Part I

State space models

1

Introduction to state space time series analysis

James Durbin

Department of Statistics, London School of Economics and Political Science

Abstract

The paper presents a broad general review of the state space approach to time series analysis. It begins with an introduction to the linear Gaussian state space model. Applications to problems in practical time series analysis are considered. The state space approach is briefly compared with the Box–Jenkins approach. The Kalman filter and smoother and the simulation smoother are described. Missing observations, forecasting and initialisation are considered. A representation of a multivariate series as a univariate series is displayed. The construction and maximisation of the likelihood function are discussed. An application to real data is presented. The treatment is extended to non-Gaussian and nonlinear state space models. A simulation technique based on importance sampling is described for analysing these models. The use of antithetic variables in the simulation is considered. Bayesian analysis of the models is developed based on an extension of the importance sampling technique. Classical and Bayesian methods are applied to a real time series.

State Space and Unobserved Component Models: Theory and Applications, eds. Andrew C. Harvey, Siem Jan Koopman and Neil Shephard. Published by Cambridge University Press.

1.1 Introduction to state space models

1.1.1 Basic ideas

The organisers have asked me to provide a broad, general introduction to state space time series analysis. In the pursuit of this objective I will try to make the exposition understandable for those who have relatively little prior knowledge of the subject, while at the same time including some results of recent research. My starting point is the claim that state space models provide an effective basis for practical time series analysis in a wide range of fields including statistics, econometrics and engineering.

I will base my exposition on the recent book by Durbin and Koopman (2001), referred to from now on as the DK book, which provides a comprehensive treatment of the subject. Readers may wish to refer to the website `http://www.ssfpack.com/dkbook/` for further information about the book. Other books that provide treatments of state space models and techniques include Harvey (1989), West and Harrison (1997), Kitagawa and Gersch (1996) and Kim and Nelson (1999). More general books on time series analysis with substantial treatments of state space methods are, for example, Brockwell and Davis (1987), Hamilton (1994) and Shumway and Stoffer (2000).

I will begin with a particular example that I will use to introduce the basic ideas that underlie state space time series analysis. This refers to logged monthly numbers of car drivers who were killed or seriously injured in road accidents in Great Britain, 1969–84. These data come from a study by Andrew Harvey and me, undertaken on behalf of the British Department of Transport, regarding the effect on road casualties of the seat belt law that was introduced in February 1983; for details see Durbin and Harvey (1985) and Harvey and Durbin (1986).

Inspection of Figure 1.1 reveals that the series is made up of a trend which initially is increasing, then decreases and subsequently flattens out, plus a seasonal effect which is high in the winter and low in the summer, together with a sudden drop in early 1983 seemingly due to the introduction of the seat belt law. Other features that could be present, though they are not apparent from visual inspection, include cycles and regression effects due to the influence of such factors as the price of petrol, weather variations and traffic density. Thus we arrive at the following model:

$$y_t = \mu_t + \gamma_t + c_t + r_t + i_t + \varepsilon_t, \tag{1.1}$$

where

y_t = observation (often logged, possibly a vector)
μ_t = trend (slow change in level)
γ_t = seasonal (pattern can change over time)
c_t = cycle (of longer period than seasonal)
r_t = regression component (coefficients can vary over time)
i_t = intervention effect (e.g. seat belt law)
ε_t = random error or disturbance or noise.

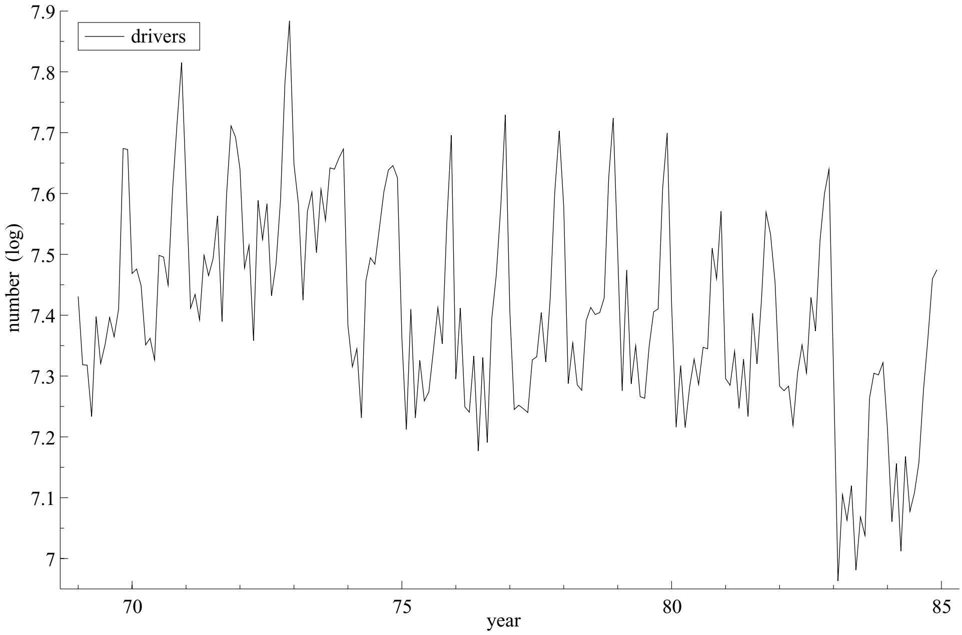

Fig. 1.1. Monthly numbers (logged) of car drivers who were killed or seriously injured in road accidents in Great Britain.

In the state space approach we construct submodels designed to model the behaviour of each component such as trend, seasonal, etc. separately and we put these submodels together to form a single matrix model called a *state space model*. The model used by Harvey and Durbin for the analysis of the data of Figure 1.1 included all the components of (1.1) except c_t; some of the results of their analysis will be presented later.

1.1.2 Special cases of the basic model

We consider the following two special cases.

1.1.2.1 The local level model

This is specified by

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim N(0, \sigma_\varepsilon^2), \\ \mu_{t+1} &= \mu_t + \eta_t, & \eta_t &\sim N(0, \sigma_\eta^2), \end{aligned} \tag{1.2}$$

for $t = 1, \ldots, n$, where the ε_ts and η_ts are all mutually independent and are also independent of μ_1.

The objective of this model is to represent a series with no trend or seasonal whose level μ_t is allowed to vary over time. The second equation of the model is a *random walk*; random walks are basic elements in many state space time series models. Although it is simple, the local level model is not an artificial model and it provides the basis for the treatment of important series in practice. It is employed to explain the ideas underlying state space time series analysis in an elementary way in Chapter 2 of the DK book. The properties of time series that are generated by a local level model are studied in detail in Harvey (1989).

1.1.2.2 The local linear trend model

This is specified by

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim N(0, \sigma_\varepsilon^2), \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim N(0, \sigma_\xi^2), \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\sim N(0, \sigma_\zeta^2). \end{aligned} \tag{1.3}$$

This extends the local level model to the case where there is a trend with a slope ν_t where both level and slope are allowed to vary over time. It is worth noting that when both ξ_t and ζ_t are zero, the model reduces to the classical linear trend plus noise model, $y_t = \alpha + \beta t +$ error. It is sometimes useful to smooth the trend by putting $\xi_t = 0$ in (1.3). Details of the model and its extensions to the general class of structural time series models are given in the DK book Section 3.2 and in Harvey (1989).

The matrix form of the local linear trend model is

$$\begin{aligned} y_t &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \varepsilon_t, \\ \begin{pmatrix} \mu_{t+1} \\ \nu_{t+1} \end{pmatrix} &= \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \begin{pmatrix} \xi_t \\ \zeta_t \end{pmatrix}. \end{aligned}$$

By considering this and other special cases in matrix form we are led to the following general model which provides the basis for much of our further treatment of state space models.

1.1.3 The linear Gaussian state space model

This has the form

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, & \varepsilon_t &\sim N(0, H_t), & & \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t, & \eta_t &\sim N(0, Q_t), & t &= 1, \dots, n, \\ \alpha_1 &\sim N(a_1, P_1). & & & & \end{aligned} \tag{1.4}$$

Matrices Z_t, H_t, T_t, R_t and Q_t are assumed known. Initially, a_1 and P_1 are assumed known; we will consider later what to do when some elements of them are unknown. The $p \times 1$ vector y_t is the observation. The unobserved $m \times 1$ vector α_t is called the *state*. The disturbances ε_t and η_t are independent sequences of independent normal vectors. The matrix R_t, when it is not the identity, is usually a selection matrix, that is, a matrix whose columns are a subset of the columns of the identity matrix; it is needed when the dimensionality of α_t is greater than that of the disturbance vector η_t. The first equation is called the *observation equation* and the second equation is called the *state equation*.

The structure of model (1.4) is a natural one for representing the behaviour of many time series as a first approximation. The first equation is a standard multivariate linear regression model whose coefficient vector α_t varies over time; the development over time of α_t is determined by the first-order vector autoregression given in the second equation. The Markovian nature of the model accounts for many of its elegant properties.

In spite of the conceptual simplicity of this model it is highly flexible and has a remarkably wide range of applications to problems in practical time series analysis. I will mention just a few.

(i) Structural time series models. These are models of the basic form (1.1) where the submodels for the components are chosen to be compatible with the state space form (1.4). The local level model and the local linear trend model are simple special cases. The models are sometimes called *dynamic linear models* .

(ii) ARMA and Box–Jenkins (BJ) ARIMA models. These can be put in state space form as described in the DK book, Section 3.3. This means that ARIMA models can be treated as special cases of state space models. I will make a few remarks at this point on the relative merits of the BJ approach and the state space approach for practical time series analysis.

  (a) BJ is a ‘black box’ approach in which the model is determined purely by the data without regard to the structure underlying the data, whereas state space fits the data to the structure of the system which generated the data.

(b) BJ eliminates trend and seasonal by differencing. However, in many cases these components have intrinsic interest and in state space they can be estimated directly. While in BJ estimates can be 'recovered' from the differenced series by maximising the residual mean square, this seems an artificial procedure.

(c) The BJ identification procedure need not lead to a unique model; in some cases several apparently quite different models can appear to fit the data equally well.

(d) In BJ it is difficult to handle regression effects, missing observations, calendar effects, multivariate observations and changes in coefficients over time; these are all straightforward in state space.

A fuller discussion of the relative merits of state space and BJ is given in the DK book, Section 3.5. The comparison is strongly in favour of state space.

(iii) Model (1.4) handles time-varying regression and regression with autocorrelated errors straightforwardly.

(iv) State space models can deal with problems in spline smoothing in discrete and continuos time on a proper modelling basis in which parameters can be estimated by standard methods, as compared with customary *ad hoc* methods.

1.2 Basic theory for state space analysis

1.2.1 Introduction

In this section we consider the main elements of the methodology required for time series analysis based on the linear Gaussian model (1.4). Let $Y_t = \{y_1, \ldots, y_t\}$, $t = 1, \ldots, n$. We will focus on the following items:

- *Kalman filter.* This recursively computes $a_{t+1} = E(\alpha_{t+1}|Y_t)$ and $P_{t+1} = Var(\alpha_{t+1}|Y_t)$ for $t = 1, \ldots, n$. Since distributions are normal, these quantities specify the distribution of α_{t+1} given data up to time t.
- *State smoother.* This estimates $\hat{\alpha}_t = E(\alpha_t \,|\, Y_n)$ and $V_t = \mathrm{Var}(\alpha_t \,|\, Y_n)$ and hence the conditional distribution of α_t given all the observations for $t = 1, \ldots, n$.
- *Simulation smoother.* An algorithm for generating draws from

$$p(\alpha_1, \ldots, \alpha_n|Y_n).$$

This is an essential element in the analysis of non-Gaussian and nonlinear models as described in Section 1.4.

- *Missing observations.* We show that the treatment of missing observations is particularly simple in the state space approach.
- *Forecasting* is simply treated as a special case of missing observations.
- *Initialisation.* This deals with the case where some elements of $a_1 = E(\alpha_1)$ and $V_1 = \mathrm{Var}(\alpha_1)$ are unknown.
- *Univariate treatment of multivariate series.* This puts a multivariate model into univariate form, which can simplify substantially the treatment of large complex models.
- *Parameter estimation.* We show that the likelihood function is easily constructed using the Kalman filter.

1.2.2 Kalman filter

We calculate

$$a_{t+1} = E(\alpha_{t+1}|Y_t), \qquad P_{t+1} = Var(\alpha_{t+1}|Y_t),$$

by the recursion

$$\begin{aligned} v_t &= y_t - Z_t a_t, \\ F_t &= Z_t P_t Z_t' + H_t, \\ K_t &= T_t P_t Z_t' F_t^{-1}, \\ L_t &= T_t - K_t Z_t, \\ a_{t+1} &= T_t a_t + K_t v_t, \\ P_{t+1} &= T_t P_t L_t' + R_t Q_t R_t' \qquad t = 1, \ldots, n, \end{aligned} \tag{1.5}$$

with a_1 and P_1 as the mean vector and the variance matrix of α_1.

1.2.3 State smoother

We calculate

$$\hat{\alpha}_t = E(\alpha_t \mid Y_n), \quad V_t = Var(\alpha_t \mid Y_n),$$

by the backwards recursion

$$\begin{aligned} r_{t-1} &= Z_t' F_t^{-1} v_t + L_t' r_t, \\ N_{t-1} &= Z_t' F_t^{-1} Z_t + L_t' N_t L_t, \\ \hat{\alpha}_t &= a_t + P_t r_{t-1}, \\ V_t &= P_t - P_t N_{t-1} P_t \qquad t = n, \ldots, 1, \end{aligned} \tag{1.6}$$

with $r_n = 0$ and $N_n = 0$. The recursive nature of formulae (1.5) and (1.6), which arises from the Markovian nature of model (1.4), implies that calculations based on them are very fast on modern computers.

The proofs of these and many related results in state space theory can be derived very simply by the use of the following elementary lemma in multivariate normal regression theory. Suppose that x, y and z are random vectors of arbitrary orders that are jointly normally distributed with means μ_p and covariance matrices

$$\Sigma_{pq} = E[(p - \mu_p)(q - \mu_q)']$$

for $p, q = x, y$ and z with $\mu_z = 0$ and $\Sigma_{yz} = 0$. The symbols x, y, z, p and q are employed for convenience and their use here is unrelated to their use in other parts of the paper.

Lemma

$$\begin{aligned} E(x|y,z) &= E(x|y) + \Sigma_{xz}\Sigma_{zz}^{-1}z, \\ \mathrm{Var}(x|y,z) &= \mathrm{Var}(x|y) - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{xz}'. \end{aligned}$$

The proof of this familiar lemma can be obtained straightforwardly from elementary multivariate normal regression theory; see, for example, the DK book Section 2.13 for details. Proofs of the Kalman filter and smoother are given in the DK book, Sections 4.2 and 4.3. The elementary nature of this lemma drives home the point that the theoretical basis of state space analysis is very simple.

1.2.4 Simulation smoothing

A simulation smoother in Gaussian state space time series analysis draws samples from the Gaussian conditional distribution of state or disturbance vectors given the observations. This has proved important in practice for the analysis of non-Gaussian models and for carrying out Bayesian inference. Recently a new technique for implementing this has been proposed by Durbin and Koopman 2002 which is both simple and computationally efficient and which we now describe.

The construction of a simulation smoother for the state vector α_t is relatively simple given the lemma in Section 1.2.3. Since the state space model (1.4) is linear and Gaussian, the density $p(\alpha_1, \ldots, \alpha_n|Y_n)$ is multivariate normal. Its variance matrix has the important property that it does not depend upon Y_n; this follows immediately from the general result that in a multivariate normal distribution the conditional variance matrix of a vector given that a second vector is fixed does not depend on the second vector. These observations lead to a straightforward derivation of the following algorithm for drawing random vectors $\tilde{\alpha}_t$ from $p(\alpha|Y_n)$:

Step 1. Obtain random draws ε_t^+ and η_t^+ from densities $N(0, H_t)$ and $N(0, Q_t)$, respectively, for $t = 1, \ldots, n$. Generate α_t^+ and y_t^+ by means of recursion (1.4) with ε_t, η_t replaced by ε_t^+, η_t^+ where the recursion is initialised by the draw $\alpha_1^+ \sim N(a_1, P_1)$.

Step 2. Compute $\hat{\alpha}_t = E(\alpha_t|Y_n)$ and $\hat{\alpha}_t^+ = E(\alpha_t|Y_n^+)$ where $Y_n^+ = \{y_1^+, \ldots, y_n^+\}$ by means of standard filtering and smoothing using (1.5) forwards and (1.6) backwards.

Step 3. Take

$$\tilde{\alpha}_t = \hat{\alpha}_t - \hat{\alpha}_t^+ + \alpha_t^+,$$

for $t = 1, \ldots, n$.

This algorithm for generating $\tilde{\alpha}_t$ only requires standard Kalman filtering and state smoothing applied to the constructed series $y_1^+, \ldots, y_n^+$ and is therefore easy to incorporate in new software; special algorithms for simulation smoothing such as the ones developed by Frühwirth-Schnatter (1994c), Carter and Kohn (1994) and de Jong and Shephard (1995) are not required. The algorithm and similar ones for the disturbances are intended to replace those given in Section 4.7 of the DK book.

1.2.5 Missing observations

These are easy to handle in state space analysis. If observation y_j is missing for any j from 2 to $n-1$, all we have to do is put $v_j = 0$ and $K_j = 0$ in equations (1.5) and (1.6). The proof is given in Section 4.8 of the DK book.

1.2.6 Forecasting

This also is very easy in state space analysis. Suppose we want to forecast $y_{n+1}, \ldots, y_{n+k}$ given $y_1, \ldots, y_n$ and calculate mean square forecast errors. We merely treat $y_{n+1}, \ldots, y_{n+k}$ as missing observations and proceed using (1.5) as in Section 1.2.5. We use

$$Z_{n+1}a_{n+1}, \ldots, Z_{n+k}a_{n+k}$$

as the forecasts and use

$$V_{n+1}, \ldots, V_{n+k}$$

to provide mean square errors; for details see the DK book, Section 4.9.