boldsymbol{\beta}_t-\bar{\boldsymbol{\beta}})+\boldsymbol{v}_{t+1}. \tag{2.30}$$

Assuming the eigenvalues of $\boldsymbol{F}$ are all inside the unit circle, $\bar{\boldsymbol{\beta}}$ has the interpretation as the average or steady-state coefficient vector. Equation (2.30) will be recognized as a state equation of the form of (2.1) with $\boldsymbol{\xi}_t=(\boldsymbol{\beta}_t-\bar{\boldsymbol{\beta}})$. Equation (2.29) can then be written as

$$\boldsymbol{y}_t=\boldsymbol{x}_t'\bar{\boldsymbol{\beta}}+\boldsymbol{x}_t'\boldsymbol{\xi}_t+\boldsymbol{w}_t, \tag{2.31}$$

which is in the form of the observation equation (2.26) with $\boldsymbol{a}(\boldsymbol{x}_t)=\boldsymbol{x}_t'\bar{\boldsymbol{\beta}}$ and $[\boldsymbol{H}(\boldsymbol{x}_t)]'=\boldsymbol{x}_t'$. Higher-order dynamics for $\boldsymbol{\beta}_t$ are easily incorporated by, instead, defining $\boldsymbol{\xi}_t'\equiv[(\boldsymbol{\beta}_t-\bar{\boldsymbol{\beta}})',(\boldsymbol{\beta}_{t-1}-\bar{\boldsymbol{\beta}})',\ldots,(\boldsymbol{\beta}_{t-p+1}-\bar{\boldsymbol{\beta}})']$ as in Nicholls and Pagan (1985, p. 437).

Excellent surveys of time-varying parameter regressions include Raj and Ullah (1981), Chow (1984) and Nicholls and Pagan (1985). Applications to vector autoregressions have been explored by Sims (1982) and Doan et al. (1984).

2.7. *Other extensions*

The derivations above assumed no correlation between $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$, though this is straightforward to generalize; see, for example, Anderson and Moore (1979, p. 108). Predetermined or exogenous variables can also be added to the state equation with few adjustments.

The Kalman filter is a very convenient algorithm for handling missing observations. If $\boldsymbol{y}_t$ is unobserved for some date t, one can simply skip the updating

equations (2.12) and (2.13) for that date and replace them with $\hat{\boldsymbol{\xi}}_{t|t}=\hat{\boldsymbol{\xi}}_{t|t-1}$ and $\boldsymbol{P}_{t|t}=\boldsymbol{P}_{t|t-1}$; see Jones (1980), Harvey and Pierse (1984) and Kohn and Ansley (1986) for further discussion. Modifications of the Kalman filtering and smoothing algorithms to allow for singular or infinite $\boldsymbol{P}_{t|t}$ are described in De Jong (1989, 1991).

3. Statistical inference about unknown parameters using the Kalman filter

3.1. *Maximum likelihood estimation*

The calculations described in Section 2 are implemented by computer, using the known numerical values for the coefficients in the matrices $\boldsymbol{F}$, $\boldsymbol{Q}$, $\boldsymbol{A}$, $\boldsymbol{H}$ and $\boldsymbol{R}$. When the values of the matrices are unknown we can proceed as follows. Collect the unknown elements of these matrices in a vector $\boldsymbol{\theta}$. For example, to estimate the ARMA$(p,p-1)$ process (1.15)–(1.16), $\boldsymbol{\theta}=(\phi_1,\phi_2,\ldots,\phi_p,\theta_1,\theta_2,\ldots,\theta_{p-1},\mu,\sigma)'$. Make an arbitrary initial guess as to the value of $\boldsymbol{\theta}$, denoted $\boldsymbol{\theta}^{(0)}$, and calculate the sequences $\{\hat{\boldsymbol{\xi}}_{t|t-1}(\boldsymbol{\theta}^{(0)})\}_{t=1}^{T}$ and $\{\boldsymbol{P}_{t|t-1}(\boldsymbol{\theta}^{(0)})\}_{t=1}^{T}$ that result from this value in (2.16) and (2.17). Recall from (2.11) that if the data were really generated from the model (2.1)–(2.2) with this value of $\boldsymbol{\theta}$, then

$$\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\zeta}_{t-1};\boldsymbol{\theta}^{(0)}\sim N(\boldsymbol{\mu}_t(\boldsymbol{\theta}^{(0)}),\boldsymbol{\Sigma}_t(\boldsymbol{\theta}^{(0)})), \tag{3.1}$$

where

$$\boldsymbol{\mu}_t(\boldsymbol{\theta}^{(0)})=[\boldsymbol{A}(\boldsymbol{\theta}^{(0)})]'\boldsymbol{x}_t+[\boldsymbol{H}(\boldsymbol{\theta}^{(0)})]'\hat{\boldsymbol{\xi}}_{t|t-1}(\boldsymbol{\theta}^{(0)}), \tag{3.2}$$

$$\boldsymbol{\Sigma}_t(\boldsymbol{\theta}^{(0)})=[\boldsymbol{H}(\boldsymbol{\theta}^{(0)})]'[\boldsymbol{P}_{t|t-1}(\boldsymbol{\theta}^{(0)})][\boldsymbol{H}(\boldsymbol{\theta}^{(0)})]+\boldsymbol{R}(\boldsymbol{\theta}^{(0)}). \tag{3.3}$$

The value of the log likelihood is then

$$\sum_{t=1}^{T}\log f(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\zeta}_{t-1};\boldsymbol{\theta}^{(0)})=-\frac{Tn}{2}\log(2\pi)-\frac{1}{2}\sum_{t=1}^{T}\log|\boldsymbol{\Sigma}_t(\boldsymbol{\theta}^{(0)})| -\frac{1}{2}\sum_{t=1}^{T}[\boldsymbol{y}_t-\boldsymbol{\mu}_t(\boldsymbol{\theta}^{(0)})]'[\boldsymbol{\Sigma}_t(\boldsymbol{\theta}^{(0)})]^{-1}[\boldsymbol{y}_t-\boldsymbol{\mu}_t(\boldsymbol{\theta}^{(0)})], \tag{3.4}$$

which reflects how likely it would have been to have observed the data if $\boldsymbol{\theta}^{(0)}$ were the true value for $\boldsymbol{\theta}$. We then make an alternative guess $\boldsymbol{\theta}^{(1)}$ so as to try to achieve a bigger value of (3.4), and proceed to maximize (3.4) with respect to $\boldsymbol{\theta}$ by numerical methods such as those described in Quandt (1983), Nash and Walker-Smith (1987) or Hamilton (1994, Section 5.7).

Many numerical optimization techniques require the gradient vector, or the derivative of (3.4) with respect to $\boldsymbol{\theta}$. The derivative with respect to the ith element of $\boldsymbol{\theta}$ could be calculated numerically by making a small change in the ith element

of $\boldsymbol{\theta}^{(0)}$ and seeing what happens to (3.4). Alternatively, one can differentiate the recursions (2.16) and (2.17) analytically with respect to θ_i so as to generate an analytical expression for the sequence of $(r \times 1)$ vectors $\{\partial \hat{\boldsymbol{\xi}}_{t|t-1}(\boldsymbol{\theta})/\partial\theta_i\}_{t=1}^{T}$ and the sequence of $(r \times r)$ matrices $\{\partial \boldsymbol{P}_{t|t-1}(\boldsymbol{\theta})/\partial\theta_i\}_{t=1}^{T}$. Using these, the derivatives of $\boldsymbol{\mu}_t(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_t(\boldsymbol{\theta})$ can be found from (3.2) and (3.3); see, for example, Caines (1988, pp. 585–586).

Estimation by the method of scoring is described in Pagan (1980) and Watson and Engle (1983). Another attractive option is to use the EM algorithm developed by Shumway and Stoffer (1982) and Watson and Engle (1983).

Since models involving ARMA processes are readily cast in state-space form, expression (3.4) offers a convenient way to calculate the exact likelihood function, as stressed by Harvey and Phillips (1979). For example, consider the linear regression model

$$y_t = \boldsymbol{\beta}'\boldsymbol{x}_t + u_t, \tag{3.5}$$

where $\boldsymbol{x}_t$ is a vector of explanatory variables that is independent of u_τ for all τ. Suppose that u_t follows an MA(2) process

$$u_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}, \tag{3.6}$$

where $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. Recalling the analysis of (1.20), this can be written in state-space form with $\boldsymbol{\xi}_t = (\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2})'$:

state equation

$$\begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \\ \varepsilon_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{t+1} \\ 0 \\ 0 \end{bmatrix},$$

observation equation

$$y_t = \boldsymbol{\beta}'\boldsymbol{x}_t + [1 \quad \theta_1 \quad \theta_2] \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \\ \varepsilon_{t-2} \end{bmatrix}.$$

That is

$$\boldsymbol{F} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad \boldsymbol{Q} = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \boldsymbol{A}' = \boldsymbol{\beta}'$$

$$\boldsymbol{H}' = [1 \quad \theta_1 \quad \theta_2] \qquad \boldsymbol{R} = 0.$$

The vector of parameters to be estimated is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \theta_1, \theta_2, \sigma)'$. By making an arbitrary guess[4] at the value of $\boldsymbol{\theta}$, we can calculate the sequences $\{\hat{\boldsymbol{\xi}}_{t|t-1}(\boldsymbol{\theta})\}_{t=1}^{T}$ and $\{\boldsymbol{P}_{t|t-1}(\boldsymbol{\theta})\}_{t=1}^{T}$ in (2.16) and (2.17). The starting value for (2.16) is the unconditional mean of $\boldsymbol{\xi}_1$,

$$\hat{\boldsymbol{\xi}}_{1|0} = E\begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \\ \varepsilon_{t-2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

while (2.17) is started with the unconditional variance,

$$\boldsymbol{P}_{1|0} = E\begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \\ \varepsilon_{t-2} \end{bmatrix} [\varepsilon_t \quad \varepsilon_{t-1} \quad \varepsilon_{t-2}] = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}.$$

From these sequences, $\boldsymbol{\mu}_t(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_t(\boldsymbol{\theta})$ can be calculated in (3.2) and (3.3), and (3.4) then provides

$$\log f(y_T, y_{T-1}, \ldots, y_1 | \boldsymbol{x}_T, \boldsymbol{x}_{T-1}, \ldots, \boldsymbol{x}_1; \boldsymbol{\theta}). \tag{3.7}$$

Note that this calculation gives the exact log likelihood, not an approximation, and is valid regardless of whether θ_1 and θ_2 are associated with an invertible MA(2) representation. The parameter estimates $\hat{\boldsymbol{\beta}}, \hat{\theta}_1, \hat{\theta}_2$ and $\hat{\sigma}$ are the values that make (3.7) as large as possible.

3.2. *Identification*

The maximum likelihood estimation procedure, just described, presupposes that the model is identified, that is, it assumes that a change in any of the parameters would imply a different probability distribution for $\{y_t\}_{t=1}^{\infty}$.

One approach to checking for identification is to rewrite the state-space model in an alternative form that is better known to econometricians. For example, since the state-space model (1.15)–(1.16) is just another way of writing an ARMA$(p, p-1)$ process, the unknown parameters $(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_{p-1}, \mu, \sigma)$ can be consistently estimated provided that the roots of $(1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_{p-1}z^{p-1}) = 0$ are normalized to lie on or outside the unit circle, and are distinct from the roots of $(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) = 0$ (assuming these to lie outside the unit circle as well). An illustration of this general idea is provided in Hamilton (1985). As another

[4] Numerical algorithms are usually much better behaved if an intelligent initial guess for $\boldsymbol{\theta}^{(0)}$ is used. A good way to proceed in this instance is to use OLS estimates of (3.5) to calculate an initial guess for $\boldsymbol{\beta}$, and use the estimated variance s^2 and autocorrelations $\hat{\rho}_1$ and $\hat{\rho}_2$ of the OLS residuals to construct initial guesses for θ_1, θ_2 and σ using the results in Box and Jenkins (1976, pp. 187 and 519).

example, the time-varying coefficient regression model (2.31) can be written

$$y_t = x_t'\bar{\beta} + u_t, \tag{3.8}$$

where

$$u_t \equiv x_t'\xi_t + w_t.$$

If x_t is deterministic, equation (3.8) describes a generalized least squares regression model in which the variance–covariance matrix of the residuals can be inferred from the state equation describing ξ_t. Thus, assuming that eigenvalues of F are all inside the unit circle, $\bar{\beta}$ can be estimated consistently as long as $(1/T)\sum_{t=1}^{T} x_t x_t'$ converges to a nonsingular matrix; other parameters can be consistently estimated if higher moments of x_t satisfy certain conditions [see Nicholls and Pagan (1985, p. 431)].

The question of identification has also been extensively investigated in the literature on linear systems; see Gevers and Wertz (1984) and Wall (1987) for a survey of some of the approaches, and Burmeister et al. (1986) for an illustration of how these results can be applied.

3.3. *Asymptotic properties of maximum likelihood estimates*

Under suitable conditions, the estimate $\hat{\theta}$ that maximizes (3.4) is consistent and asymptotically normal. Typical conditions require θ to be identified, eigenvalues of F to be inside the unit circle, the exogenous variable x_t to behave asymptotically like a full rank linearly nondeterministic covariance-stationary process, and the true value of θ to not fall on the boudary of the allowable parameter space; see Caines (1988, Chapter 7) for a thorough discussion. Pagan (1980, Theorem 4) and Ghosh (1989) demonstrated that for particular examples of state-space models

$$\sqrt{T}\mathcal{I}_{2D,T}^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{L} N(0, I) \tag{3.9}$$

where $\mathcal{I}_{2D,T}$ is the information matrix for a sample of size T as calculated from second derivatives of the log likelihood function:

$$\mathcal{I}_{2D,T} = -\frac{1}{T}E\left[\sum_{t=1}^{T} \frac{\partial^2 \log f(y_t|\zeta_{t-1}, x_t; \theta)}{\partial\theta\partial\theta'}\bigg|_{\theta=\theta_0}\right]. \tag{3.10}$$

Engle and Watson (1981) showed that the row i, column j element of $\mathcal{I}_{2D,T}$ is given by

$$\frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{2}\operatorname{trace}\left[[\Sigma_t(\theta)]^{-1}\frac{\partial\Sigma_t(\theta)}{\partial\theta_i}[\Sigma_t(\theta)]^{-1}\frac{\partial\Sigma_t(\theta)}{\partial\theta_j}\right] + E\left[\frac{\partial[\mu_t(\theta)]'}{\partial\theta_i}[\Sigma_t(\theta)]^{-1}\frac{\partial\mu_t(\theta)}{\partial\theta_i}\right]\right\}. \tag{3.11}$$

One option is to estimate (3.10) by (3.11) with the expectation operator dropped from (3.11). Another common practice is to assume that the limit of $\mathcal{I}_{2D,T}$ as $T \to \infty$ is the same as the plim of

$$\hat{\mathcal{I}} = -\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \log f(y_t|\zeta_{t-1}, x_t; \theta)}{\partial\theta\partial\theta'}\bigg|_{\theta=\hat{\theta}}, \tag{3.12}$$

which can be calculated analytically or numerically by differentiating (3.4). Reported standard errors for $\hat{\theta}$ are then square roots of diagonal elements of $(1/T)(\hat{\mathcal{I}})^{-1}$.

It was noted above that the Kalman filter can be motivated by linear projection arguments even without normal distributions. It is thus of interest to consider as in White (1982) what happens if we use as an estimate of θ the value that maximizes (3.4), even though the true distribution is not normal. Under certain conditions such quasi-maximum likelihood estimates give consistent and asymptotically normal estimates of the true value of θ, with

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{L} N(0, [\mathcal{I}_{2D}\mathcal{I}_{0P}^{-1}\mathcal{I}_{2D}]^{-1}), \tag{3.13}$$

where $\mathcal{I}_{2D}$ is the plim of (3.12) when evaluated at the true value θ_0 and $\mathcal{I}_{0P}$ is the limit of $(1/T)\sum_{t=1}^{T}[s_t(\theta_0)][s_t(\theta_0)]'$ where

$$s_t(\theta_0) \equiv \left[\frac{\partial \log f(y_t|\zeta_{t-1}, x_t; \theta)}{\partial\theta}\bigg|_{\theta=\theta_0}\right].$$

An important hypothesis test for which (3.9) clearly is not valid is testing the constancy of regression coefficients [see Tanaka (1983) and Watson and Engle (1985)]. One can think of the constant-coefficient model as being embedded as a special case of (2.30) and (2.31) in which $E(v_{t+1}v_{t+1}') = 0$ and $\beta_1 = \bar{\beta}$. However, such a specification violates two of the conditions for asymptotic normality mentioned above. First, under the null hypothesis Q falls on the boundary of the allowable parameter space. Second, the parameters of F are unidentified under the null. Watson and Engle (1985) proposed an appropriate test based on the general procedure of Davies (1977). The results in Davies have recently been extended by Hansen (1993). Given the computational demands of these tests, Nicholls and

Pagan (1985, p. 429) recommended Lagrange multiplier tests for heteroskedasticity based on OLS estimation of the constant-parameter model as a useful practical approach. Other approaches are described in Nabeya and Tanaka (1988) and Leybourne and McCabe (1989).

3.4. *Confidence intervals for smoothed estimates and forecasts*

Let $\hat{\xi}_{\tau|T}(\boldsymbol{\theta}_0)$ denote the optimal inference about ξ_τ conditional on observation of all data through date T assuming that $\boldsymbol{\theta}_0$ is known. Thus, for $\tau \leq T$, $\hat{\xi}_{\tau|T}(\boldsymbol{\theta}_0)$ is the smoothed inference (2.22), while for $\tau > T$, $\hat{\xi}_{\tau|T}(\boldsymbol{\theta}_0)$ is the forecast (2.18). If $\boldsymbol{\theta}_0$ were known with certainty, the mean squared error of this inference, denoted $\boldsymbol{P}_{\tau|T}(\boldsymbol{\theta}_0)$, would be given by (2.23) for $\tau \leq T$ and (2.19) for $\tau > T$.

In the case where the true value of $\boldsymbol{\theta}$ is unknown, this optimal inference is approximated by $\hat{\xi}_{\tau|T}(\hat{\boldsymbol{\theta}})$ for $\hat{\boldsymbol{\theta}}$ the maximum likelihood estimate. To describe the consequences of this, it is convenient to adopt the Bayesian perspective that $\boldsymbol{\theta}$ is a random variable. Conditional on having observed all the data ζ_T, the posterior distribution might be approximated by

$$\boldsymbol{\theta}|\zeta_T \sim N(\hat{\boldsymbol{\theta}}, (1/T)(\hat{\mathcal{I}})^{-1}). \tag{3.14}$$

Hamilton (1986) showed that

$$E\{[\xi_\tau - \hat{\xi}_{\tau|T}(\hat{\boldsymbol{\theta}})][\xi_\tau - \hat{\xi}_{\tau|T}(\hat{\boldsymbol{\theta}})]'|\zeta_T\} = E_{\boldsymbol{\theta}|\zeta_T}\{[\xi_\tau - \hat{\xi}_{\tau|T}(\boldsymbol{\theta})][\xi_\tau - \hat{\xi}_{\tau|T}(\boldsymbol{\theta})]'|\zeta_T\}$$
$$+ E_{\boldsymbol{\theta}|\zeta_T}\{[\hat{\xi}_{\tau|T}(\boldsymbol{\theta}) - \hat{\xi}_{\tau|T}(\hat{\boldsymbol{\theta}})][\hat{\xi}_{\tau|T}(\boldsymbol{\theta}) - \hat{\xi}_{\tau|T}(\hat{\boldsymbol{\theta}})]'|\zeta_T\}, \tag{3.15}$$

where $E_{\boldsymbol{\theta}|\zeta_T}(\cdot)$ denotes the expectation of $(\cdot)$ with respect to the distribution in (3.14). Thus the mean squared error of an inference based on estimated parameters is the sum of two terms. The first term can be written as $E_{\boldsymbol{\theta}|\zeta_T}\{\boldsymbol{P}_{\tau|T}(\boldsymbol{\theta})\}$, and might be described as "filter uncertainty". A convenient way to calculate this would be to generate, say, 10,000 Monte Carlo draws of $\boldsymbol{\theta}$ from the distribution (3.14), run through the Kalman filter iterations implied by each draw, and calculate the average value of $\boldsymbol{P}_{\tau|T}(\boldsymbol{\theta})$ across draws. The second term, which might be described as "parameter uncertainty", could be estimated from the outer product of $[\hat{\xi}_{\tau|T}(\boldsymbol{\theta}_i) - \hat{\xi}_{\tau|T}(\hat{\boldsymbol{\theta}})]$ with itself for the ith Monte Carlo draw, and again averaging across Monte Carlo realizations.

Similar corrections to (2.21) can be used to generate a mean squared error for the forecast of y_{t+m} in (2.20).

3.5. *Empirical application – an analysis of the real interest rate*

As an illustration of these methods, consider Fama and Gibbons's (1982) real interest rate example discussed in equations (1.21) and (1.22). Let $y_t = i_t - \pi_t$ denote

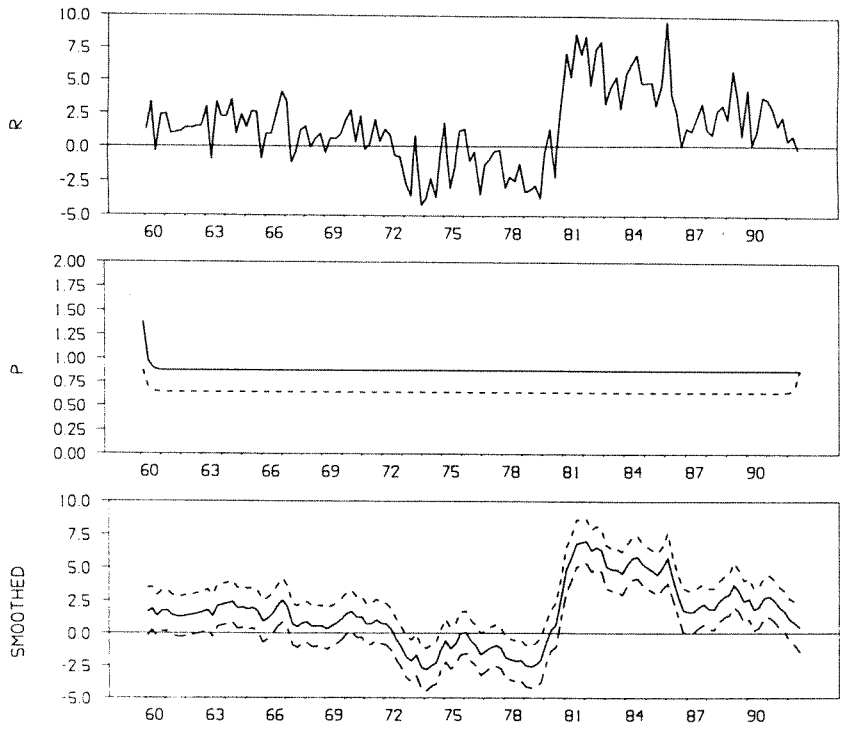

Figure 1. *Top panel.* Ex post real interest rate for the United States, quarterly from 1960:I to 1992:III, quoted at an annual rate. *Middle panel.* Filter uncertainty. Solid line: $P_{t|t}(\hat{\boldsymbol{\theta}})$. Dashed line: $P_{t|T}(\hat{\boldsymbol{\theta}})$. *Bottom panel.* Smoothed inferences $\hat{\xi}_{t|T}(\hat{\boldsymbol{\theta}})$ along with 95 percent confidence intervals.

the observed ex post real interest rate, where i_t is the nominal interest rate on 3-month U.S. Treasury bills for the third month of quarter t (expressed at an annual rate) and π_t is the inflation rate between the third month of quarter t and the third month of $t+1$, measured as 400 times the change in the natural logarithm of the consumer price index. Quarterly data for y_t are plotted in the top panel of Figure 1 for $t =$ 1960:I to 1992:III.

The maximum likelihood estimates for the parameters of this model are as follows, with standard errors in parentheses,

$$\underset{(0.041)}{\xi_t = 0.914\xi_{t-1} + v_t} \qquad \underset{(0.177)}{\hat{\sigma}_v = 0.977}$$

$$\underset{(0.93)}{y_t = 1.43 + \xi_t + w_t} \qquad \underset{(0.14)}{\hat{\sigma}_w = 1.34}\,.$$

Here the state variable $\xi_t = i_t - \pi_t^e - \mu$ has the interpretation as the deviation of the unobserved ex ante real interest rate from its population mean μ.

Even if the population parameter vector $\boldsymbol{\theta} = (\phi, \sigma_v, \mu, \sigma_w)'$ were known with certainty, the econometrician still would not know the value of the ex ante real interest rate, since the market's expected inflation π_t^e is unobserved. However, the econometrician can make an educated guess as to the value of ξ_t based on observations of the ex post real rate through date t, treating the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ as if known with certainty. This guess is the magnitude $\hat{\xi}_{t|t}(\hat{\boldsymbol{\theta}})$, and its mean squared error $P_{t|t}(\hat{\boldsymbol{\theta}})$ is plotted as the solid line in the middle panel of Figure 1. The mean squared error quickly asymptotes to

$$P(\hat{\boldsymbol{\theta}}) = E[\xi_t - E(\xi_t | y_t, y_{t-1}, y_{t-2}, \ldots; \hat{\boldsymbol{\theta}})]^2,$$

which is a fixed constant owing to the stationarity of the process.

The middle panel of Figure 1 also plots the mean squared error for the smoothed inference, $P_{t|T}(\hat{\boldsymbol{\theta}})$. For observations in the middle of the sample this is essentially the mean squared error (MSE) of the doubly-infinite projection

$$S(\hat{\boldsymbol{\theta}}) = E[\xi_t - E(\xi_t | \ldots, y_{t-1}, y_t, y_{t+1}, \ldots; \hat{\boldsymbol{\theta}})]^2.$$

The mean squared error for the smoothed inference is slightly higher for observations near the beginning of the sample (for which the smoothed inference is unable to exploit relevant data on $y_0, y_{-1}, \ldots$) and near the end of the sample for which knowledge of $y_{T+1}, y_{T+2}, \ldots$ would be useful).

The bottom panel of Figure 1 plots the econometrician's best guess as to the value of the ex ante real interest rate based on all of the data observed:

$$i_t - \hat{\pi}_t^e = \hat{\mu} + \hat{\xi}_{t|T}.$$

Ninety-five percent confidence intervals for this inference that take account of both the filter uncertainty $P_{t|T}(\hat{\boldsymbol{\theta}})$ and parameter uncertainty due to the random nature of $\hat{\boldsymbol{\theta}}$ are also plotted. Negative ex ante real interest rates during the 1970's and very high ex ante real interest rates during the early 1980's both appear to be statistically significant. Hamilton (1985) obtained similar results from a more complicated representation for the ex ante real interest rate.

4. Discrete-valued state variables

The time-varying coefficients model was advocated by Sims (1982) as a useful way of dealing with changes occurring all the time in government policy and economic institutions. Often, however, these changes take the form of dramatic, discrete events, such as major wars, financial panics or significant changes in the policy

objectives of the central bank or taxing authority. It is thus of interest to consider time-series models in which the coefficients change only occasionally as a result of such changes in regime.

Consider an unobserved scalar s_t that can take on integer values $1, 2, \ldots, N$ corresponding to N different possible regimes. We can then think of a time-varying coefficient regression model of the form of (2.29),

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta}_{s_t} + w_t \tag{4.1}$$

for $\boldsymbol{x}_t$ a $(k \times 1)$ vector of predetermined or exogenous variables and $w_t \sim$ i.i.d. $N(0, \sigma^2)$. Thus in the regime represented by $s_t = 1$, the regression coefficients are given by $\boldsymbol{\beta}_1$, when $s_t = 2$, the coefficients are $\boldsymbol{\beta}_2$, and so on. The variable s_t summarizes the "state" of the system. The discrete analog to (2.1), the state transition equation for a continuous-valued state variable, is a Markov chain in which the probability distribution of s_{t+1} depends on past events only through the value of s_t. If, as before, observations through date t are summarized by the vector

$$\boldsymbol{\zeta}_t \equiv (y_t, y_{t-1}, \ldots, y_1, \boldsymbol{x}_t', \boldsymbol{x}_{t-1}', \ldots, \boldsymbol{x}_1')',$$

the assumption is that

$$\begin{aligned} \text{Prob}(s_{t+1} = j | s_t = i, s_{t-1} = i_1, s_{t-2} = i_2, \ldots, \boldsymbol{\zeta}_t) &= \text{Prob}(s_{t+1} = j | s_t = i) \\ &\equiv p_{ij}. \end{aligned} \tag{4.2}$$

When this probability does not depend on the previous state ($p_{ij} = p_{lj}$ for all i, j, and l), the system (4.1)–(4.2) is the switching regression model of Quandt (1958); with general transition probabilities it is the Markov-switching regression model developed by Goldfeld and Quandt (1973) and Cosslett and Lee (1985). When $\boldsymbol{x}_t$ includes lagged values of y, (4.1)–(4.2) describes the Markov-switching time-series model of Hamilton (1989).

4.1. *Linear state-space representation of the Markov-switching model*

The parallel between (4.2)–(4.1) and (2.1)–(2.2) is instructive. Let $\boldsymbol{F}$ denote an $(N \times N)$ matrix whose row i, column j element is given by p_{ji}:

$$\boldsymbol{F} = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{N1} \\ p_{12} & p_{22} & \cdots & p_{N2} \\ \vdots & \vdots & \cdots & \vdots \\ p_{1N} & p_{2N} & \cdots & p_{NN} \end{bmatrix}. \tag{4.3}$$

Let e_i denote the ith column of the $(N \times N)$ identity matrix and construct an $(N \times 1)$ vector ξ_t that is equal to e_i when $s_t = i$. Then the expectation of ξ_{t+1} is an $(N \times 1)$ vector whose ith element is the probability that $s_{t+1} = i$. In particular, the expectation of ξ_{t+1} conditional on knowing that $s_t = 1$ is the first column of F. More generally,

$$E(\xi_{t+1}|\xi_t, \xi_{t-1}, \ldots, \xi_1, \zeta_t) = F\xi_t. \tag{4.4}$$

The Markov chain (4.2) thus implies the linear state equation

$$\xi_{t+1} = F\xi_t + v_{t+1}, \tag{4.5}$$

where v_{t+1} is uncorrelated with past values of ξ, y or x.

The probability that $s_{t+2} = j$ given $s_t = i$ can be calculated from

$$\begin{aligned}\text{Prob}(s_{t+2} = j|s_t = i) &= p_{i1}p_{1j} + p_{i2}p_{2j} + \cdots + p_{iN}p_{Nj} \\ &= p_{1j}p_{i1} + p_{2j}p_{i2} + \cdots + p_{Nj}p_{iN},\end{aligned}$$

which will be recognized as the row j, column i element of F^2. In general, the probability that $s_{t+m} = j$ given $s_t = i$ is given by the row j, column i element of F^m, and

$$E(\xi_{t+m}|\xi_t, \xi_{t-1}, \ldots, \xi_1, \zeta_t) = F^m\xi_t. \tag{4.6}$$

Moreover, the regression equation (4.1) can be written

$$y_t = x_t'B\xi_t + w_t, \tag{4.7}$$

where B is a $(k \times N)$ matrix whose ith column is given by β_i. Equation (4.7) will be recognized as an observation equation of the form of (2.26) with $[H(x_t)]' = x_t'B$.

Thus the model (4.1)–(4.2) can be represented by the linear state-space model (2.1) and (2.26). However, the disturbance in the state equation v_{t+1} can only take on a set of N^2 possible discrete values, and is thus no longer normal, so that the Kalman filter applied to this system does not generate optimal forecasts or evaluation of the likelihood function.

4.2. *Optimal filter when the state variable follows a Markov chain*

The Kalman filter was described above as an iterative algorithm for calculating the distribution of the state vector ξ_t conditional on ζ_{t-1}. When ξ_t is a continuous normal variable, this distribution is summarized by its mean and variance. When the state variable is the discrete scalar s_t, its conditional distribution is, instead, summarized by

$$\text{Prob}(s_t = i|\zeta_{t-1}) \quad \text{for } i = 1, 2, \ldots, N. \tag{4.8}$$

Expression (4.8) describes a set of N numbers which sum to unity. Hamilton (1989) presented an algorithm for calculating these numbers, which might be viewed as a discrete version of the Kalman filter. This is an iterative algorithm whose input at step t is the set of N numbers $\{\text{Prob}(s_t = i|\zeta_{t-1})\}_{i=1}^N$ and whose output is $\{\text{Prob}(s_{t+1} = i|\zeta_t)\}_{i=1}^N$. In motivating the Kalman filter, we initially assumed that the values of F, Q, A, H and R were known with certainty, but then showed how the filter could be used to evaluate the likelihood function and estimate these parameters. Similarly, in describing the discrete analog, we will initially assume that the values of $\beta_1, \beta_2, \ldots, \beta_N, \sigma$, and $\{p_{ij}\}_{i,j=1}^N$ are known with certainty, but will then see how the filter facilitates maximum likelihood estimation of these parameters. A key difference is that, whereas the Kalman filter produces forecasts that are linear in the data, the discrete-state algorithm, described below, is nonlinear.

If the Markov chain is stationary and ergodic, the iteration to evaluate (4.8) can be started at date $t = 1$ with the unconditional probabilities. Let π_i denote the unconditional probability that $s_t = i$ and collect these in an $(N \times 1)$ vector $\pi \equiv (\pi_1, \pi_2, \ldots, \pi_N)'$. Noticing that π can be interpreted as $E(\xi_t)$, this vector can be found by taking expectations of (4.5):

$$\pi = F\pi. \tag{4.9}$$

Although this represents a system of N equations in N unknowns, it cannot be solved for π; the matrix $(I_N - F)$ is singular, since each of its columns sums to zero. However, if the chain is stationary and ergodic, the system of $(N + 1)$ equations represented by (4.9) along with the equation

$$\mathbf{1}'\pi = 1 \tag{4.10}$$

can be solved uniquely for the ergodic probabilities (here "$\mathbf{1}$" denotes an $(N \times 1)$ vector, all of whose elements are unity). For $N = 2$, the solution is

$$\pi_1 = (1 - p_{22})/(1 - p_{11} + 1 - p_{22}), \tag{4.11}$$

$$\pi_2 = (1 - p_{11})/(1 - p_{11} + 1 - p_{22}). \tag{4.12}$$

A general solution for π can be calculated from the $(N + 1)$th column of the matrix $(A'A)^{-1}A'$ where

$$\underset{(N+1)\times N}{A} = \begin{bmatrix} I_N - F \\ \mathbf{1}' \end{bmatrix}.$$

The input for step t of the algorithm is $\{\text{Prob}(s_t = i|\zeta_{t-1})\}_{i=1}^N$, whose ith entry under the assumption of predetermined or exogenous x_t is the same as

$$\text{Prob}(s_t = i|x_t, \zeta_{t-1}). \tag{4.13}$$

The assumption in (4.1) was that

$$f(y_t|s_t = i, x_t, \zeta_{t-1}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[\frac{-(y_t - x_t'\beta_i)^2}{2\sigma^2}\right]. \tag{4.14}$$

For given i, x_t, y_t, β_i, and σ, the right-hand side of (4.14) is a number that can be calculated. This number can be multiplied by (4.13) to produce the likelihood of jointly observing $s_t = i$ and y_t:

$$f(y_t, s_t = i|x_t, \zeta_{t-1}) = f(y_t|s_t = i, x_t, \zeta_{t-1}) \cdot \text{Prob}(s_t = i|x_t, \zeta_{t-1}). \tag{4.15}$$

Expression (4.15) describes a set of N numbers (for $i = 1, 2, \ldots, N$) whose sum is the density of y_t conditional on x_t and ζ_{t-1}:

$$f(y_t|x_t, \zeta_{t-1}) = \sum_{i=1}^{N} f(y_t, s_t = i|x_t, \zeta_{t-1}). \tag{4.16}$$

If each of the N numbers in (4.15) is divided by the magnitude in (4.16), the result is the optimal inference about s_t based on observation of $\zeta_t \equiv \{y_t, x_t, \zeta_{t-1}\}$:

$$\text{Prob}(s_t = i|\zeta_t) = \frac{f(y_t, s_t = i|x_t, \zeta_{t-1})}{f(y_t|x_t, \zeta_{t-1})}. \tag{4.17}$$

The output for the jth iteration can then be calculated from

$$\begin{aligned}\text{Prob}(s_{t+1} = j|\zeta_t) &= \sum_{i=1}^{N} \text{Prob}(s_{t+1} = j, s_t = i|\zeta_t) \\ &= \sum_{i=1}^{N} \text{Prob}(s_{t+1} = j|s_t = i, \zeta_t) \cdot \text{Prob}(s_t = i|\zeta_t) \\ &= \sum_{i=1}^{N} p_{ij} \cdot \text{Prob}(s_t = i|\zeta_t).\end{aligned} \tag{4.18}$$

To summarize, let $\hat{\xi}_{t|t-1}$ denote an $(N \times 1)$ vector whose ith element represents $\text{Prob}(s_t = i|\zeta_{t-1})$ and let η_t denote an $(N \times 1)$ vector whose ith element is given by (4.14). Then the sequence $\{\hat{\xi}_{t|t-1}\}_{t=1}^T$ can be found by iterating on

$$\hat{\xi}_{t+1|t} = \frac{F \cdot (\hat{\xi}_{t|t-1} \odot \eta_t)}{\mathbf{1}'(\hat{\xi}_{t|t-1} \odot \eta_t)}, \tag{4.19}$$

where "$\odot$" denotes element-by-element multiplication and $\mathbf{1}$ represents an $(N \times 1)$ vector of ones. The iteration is started with $\hat{\xi}_{1|0} = \pi$ where π is given by the solution to (4.9) and (4.10). The contemporaneous inference $\hat{\xi}_{t|t}$ is given by $(\hat{\xi}_{t|t-1} \odot \eta_t)/[\mathbf{1}'(\hat{\xi}_{t|t-1} \odot \eta_t)]$.

4.3. *Extensions*

The assumption that y_t depends only on the current value s_t of a first-order Markov chain is not really restrictive. For example, the model estimated in Hamilton (1989) was

$$y_t - \mu_{s_t^*} = \phi_1(y_{t-1} - \mu_{s_{t-1}^*}) + \phi_2(y_{t-2} - \mu_{s_{t-2}^*}) + \cdots + \phi_p(y_{t-p} - \mu_{s_{t-p}^*}) + \varepsilon_t, \tag{4.20}$$

where s_t^* can take on the values 1 or 0, and follows a Markov chain with $\text{Prob}(s_{t+1}^* = j|s_t^* = i) = p_{ij}^*$. This can be written in the form of (4.1)–(4.2) by letting $N = 2^{p+1}$ and defining

$$\begin{aligned}
&s_t = 1 \quad \text{if} \quad (s_t^* = 1, s_{t-1}^* = 1, \ldots, \text{ and } s_{t-p}^* = 1), \\
&s_t = 2 \quad \text{if} \quad (s_t^* = 0, s_{t-1}^* = 1, \ldots, \text{ and } s_{t-p}^* = 1), \\
&\vdots \\
&s_t = N-1 \quad \text{if} \quad (s_t^* = 1, s_{t-1}^* = 0, \ldots, \text{ and } s_{t-p}^* = 0), \\
&s_t = N \quad \text{if} \quad (s_t^* = 0, s_{t-1}^* = 0, \ldots, \text{ and } s_{t-p}^* = 0).
\end{aligned} \tag{4.21}$$

For illustration, the matrix of transition probabilities when $p = 2$ is

$$\underset{(8\times 8)}{F} = \begin{bmatrix}
p_{11}^* & 0 & 0 & 0 & p_{11}^* & 0 & 0 & 0 \\
p_{10}^* & 0 & 0 & 0 & p_{10}^* & 0 & 0 & 0 \\
0 & p_{01}^* & 0 & 0 & 0 & p_{01}^* & 0 & 0 \\
0 & p_{00}^* & 0 & 0 & 0 & p_{00}^* & 0 & 0 \\
0 & 0 & p_{11}^* & 0 & 0 & 0 & p_{11}^* & 0 \\
0 & 0 & p_{10}^* & 0 & 0 & 0 & p_{10}^* & 0 \\
0 & 0 & 0 & p_{01}^* & 0 & 0 & 0 & p_{01}^* \\
0 & 0 & 0 & p_{00}^* & 0 & 0 & 0 & p_{00}^*
\end{bmatrix}. \tag{4.22}$$

There is also no difficulty in generalizing the above method to $(n \times 1)$ vector processes $\boldsymbol{y}_t$ with changing coefficients or variances. Suppose that when the process is in state s_t,

$$\boldsymbol{y}_t | s_t, \boldsymbol{x}_t \sim N(\boldsymbol{\Pi}'_{s_t} \boldsymbol{x}_t, \boldsymbol{\Omega}_{s_t}) \tag{4.23}$$

where $\boldsymbol{\Pi}'_1$, for example, is an $(n \times k)$ matrix of regression coefficients appropriate when $s_t = 1$. Then we simply replace (4.14) with

$$f(\boldsymbol{y}_t | s_t = i, \boldsymbol{x}_t, \zeta_{t-1}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Omega}_i|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{y}_t - \boldsymbol{\Pi}'_i \boldsymbol{x}_t)' \boldsymbol{\Omega}_i^{-1} (\boldsymbol{y}_t - \boldsymbol{\Pi}'_i \boldsymbol{x}_t) \right], \tag{4.24}$$

with other details of the recursion identical.

It is more difficult to incorporate changes in regime in a moving average process such as $y_t = \varepsilon_t + \theta_{s_t^*} \varepsilon_{t-1}$. For such a process the distribution of y_t depends on the complete history $(y_{t-1}, y_{t-2}, \ldots, y_1, s_t^*, s_{t-1}^*, \ldots, s_1^*)$, and N, in a representation such as (4.21), grows with the sample size T. Lam (1990) successfully estimated a related model by truncating the calculations for negligible probabilities. Approximations to the optimal filter for a linear state-space model with changing coefficient matrices have been proposed by Gordon and Smith (1990), Shumway and Stoffer (1991) and Kim (1994).

4.4. *Forecasting*

Applying the law of iterated expectations to (4.6), the optimal forecast of ξ_{t+m} based on data observed through date t is

$$E(\xi_{t+m} | \zeta_t) = \boldsymbol{F}^m \hat{\xi}_{t|t}, \tag{4.25}$$

where $\hat{\xi}_{t|t}$ is the optimal inference calculated by the filter.

As an example of using (4.25) to forecast y_t, consider again the example in (4.20). This can be written as

$$y_t = \mu_{s_t^*} + z_t, \tag{4.26}$$

where $z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + \varepsilon_t$. If $\{s_t^*\}$ were observed, an m-period-ahead forecast of the first term in (4.26) turns out to be

$$E(\mu_{s_{t+m}^*} | s_t^*) = \mu_0 + \{\pi_1 + \lambda^m (s_t^* - \pi_1)\}(\mu_1 - \mu_0), \tag{4.27}$$

where $\lambda \equiv (-1 + p_{11}^* + p_{00}^*)$ and $\pi_1 = (1 - p_{00}^*)/(1 - p_{11}^* + 1 - p_{00}^*)$. If p_{11}^* and p_{00}^* are both greater than $\frac{1}{2}$, then $0 < \lambda < 1$ and there is a smooth decay toward the steady-state probabilities. Similarly, the optimal forecast of z_{t+m} based on its own lagged values can be deduced from (1.9):

$$E(z_{t+m} | z_t, z_{t-1}, \ldots, z_{t-p+1}) = \boldsymbol{e}'_1 \boldsymbol{\Phi}^m [z_t \quad z_{t-1} \quad \cdots \quad z_{t-p+1}]' \tag{4.28}$$

where $\boldsymbol{e}'_1$ denotes the first row of the $(p \times p)$ identity matrix and $\boldsymbol{\Phi}$ denotes the $(p \times p)$ matrix on the right-hand side of (1.12). Recalling that $z_t = y_t - \mu_{s_t^*}$ is known if y_t and s_t^* are known, we can substitute (4.27) and (4.28) into (4.26) to conclude

$$\begin{aligned} E(y_{t+m} | s_t, \zeta_t) = \mu_0 &+ \{\pi_1 + \lambda^m (s_t^* - \pi_1)\}(\mu_1 - \mu_0) \\ &+ \boldsymbol{e}'_1 \boldsymbol{\Phi}^m [(y_t - \mu_{s_t^*}) \quad (y_{t-1} - \mu_{s_{t-1}^*}) \quad \cdots \quad (y_{t-p+1} - \mu_{s_{t-p+1}^*})]'. \end{aligned} \tag{4.29}$$

Since (4.29) is linear in $\{s_t^*\}$, the forecast based solely on the observed variables ζ_t can be found by applying the law of iterated expectations to (4.29):

$$E(y_{t+m} | \zeta_t) = \mu_0 + \{\pi_1 + \lambda^m [\text{Prob}(s_t^* = 1 | \zeta_t) - \pi_1]\}(\mu_1 - \mu_0) + \boldsymbol{e}'_1 \boldsymbol{\Phi}^m \tilde{\boldsymbol{y}}_t, \tag{4.30}$$

where the ith element of the $(p \times 1)$ vector $\tilde{\boldsymbol{y}}_t$ is given by

$$\tilde{y}_{it} = y_{t-i+1} - \mu_0 \,\text{Prob}(s_{t-i+1}^* = 0 | \zeta_t) - \mu_1 \,\text{Prob}(s_{t-i+1}^* = 1 | \zeta_t).$$

The ease of forecasting makes this class of models very convenient for rational-expectations analysis; for applications see Hamilton (1988), Cecchetti et al. (1990) and Engel and Hamilton (1990).

4.5. *Smoothed probabilities*

We have assumed that the current value of s_t contains all the information in the history of states through date t that is needed to describe the probability laws for y and s:

$$f(\boldsymbol{y}_t | s_t, \boldsymbol{x}_t, \zeta_{t-1}) = f(\boldsymbol{y}_t | s_t, s_{t-1}, \ldots, s_1, \boldsymbol{x}_t, \zeta_{t-1}),$$

$$\text{Prob}(s_{t+1} = j | s_t = i) = \text{Prob}(s_{t+1} = j | s_t = i, s_{t-1} = i_{t-1}, \ldots, s_1 = i_1).$$

Under these assumptions we have, as in Kitagawa (1987, p. 1033) and Kim (1994),

that

$$\begin{aligned}\text{Prob}(s_t = j, s_{t+1} = i|\zeta_T) &= \text{Prob}(s_{t+1} = i|\zeta_T)\text{Prob}(s_t = j|s_{t+1} = i, \zeta_T)\\ &= \text{Prob}(s_{t+1} = i|\zeta_T)\text{Prob}(s_t = j|s_{t+1} = i, \zeta_t)\\ &= \text{Prob}(s_{t+1} = i|\zeta_T)\frac{\text{Prob}(s_t = j, s_{t+1} = i|\zeta_t)}{\text{Prob}(s_{t+1} = i|\zeta_t)}\\ &= \text{Prob}(s_{t+1} = i|\zeta_T)\frac{\text{Prob}(s_t = j|\zeta_t)\text{Prob}(s_{t+1} = i|s_t = j)}{\text{Prob}(s_{t+1} = i|\zeta_t)}.\end{aligned} \tag{4.31}$$

Sum (4.31) over $i = 1, \ldots, N$ and collect the resulting equations for $j = 1, \ldots, N$ in a vector $\hat{\xi}_{t|T}$, whose jth element is $\text{Prob}(s_t = j|\zeta_T)$:

$$\hat{\xi}_{t|T} = \hat{\xi}_{t|t} \odot \{\boldsymbol{F}'(\hat{\xi}_{t+1|T}(\div)\hat{\xi}_{t+1|t})\}, \tag{4.32}$$

where "$(\div)$" denotes element-by-element division. The smoothed probabilities are thus found by iterating on (4.32) backwards for $t = T-1, T-2, \ldots, 1$.

4.6. *Maximum likelihood estimation*

For given numerical values of the transition probabilities in $\boldsymbol{F}$ and the regression parameters such as $(\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_N, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_N)$ in (4.24), the value of the log likelihood function of the observed data is $\sum_{t=1}^{T} \log f(\boldsymbol{y}_t|\boldsymbol{x}_t, \zeta_{t-1})$ for $f(\boldsymbol{y}_t|\boldsymbol{x}_t, \zeta_{t-1})$ given by (4.16). This can be maximized numerically. Again, the EM elgorithm is often an efficient approach [see Baum et al. (1970), Kiefer (1980) and Hamilton (1990)]. For the model given in (4.24), the EM algorithm is implemented by making an arbitrary initial guess at the parameters and calculating the smoothed probabilities. OLS regression of $\boldsymbol{y}_t\sqrt{\text{Prob}(s_t = 1|\zeta_T)}$ on $\boldsymbol{x}_t\sqrt{\text{Prob}(s_t = 1|\zeta_T)}$ gives a new estimate of $\boldsymbol{\Pi}_1$ and a new estimate of $\boldsymbol{\Omega}_1$ is provided by the sample variance matrix of these OLS residuals. Smoothed probabilities for state 2 are used to estimate $\boldsymbol{\Pi}_2$ and $\boldsymbol{\Omega}_2$, and so on. New estimates for p_{ij} are inferred from

$$p_{ij} = \frac{\left[\sum_{t=2}^{T} \text{Prob}(s_t = j, s_{t-1} = i|\zeta_T)\right]}{\left[\sum_{t=2}^{T} \text{Prob}(s_{t-1} = i|\zeta_T)\right]},$$

with the probability of the initial state calculated from $\pi_i = \text{Prob}(s_1 = i|\zeta_T)$ rather than (4.9)–(4.10). These new parameter values are then used to recalculate the smoothed probabilities, and the procedure continues until convergence.

When the variance depends on the state as in (4.24), there is an essential singularity in the likelihood function at $\boldsymbol{\Omega}_1 = \boldsymbol{0}$. This can be safely ignored without consequences; for further discussion, see Hamilton (1991).

4.7. *Asymptotic properties of maximum likelihood estimates*

It is typically assumed that the usual asymptotic distribution theory motivating (3.9) holds for this class of models, though we are aware of no formal demonstration of this apart from Kiefer's (1978) analysis of i.i.d. switching regressions. Hamilton (1993) examined specification tests derived under the assumption that (3.9) holds.

Two cases in which (3.9) is clearly invalid should be mentioned. First, the maximum likelihood estimate $\hat{p}_{ij}$ may well be at a boundary of the allowable parameter space (zero or one), in which case the information matrix in (3.12) need not even be positive definite. One approach in this case is to regard the value of p_{ij} as fixed at zero or one and calculate the information matrix with respect to other parameters.

Another case in which standard asymptotic distribution theory cannot be invoked is to test for the number of states. The parameter p_{12} is unidentified under the null hypothesis that the distribution under state one is the same as under state two. A solution to this problem was provided by Hansen (1992). Testing the specification with fewer states for evidence of omitted heteroskedasticity affords a simple alternative.

4.8. *Empirical application – another look at the real interest rate*

We illustrate these methods with a simplified version of Garcia and Perron's (1993) analysis of the real interest rate. Let y_t denote the ex post real interest rate data described in Section 3.5. Garcia and Perron concluded that a similar data set was well described by $N = 3$ different states. Maximum likelihood estimates for our data are as follows, with standard errors in parentheses:[5]

$$\begin{aligned} y_t|s_t = 1 &\sim N(\underset{(0.41)}{5.69}, \underset{(1.11)}{3.72}),\\ y_t|s_t = 2 &\sim N(\underset{(0.16)}{1.58}, \underset{(0.32)}{1.93}),\\ y_t|s_t = 3 &\sim N(\underset{(0.30)}{-1.58}, \underset{(0.72)}{2.83}),\end{aligned}$$

[5] Garcia and Perron also included $p = 2$ autoregressive terms as in (4.20), which were omitted from the analysis described here.

$$\hat{F} = \begin{bmatrix} 0.950 & 0 & 0.036 \\ (0.044) & & (0.030) \\ 0.050 & 0.990 & 0 \\ (0.044) & (0.010) & \\ 0 & 0.010 & 0.964 \\ & (0.010) & (0.030) \end{bmatrix}.$$

The unrestricted maximum likelihood estimates for the transition probabilities occur at the boundaries with $\hat{p}_{13} = \hat{p}_{21} = \hat{p}_{32} = 0$. These values were then imposed a priori and derivatives were taken with respect to the remaining free parameters $\boldsymbol{\theta} = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, p_{11}, p_{22}, p_{33})'$ to calculate standard errors.

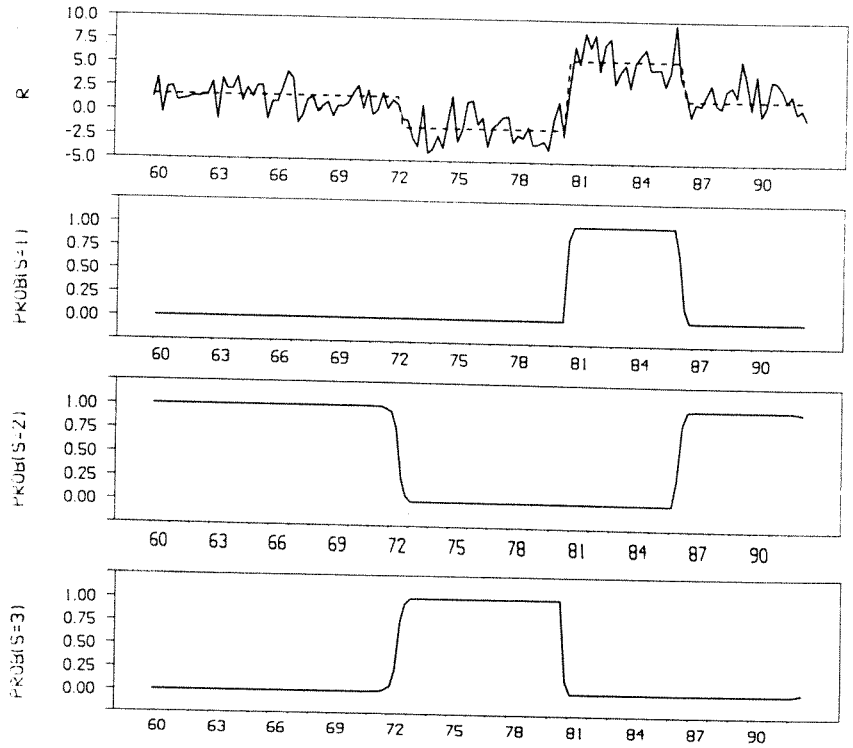

Figure 2. *Top panel.* Solid line: ex post real interest rate. Dashed line: $\hat{\mu}_i \hat{\delta}_{i,t}$, where $\hat{\delta}_{i,t} = 1$ if $\text{Prob}(s_t = i|\zeta_T; \hat{\boldsymbol{\theta}}) > 0.5$ and $\hat{\delta}_{i,t} = 0$ otherwise. *Second panel.* $\text{Prob}(s_t = 1|\zeta_T; \hat{\boldsymbol{\theta}})$. *Third panel.* $\text{Prob}(s_t = 2|\zeta_T; \hat{\boldsymbol{\theta}})$. *Fourth panel.* $\text{Prob}(s_t = 3|\zeta_T; \hat{\boldsymbol{\theta}})$.

Regime 1 is characterized by average real interest rates in excess of 5 percent, while regime 3 is characterized by negative real interest rates. Regime 2 represents the more typical experience of an average real interest rate of 1.58 percent.

The bottom three panels of Figure 2 plot the smoothed probabilities $\text{Prob}(s_t = i|\zeta_T; \hat{\boldsymbol{\theta}})$ for $i = 1, 2$ and 3, respectively. The high interest rate regime lasted from 1980:IV to 1986:II, while the negative real interest rate regime occurred during 1972:3 to 1980:III.

Regime 1 only occurred once during the sample, and yet the asymptotic standard errors reported above suggest that the transition probability $\hat{p}_{11}$ has a standard error of only 0.044. This is because there is in fact not just one observation useful for estimating p_{11}, but, rather, 23 observations. It is exceedingly unlikely that one could have flipped a fair coin once each quarter from 1980:IV through 1986:II and have it come up heads each time; thus the possibility that p_{11} might be as low as 0.5 can easily be dismissed.

The means $\hat{\mu}_1, \hat{\mu}_2$ and $\hat{\mu}_3$ corresponding to the imputed regime for each date are plotted along with the actual data for y_t in the top panel of Figure 2. Garcia and Perron noted that the timing of the high real interest rate episode suggests that fiscal policy may have been more important than monetary policy in producing this unusual episode.

5. Non-normal and nonlinear state-space models

A variety of approximating techniques have been suggested for the case when the disturbances $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ come from a general non-normal distribution or when the state or observation equations are nonlinear. This section reviews two approaches. The first approximates the optimal filter using a finite grid and the second is known as the extended Kalman filter.

5.1. *Kitagawa's grid approximation for nonlinear, non-normal state-space models*

Kitagawa (1987) suggested the following general approach for nonlinear or non-normal filtering. Although the approach in principle can be applied to vector systems, the notation and computations are simplest when the observed variable (y_t) and the state variable (ξ_t) are both scalars. Thus consider

$$\xi_{t+1} = \phi(\xi_t) + v_{t+1}, \tag{5.1}$$

$$y_t = h(\xi_t) + w_t. \tag{5.2}$$

The disturbances v_t and w_t are each i.i.d. and mutually independent and have

densities denoted $q(v_t)$ and $r(w_t)$, respectively. These densities need not be normal, but they are assumed to be of a known form; for example, we may postulate that v_t has a t distribution with ν degrees of freedom:

$$q(v_t) = c(1 + (v_t^2/\nu))^{-(\nu+1)/2},$$

where c is a normalizing constant. Similarly $\phi(\cdot)$ and $h(\cdot)$ represent parametric functions of some known form; for example, $\phi(\cdot)$ might be the logistic function, in which case (5.1) would be

$$\xi_{t+1} = \frac{1}{1 + a\exp(-b\xi_t)} + v_{t+1}. \tag{5.3}$$

Step t of the Kalman filter accepted as input the distribution of ξ_t conditional on $\zeta_{t-1} \equiv (y_{t-1}, y_{t-2}, \ldots, y_1)'$ and produced as output the distribution of ξ_{t+1} conditional on ζ_t. Under the normality assumption the input distribution was completely summarized by the mean $\hat{\xi}_{t|t-1}$ and variance $P_{t|t-1}$. More generally, we can imagine a recursion whose input is the density $f(\xi_t|\zeta_{t-1})$ and whose output is $f(\xi_{t+1}|\zeta_t)$. These, in general, would be continuous functions, though they can be summarized by their values at a finite grid of points, denoted $\xi^{(0)}, \xi^{(1)}, \ldots, \xi^{(N)}$. Thus the input for Kitagawa's filter is the set of $(N+1)$ numbers

$$f(\xi_t|\zeta_{t-1})|_{\xi_t = \xi^{(i)}} \qquad i = 0, 1, \ldots, N \tag{5.4}$$

and the output is (5.4) with t replaced by $t+1$.

To derive the filter, first notice that under the assumed structure, ξ_t summarizes everything about the past that matters for y_t:

$$f(y_t|\xi_t) = f(y_t|\xi_t, \zeta_{t-1}).$$

Thus

$$\begin{aligned} f(y_t, \xi_t|\zeta_{t-1}) &= f(y_t|\xi_t) f(\xi_t|\zeta_{t-1}) \\ &= r[y_t - h(\xi_t)] f(\xi_t|\zeta_{t-1}) \end{aligned} \tag{5.5}$$

and

$$f(y_t|\zeta_{t-1}) = \int_{-\infty}^{\infty} f(y_t, \xi_t|\zeta_{t-1})\, d\xi_t. \tag{5.6}$$

Given the observed y_t and the known form for $r(\cdot)$ and $h(\cdot)$, the joint density (5.5) can be calculated for each $\xi_t = \xi^{(i)}$, $i = 0, 1, \ldots, N$, and these values can then be

used to approximate (5.6) by

$$f(y_t|\zeta_{t-1}) \simeq \sum_{i=1}^{N} \{f(y_t, \xi_t|\zeta_{t-1})|_{\xi_t = \xi^{(i)}} + f(y_t, \xi_t|\zeta_{t-1})|_{\xi_t = \xi^{(i-1)}}\} \tfrac{1}{2}\{\xi^{(i)} - \xi^{(i-1)}\}. \tag{5.7}$$

The updated density for ξ_t is obtained by dividing each of the $N+1$ numbers in (5.5) by the constant (5.7):

$$\begin{aligned} f(\xi_t|\zeta_t) &= f(\xi_t|y_t, \zeta_{t-1}) \\ &= \frac{f(y_t, \xi_t|\zeta_{t-1})}{f(y_t|\zeta_{t-1})}. \end{aligned} \tag{5.8}$$

The joint conditional density of ξ_{t+1} and ξ_t is then

$$\begin{aligned} f(\xi_{t+1}, \xi_t|\zeta_t) &= f(\xi_{t+1}|\xi_t) f(\xi_t|\zeta_t) \\ &= q[\xi_{t+1} - \phi(\xi_t)] f(\xi_t|\zeta_t). \end{aligned} \tag{5.9}$$

For any pair of values $\xi^{(i)}$ and $\xi^{(j)}$, equation (5.9) can be evaluated at $\xi_t = \xi^{(i)}$ and $\xi_{t+1} = \xi^{(j)}$ from (5.8) and the form of $q(\cdot)$ and $\phi(\cdot)$. The recursion is completed by:

$$\begin{aligned} f(\xi_{t+1}|\zeta_t)|_{\xi_{t+1} = \xi^{(j)}} &= \int_{-\infty}^{\infty} f(\xi_{t+1}, \xi_t|\zeta_t)|_{\xi_{t+1} = \xi^{(j)}}\, d\xi_t \\ &\simeq \sum_{i=1}^{N} \{f(\xi_{t+1}, \xi_t|\zeta_t)|_{\xi_{t+1} = \xi^{(j)}, \xi_t = \xi^{(i)}} \\ &\qquad + f(\xi_{t+1}, \xi_t|\zeta_t)|_{\xi_{t+1} = \xi^{(j)}, \xi_t = \xi^{(i-1)}}\} \tfrac{1}{2}\{\xi^{(i)} - \xi^{(i-1)}\}. \end{aligned} \tag{5.10}$$

An approximation to the log likelihood can be calculated from (5.6):

$$\log f(y_T, y_{T-1}, \ldots, y_1) = \sum_{t=1}^{T} \log f(y_t|\zeta_{t-1}). \tag{5.11}$$

The maximum likelihood estimates of parameters such as a, b and ν are then the values for which (5.11) is greatest.

Feyzioglu and Hassett (1991) provided an economic application of Kitagawa's approach to a nonlinear, non-normal state-space model.

5.2. *Extended Kalman filter*

Consider next a multidimensional normal state-space model

$$\xi_{t+1} = \phi(\xi_t) + v_{t+1}, \tag{5.12}$$

$$y_t = a(x_t) + h(\xi_t) + w_t, \tag{5.13}$$

where ϕ: $\mathbb{R}^r \to \mathbb{R}^r$, a: $\mathbb{R}^k \to \mathbb{R}^n$ and h: $\mathbb{R}^r \to \mathbb{R}^n$, $v_t \sim$ i.i.d. $N(0, Q)$ and $w_t \sim$ i.i.d. $N(0, R)$. Suppose $\phi(\cdot)$ in (5.12) is replaced with a first-order Taylor's approximation around $\xi_t = \hat{\xi}_{t|t}$,

$$\xi_{t+1} = \phi_t + \Phi_t(\xi_t - \hat{\xi}_{t|t}) + v_{t+1}, \tag{5.14}$$

where

$$\underset{(r\times 1)}{\phi_t} = \phi(\hat{\xi}_{t|t}) \qquad \underset{(r\times r)}{\Phi_t} \equiv \left.\frac{\partial \phi(\xi_t)}{\partial \xi_t'}\right|_{\xi_t = \hat{\xi}_{t|t}} \tag{5.15}$$

For example, suppose $r = 1$ and $\phi(\cdot)$ is the logistic function as in (5.3). Then (5.14) would be given by

$$\xi_{t+1} = \frac{1}{1 + a\exp(-b\hat{\xi}_{t|t})} + \frac{ab\exp(-b\hat{\xi}_{t|t})}{[1 + a\exp(-b\hat{\xi}_{t|t})]^2}(\xi_t - \hat{\xi}_{t|t}) + v_{t+1}. \tag{5.16}$$

If the form of ϕ and any parameters it depends on [such as a and b in (5.3)] are known, then the inference $\hat{\xi}_{t|t}$ can be constructed as a function of variables observed at date t (ζ_t) through a recursion to be described in a moment. Thus ϕ_t and Φ_t in (5.14) are directly observed at date t.

Similarly the function $h(\cdot)$ in (5.13) can be linearized around $\hat{\xi}_{t|t-1}$ to produce

$$y_t = a(x_t) + h_t + H_t'(\xi_t - \hat{\xi}_{t|t-1}) + w_t, \tag{5.17}$$

where

$$\underset{(n\times 1)}{h_t} \equiv h(\hat{\xi}_{t|t-1}) \qquad \underset{(n\times r)}{H_t'} \equiv \left.\frac{\partial h(\xi_t)}{\partial \xi_t'}\right|_{\xi_t = \hat{\xi}_{t|t-1}} \tag{5.18}$$

Again h_t and H_t are observed at date $t-1$. The function $a(\cdot)$ in (5.13) need not be linearized since x_t is observed directly.

The idea behind the extended Kalman filter is to treat (5.14) and (5.17) as if they were the true model. These will be recognized as time-varying coefficient versions of a linear state-space model, in which the observed predetermined variable

$\phi_t - \Phi_t\hat{\xi}_{t|t}$ has been added to the state equation. Retracing the logic behind the Kalman filter for this application, the input for step t of the iteration is again the forecast $\hat{\xi}_{t|t-1}$ and mean squared error $P_{t|t-1}$. Given these, the forecast of y_t is found from (5.17):

$$\begin{aligned} E(y_t|x_t, \zeta_{t-1}) &= a(x_t) + h_t \\ &= a(x_t) + h(\hat{\xi}_{t|t-1}). \end{aligned} \tag{5.19}$$

The joint distribution of ξ_t and y_t conditional on x_t and ζ_{t-1} continues to be given by (2.11), with (5.19) replacing the mean of y_t and H_t replacing H. The contemporaneous inference (2.12) goes through with the same minor modification:

$$\hat{\xi}_{t|t} = \hat{\xi}_{t|t-1} + P_{t|t-1}H_t(H_t'P_{t|t-1}H_t + R)^{-1}[y_t - a(x_t) - h(\hat{\xi}_{t|t-1})]. \tag{5.20}$$

If (5.14) described the true model, then the optimal forecast of ξ_{t+1} on the basis of ζ_t would be

$$E(\hat{\xi}_{t+1}|\zeta_t) = \phi_t = \phi(\hat{\xi}_{t|t}).$$

To summarize, step t of the extended Kalman filter uses $\hat{\xi}_{t|t-1}$ and $P_{t|t-1}$ to calculate H_t from (5.18) and $\hat{\xi}_{t|t}$ from (5.20). From these we can evaluate Φ_t in (5.15). The output for step t is then

$$\hat{\xi}_{t+1|t} = \phi(\hat{\xi}_{t|t}), \tag{5.21}$$

$$P_{t+1|t} = \Phi_t P_{t|t-1}\Phi_t' - \{\Phi_t P_{t|t-1}H_t(H_t'P_{t|t-1}H_t + R)^{-1}H_t'P_{t|t-1}\Phi_t'\} + Q. \tag{5.22}$$

The recursion is started with $\hat{\xi}_{1|0}$ and $P_{1|0}$ representing the analyst's prior information about the initial state.

5.3. *Other approaches to nonlinear state-space models*

A number of other approaches to nonlinear state-space models have been explored in the literature. See Anderson and Moore (1979, Chapter 8) and Priestly (1980, 1988) for partial surveys.

References

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

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

Aoki, M. (1987) *State Space Modeling of Time Series*. New York: Springer Verlag.
Baum, L.E., T. Petrie, G. Soules and N. Weiss (1970) "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", *Annals of Mathematical Statistics*, 41, 164–171.
Box, G.E.P. and G.M. Jenkins (1976) *Time Series Analysis: Forecasting and Control*, Second edition. San Francisco: Holden–Day.
Burmeister, E. and K.D. Wall (1982) "Kalman Filtering Estimation of Unobserved Rational Expectations with an Application to the German Hyperinflation", *Journal of Econometrics*, 20, 255–284.
Burmeister, E., K.D. Wall and J.D. Hamilton (1986) "Estimation of Unobserved Expected Monthly Inflation Using Kalman Filtering", *Journal of Business and Economic Statistics*, 4, 147–160.
Caines, P.E. (1988) *Linear Stochastic Systems*. New York: John Wiley and Sons, Inc.
Cecchetti, S.G., P.-S. Lam and N. Mark (1990) "Mean Reversion in Equilibrium Asset Prices", *American Economic Review*, 80, 398–418.
Chow, G.C. (1984) "Random and Changing Coefficient Models", in: Z. Griliches and M.D. Intriligator, eds., *Handbook of Econometrics, Vol. 2*. Amsterdam: North-Holland.
Cosslett, S.R. and L.-F. Lee (1985) "Serial Correlation in Discrete Variable Models", *Journal of Econometrics*, 27, 79–97.
Davies, R.B. (1977) "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative", *Biometrika*, 64, 247–254.
DeGroot, M.H. (1970) *Optimal Statistical Decisions*. New York: McGraw–Hill.
De Jong, P. (1988) "The Likelihood for a State Space Model", *Biometrika*, 75, 165–169.
De Jong, P. (1989) "Smoothing and Interpolation with the State-Space Model", *Journal of the American Statistical Association*, 84, 1085–1088.
De Jong, P. (1991) "The Diffuse Kalman Filter", *Annals of Statistics*, 19, 1073–1083.
Dempster, A.P., N.M. Laird and D.B. Rubin (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
Doan, T., R.B. Litterman and C.A. Sims (1984) "Forecasting and Conditional Projection Using Realistic Prior Distributions", *Econometric Reviews*, 3, 1–100.
Engel, C. and J.D. Hamilton (1990) "Long Swings in the Dollar: Are They in the Data and Do Markets Know It?", *American Economic Review*, 80, 689–713.
Engle, R.F. and M.W. Watson (1981) "A One-Factor Multivariate Time Series Model of Metropolitan Wage Rates", *Journal of the American Statistical Association*, 76, 774–781.
Engle, R.F. and M.W. Watson (1987) "The Kalman Filter: Applications to Forecasting and Rational-Expectations Models", in: T.F. Bewley, ed., *Advances in Econometrics*. Fifth World Congress, Volume I. Cambridge, England: Cambridge University Press.
Fama, E.F. and M.R. Gibbons (1982) "Inflation, Real Returns, and Capital Investment", *Journal of Monetary Economics*, 9, 297–323.
Feyzioglu, T. and K. Hassett (1991), A Nonlinear Filtering Technique for Estimating the Timing and Importance of Liquidity Constraints, Mimeographed, Georgetown University.
Garcia, R. and P. Perron (1993), An Analysis of the Real Interest Rate Under Regime Shifts, Mimeographed, University of Montreal.
Gevers, M. and V. Wertz (1984) "Uniquely Identifiable State-Space and ARMA Parameterizations for Multivariable Linear Systems", *Automatica*, 20, 333–347.
Ghosh, D. (1989) "Maximum Likelihood Estimation of the Dynamic Shock-Error Model", *Journal of Econometrics*, 41, 121–143.
Goldfeld, S.M. and R.M. Quandt (1973) "A Markov Model for Switching Regressions", *Journal of Econometrics*, 1, 3–16.
Gordon, K. and A.F.M. Smith (1990) "Modeling and Monitoring Biomedical Time Series", *Journal of the American Statistical Association*, 85, 328–337.
Hamilton, J.D. (1985) "Uncovering Financial Market Expectations of Inflation", *Journal of Political Economy*, 93, 1224–1241.
Hamilton, J.D. (1986) "A Standard Error for the Estimated State Vector of a State-Space Model", *Journal of Econometrics*, 33, 387–397.
Hamilton, J.D. (1988) "Rational-Expectations Econometric Analysis of Changes in Regime: An Investigation of the Term Structure of Interest Rates", *Journal of Economic Dynamics and Control*, 12, 385–423.

Hamilton, J.D. (1989) "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle", *Econometrica*, 57, 357–384.
Hamilton, J.D. (1990) "Analysis of Time Series Subject to Changes in Regime", *Journal of Econometrics*, 45, 39–70.
Hamilton, J.D. (1991) "A Quasi-Bayesian Approach to Estimating Parameters for Mixtures of Normal Distributions", *Journal of Business and Economic Statistics*, 9, 27–39.
Hamilton, J.D. (1993) Specification Testing in Markov-Switching Time Series Models, Mimeographed, University of California, San Diego.
Hamilton, J.D. (1994) *Time Series Analysis*. Princeton, N.J.: Princeton University Press.
Hannan, E.J. (1971) "The Identification Problem for Multiple Equation Systems with Moving Average Errors", *Econometrica*, 39, 751–765.
Hansen, B.E. (1992) "The Likelihood Ratio Test Under Non-Standard Conditions: Testing the Markov Trend Model of GNP", *Journal of Applied Econometrics*, 7, S61–S82.
Hansen, B.E. (1993) Inference When a Nuisance Parameter is Not Identified Under the Null Hypothesis, Mimeographed, University of Rochester.
Harvey, A.C. (1987) "Applications of the Kalman Filter in Econometrics", in: T.F. Bewley, ed., *Advances in Econometrics*, Fifth World Congress, Volume I. Cambridge, England: Cambridge University Press.
Harvey, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, England: Cambridge University Press.
Harvey, A.C. and G.D.A. Phillips (1979) "The Maximum Likelihood Estimation of Regression Models with Autoregressive-Moving Average Disturbances", *Biometrika*, 66, 49–58.
Harvey, A.C. and R.G. Pierse (1984) "Estimating Missing Observations in Economic Time Series", *Journal of the American Statistical Association*, 79, 125–131.
Harvey, A.C. and P.H.J. Todd (1983) "Forecasting Economic Time Series with Structural and Box–Jenkins Models: A Case Study", *Journal of Business and Economic Statistics*, 1, 299–307.
Imrohoroglu, S. (1993) "Testing for Sunspots in the German Hyperinflation", *Journal of Economic Dynamics and Control*, 17, 289–317.
Jones, R.H. (1980) "Maximum Likelihood Fitting of ARMA Models to Time Series with Missing Observations", *Technometrics*, 22, 389–395.
Kalman, R.E. (1960) "A New Approach to Linear Filtering and Prediction Problems", *Journal of Basic Engineering, Transactions of the ASME*, Series D, 82, 35–45.
Kalman, R.E. (1963) "New Methods in Wiener Filtering Theory", in: J.L. Bogdanoff and F. Kozin, eds., *Proceedings of the First Symposium of Engineering Applications of Random Function Theory and Probability*, pp. 270–388. New York: John Wiley & Sons, Inc.
Kiefer, N.M. (1978) "Discrete Parameter Variation: Efficient Estimation of a Switching Regression Model", *Econometrica*, 46, 427–434.
Kiefer, N.M. (1980) "A Note on Switching Regressions and Logistic Discrimination", *Econometrica*, 48, 1065–1069.
Kim, C.-J. (1994) "Dynamic Linear Models with Markov-Switching", *Journal of Econometrics*, 60, 1–22.
Kitagawa, G. (1987) "Non-Gaussian State-Space Modeling of Nonstationary Time Series", *Journal of the American Statistical Association*, 82, 1032–1041.
Kohn, R. and C.F. Ansley (1986) "Estimation, Prediction, and Interpolation for ARIMA Models with Missing Data", *Journal of the American Statistical Association*, 81, 751–761.
Lam, P.-S. (1990) "The Hamilton Model with a General Autoregressive Component: Estimation and Comparison with Other Models of Economic Time Series", *Journal of Monetary Economics*, 26, 409–432.
Leybourne, S.J. and B.P.M. McCabe (1989) "On the Distribution of Some Test Statistics for Coefficient Constancy", *Biometrika*, 76, 169–177.
Magnus, J.R. and H. Neudecker (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: John Wiley & Sons, Inc.
Meinhold, R.J. and N.D. Singpurwalla (1983) "Understanding the Kalman Filter", *American Statistician*, 37, 123–127.
Nabeya, S. and K. Tanaka (1988) "Asymptotic Theory for the Constance of Regression Coefficients Against the Random Walk Alternative", *Annals of Statistics*, 16, 218–235.
Nash, J.C. and M. Walker-Smith (1987) *Nonlinear Parameter Estimation: An Integrated System in Basic*. New York: Marcel Dekker.

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*. Amsterdam: North-Holland.
Pagan, A. (1980) "Some Identification and Estimation Results for Regression Models with Stochastically Varying Coefficients", *Journal of Econometrics*, 13, 341–363.
Priestly, M.B. (1980) "State-Dependent Models: A General Approach to Non-Linear Time Series Analysis", *Journal of Time Series Analysis*, 1, 47–71.
Priestly, M.B. (1988) "Current Developments in Time-Series Modelling", *Journal of Econometrics*, 37, 67–86.
Quandt, R.E. (1958) "The Estimation of Parameters of Linear Regression System Obeying Two Separate Regimes", *Journal of the American Statistical Association*, 55, 873–880.
Quandt, R.E. (1983) "Computational Problems and Methods", in: Z. Griliches and M.D. Intriligator, eds., *Handbook of Econometrics. Volume 1*. Amsterdam: North-Holland.
Raj, B. and A. Ullah (1981) *Econometrics: A Varying Coefficients Approach*. London: Croom-Helm.
Sargent, T.J. (1989) "Two Models of Measurements and the Investment Accelerator", *Journal of Political Economy*, 97, 251–287.
Shumway, R.H. and D.S. Stoffer (1982) "An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm", *Journal of Time Series Analysis*, 3, 253–263.
Shumway, R.H. and D.S. Stoffer (1991) "Dynamic Linear Models with Switching", *Journal of the American Statistical Association*, 86, 763–769.
Sims, C.A. (1982) "Policy Analysis with Econometric Models", *Brookings Papers on Economic Activity*, 1, 107–152.
Stock, J.H. and M.W. Watson (1991) "A Probability Model of the Coincident Economic Indicators", in: K. Lahiri and G.H. Moore, eds., *Leading Economic Indicators: New Approaches and Forecasting Records*, Cambridge, England: Cambridge University Press.
Tanaka, K. (1983) "Non-Normality of the Lagrange Multiplier Statistic for Testing the Constancy of Regression Coefficients", *Econometrica*, 51, 1577–1582.
Tjøstheim, D. (1986) "Some Doubly Stochastic Time Series Models", *Journal of Time Series Analysis*, 7, 51–72.
Wall, K.D. (1980) "Generalized Expectations Modeling in Macroeconometrics", *Journal of Economic Dynamics and Control*, 2, 161–184.
Wall, K.D. (1987) "Identification Theory for Varying Coefficient Regression Models", *Journal of Time Series Analysis*, 8, 359–371.
Watson, M.W. (1989) "Recursive Solution Methods for Dynamic Linear Rational Expectations Models", *Journal of Econometrics*, 41, 65–89.
Watson, M.W. and R.F. Engle (1983) "Alternative Algorithms for the Estimation of Dynamic Factor, MIMIC, and Varying Coefficient Regression Models", *Journal of Econometrics*, 23, 385–400.
Watson, M.W. and R.F. Engle (1985) "Testing for Regression Coefficient Stability with a Stationary AR(1) Alternative", *Review of Economics and Statistics*, 67, 341–346.
White, H. (1982) "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, 50, 1–25.