

Chapter 7

State Space Models

7.1 Introduction

State Space models, developed over the past 10–20 years, are alternative models for time series. They include both the ARIMA models of Chapters 3–6 and the Classical Decomposition Model of Chapter 2 as special cases, but go well beyond both.

They are important because (i) they provide a rich family of naturally interpretable models for data, and (ii) they lead to highly efficient estimation and forecasting algorithms, through the *Kalman recursions* – see §7.3.

They are widely used and, perhaps in consequence, are known under several different names: structural models (econometrics), dynamic linear models (Statistics), Bayesian forecasting models (Statistics), linear system models (engineering), Kalman filtering models (control engineering),

The essential idea is that behind the observed time series X_t there is an underlying process \mathbf{S}_t which itself is evolving through time in a way that reflects the structure of the system being observed.

7.2 The Model

Example 1: The Random Walk plus Noise Model

For a series X_t with trend, but no seasonal or cyclic variation the Classical Decomposition of §2.3 is based on

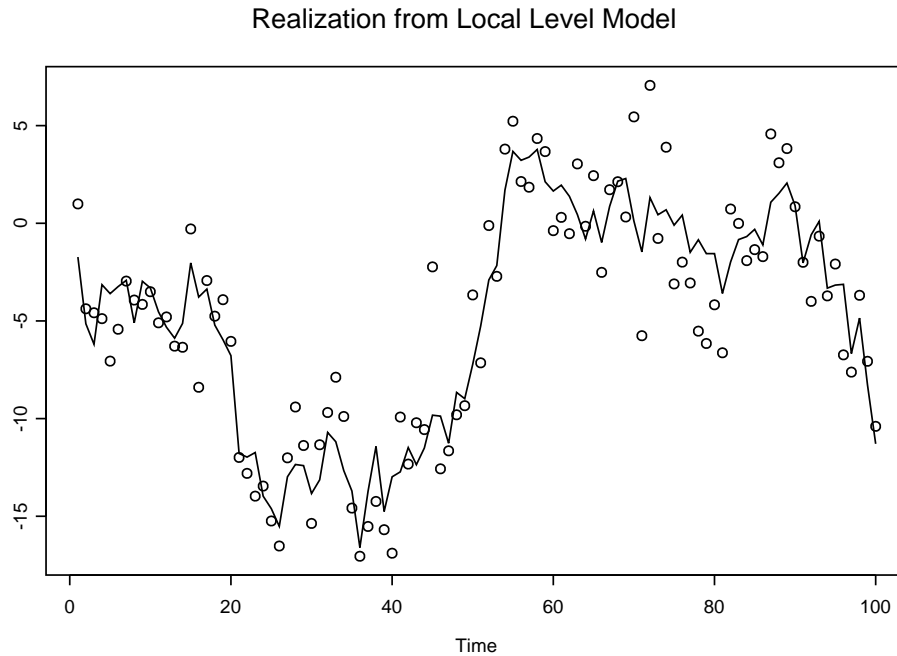
$$X_t = a_t + r_t, \quad (7.1)$$

where a_t represents deterministic trend – the underlying general level or signal at t – and r_t represents random variation or noise. To make this into a more precisely specified model we might suppose that r_t is white noise $\text{WN}(0, \sigma_r^2)$. Also, instead of supposing that a_t is deterministic we could take it to be random as well, but not changing much over time. A way of representing this would be to suppose

$$a_t = a_{t-1} + \eta_t, \quad (7.2)$$

where η_t is white noise $WN(0, \sigma_\eta^2)$, uncorrelated with the r_t . Equations (7.1) and (7.2) together define the *Random Walk plus Noise* model, also known as the *Local Level* model.

A realization from this model, with $a_0 = 0$ and $\sigma_r^2 = 6, \sigma_\eta^2 = 3$, is given below.



The model might be suitable for an industrial process, with process level a_t intended to be within design limits, but not directly observable itself. The sizes of σ_r^2 and σ_η^2 will determine local and large-scale smoothness respectively.

Review Question: How will σ_r^2 and σ_η^2 affect local and large-scale smoothness? How would the graph of a process for which σ_r^2/σ_η^2 is large differ from that of one for which it is small?

Example 2: A Seasonal Model with Noise

In the Classical Decomposition Method the seasonal/cyclic components s_t with a period c say were taken to be numbers which repeated themselves every c time units and summed to 0 over the c ‘seasons’: that is, for each t ,

$$s_{t+c} = s_t, \quad \sum_{j=1}^c s_{t+j} = 0 \text{ for each } t.$$

Given any $c - 1$ values s_1, \dots, s_{c-1} , we can generate such a sequence by setting

$$s_c = -s_1 - \dots - s_{c-1} \tag{7.3}$$

and

$$s_{t+c} = s_t \quad \text{for all } t = 1, 2, \dots \tag{7.4}$$

The result is a pattern *exactly* reproducing itself every c time units.

Note that (7.3) and (7.4) together amount to saying that s_t can be found successively from

$$s_t = -s_{t-c+1} - s_{t-c+2} - \dots - s_{t-1}, \quad t = c, c + 1, \dots \tag{7.5}$$

We can introduce some variability into the seasonal pattern by adding a white noise perturbation to (7.5):

$$s_t = -s_{t-c+1} - s_{t-c+2} - \cdots - s_{t-1} + \eta_t,$$

where η_t is, say, $\text{WN}(0, \sigma_\eta^2)$. On average (in expectation) the s_t 's will still sum to 0 over the seasons, but individual variability is now possible.

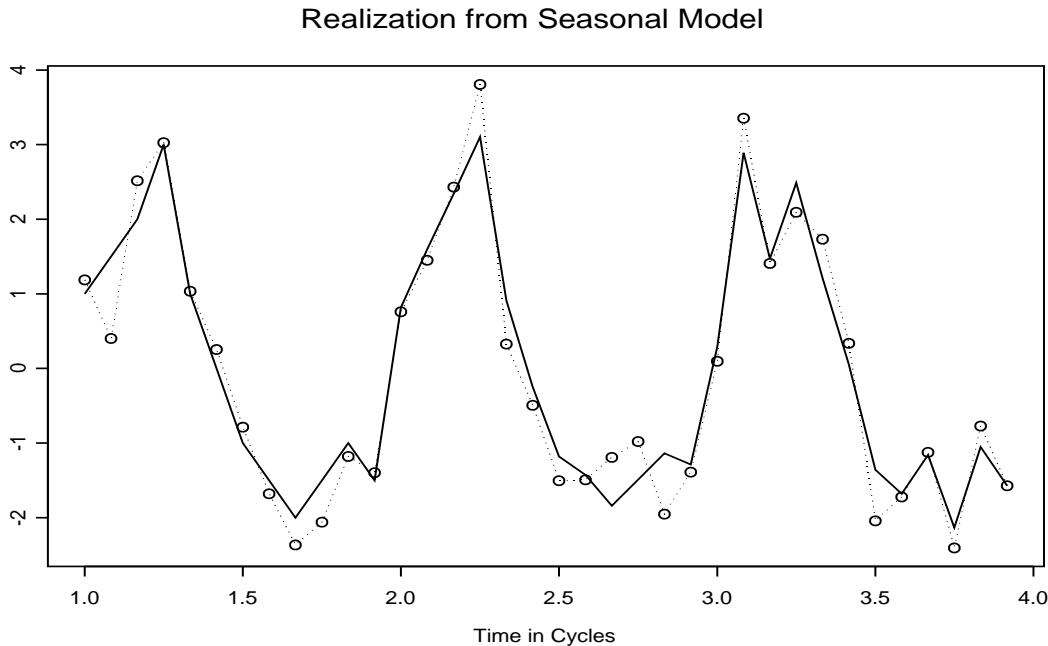
If, as in the previous example, we suppose that actual observations on the process are subject to error, we get

$$X_t = s_t + r_t \tag{7.6}$$

$$s_t = -\sum_{j=1}^{c-1} s_{t-j} + \eta_t, \tag{7.7}$$

where the observation error r_t might be taken to be $\text{WN}(0, \sigma_r^2)$.

The two equations (7.6) and (7.7) define the model. A realization with $c = 12$ is shown below.



If we write s_t, \dots, s_{t-c+2} as a vector, \mathbf{S}'_t say, then (7.7) can be written in matrix form as

$$\mathbf{S}_t = \begin{pmatrix} s_t \\ s_{t-1} \\ \vdots \\ \vdots \\ s_{t-c+2} \end{pmatrix} = \begin{pmatrix} -1 & -1 & \cdots & \cdots & -1 \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} s_{t-1} \\ s_{t-2} \\ \vdots \\ \vdots \\ s_{t-c+1} \end{pmatrix} + \begin{pmatrix} \eta_t \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix},$$

that is,

$$\mathbf{S}_t = \mathbf{F}\mathbf{S}_{t-1} + \mathbf{V}_t, \tag{7.8}$$

say, where F is the $(c-1) \times (c-1)$ matrix above and \mathbf{V}_t is the random vector $\mathbf{V}_t' = (\eta_t, 0, \dots, 0)$. The other equation defining the model, (7.6), may be written in terms of \mathbf{S}_t as

$$X_t = (1, 0, \dots, 0)\mathbf{S}_t + r_t. \quad (7.9)$$

General Form of State Space Model

In the examples each model consists of two parts: an underlying process m_t or \mathbf{S}_t , called in general the *state process* of the system, whose evolution is governed by one equation ((7.2) and (7.8) respectively), and another process – the observed time series X_t itself – called in general the *observation process*, which is related to the state process by another equation ((7.1) and (7.9) respectively).

In general a state space model consists of a pair of random quantities X_t and \mathbf{S}_t whose evolution and relationship are described by the equations:

$$X_t = G_t\mathbf{S}_t + \epsilon_t \quad (7.10)$$

$$\mathbf{S}_t = F_t\mathbf{S}_{t-1} + \mathbf{V}_t \quad (7.11)$$

where

\mathbf{S}_t denotes the state at time t ,

G_t and F_t are known matrices, possibly depending on time,

ϵ_t is $\text{WN}(0, \sigma_\epsilon^2)$,

\mathbf{V}_t is a vector of white noise processes, each uncorrelated with the ϵ_t process.

Equation (7.10) is called the *observation equation*, and equation (7.11) the *state* or *system equation*.

The component white noise processes of \mathbf{V}_t may be correlated with each other, though components of \mathbf{V}_t and \mathbf{V}_s for $t \neq s$ are taken to be uncorrelated. We use the notation $\mathbf{V}_t \sim \text{WN}(0, \{Q_t\})$ to mean that the random vector \mathbf{V}_t consists of univariate WN components (so that it has mean $\mathbf{0}$) and has variance-covariance matrix Q_t , that is:

$$E(\mathbf{V}_t\mathbf{V}_t') = Q_t.$$

In *Example 1* \mathbf{S}_t can be identified directly with a_t , and ϵ_t with r_t . G_t and F_t are both equal to the degenerate 1-dimensional unit matrix, and \mathbf{V}_t is the 1-dimensional vector with component η_t . The covariance matrix Q_t is therefore

simply σ_η^2 .

In *Example 2* the matrix $G_t = (1, 0, \dots, 0)$ as in (7.9), ϵ_t is r_t , the matrix F_t and the vector \mathbf{V}_t are as in (7.8), and Q_t is the $(c-1) \times (c-1)$ matrix with all entries zero except the top left hand one, which is equal to σ_η^2 .

Example 3: AR(1) Model

The stationary AR(1) process given by

$$X_t = \alpha X_{t-1} + \eta_t \tag{7.12}$$

is another example of a state space model. Identify the state \mathbf{S}_t with X_t itself, so that the state equation can be taken to be (7.12) if we set $F = \alpha$ and $\mathbf{V}_t = \eta_t$; and the observation equation is just $X_t = \mathbf{S}_t$, which has the form (7.10) with $G = 1$ and $\epsilon = 0$.

Notes

- (a) By iterating the state equation (7.11) we get

$$\begin{aligned} \mathbf{S}_t &= F_t \mathbf{S}_{t-1} + \mathbf{V}_t \\ &= F_t (F_{t-1} \mathbf{S}_{t-2} + \mathbf{V}_{t-1}) + \mathbf{V}_t \\ &\vdots \\ &= (F_t F_{t-1} \cdots F_2) \mathbf{S}_1 + (F_t \cdots F_3) \mathbf{V}_2 + \cdots + F_t \mathbf{V}_{t-1} + \mathbf{V}_t \\ &= f_t(\mathbf{S}_1, \mathbf{V}_2, \dots, \mathbf{V}_t) \end{aligned} \tag{7.13}$$

for a function f_t . From the observation equation therefore

$$\begin{aligned} X_t &= G_t f_t(\mathbf{S}_1, \dots) + \epsilon_t \\ &= g_t(\mathbf{S}_1, \mathbf{V}_2, \dots, \mathbf{V}_t, \epsilon_t) \end{aligned}$$

for a function g_t .

Thus the process is driven (through the G_t and F_t) by the white noise terms and the initial state \mathbf{S}_1 .

- (b) It turns out to be possible to put a large number of time series models – including, for example, all ARIMA models – into a state space form. An advantage of doing so is that the state equation gives a simple way of analysing the process \mathbf{S}_t , and from that it is easy via the observation equation to find out about the observation process X_t . If \mathbf{S}_1 and $\mathbf{V}_2, \dots, \mathbf{V}_t$ are independent (as opposed to just being uncorrelated) then \mathbf{S}_t has the *Markov property*, that is, the distribution of \mathbf{S}_t given $\mathbf{S}_{t-1}, \mathbf{S}_{t-2}, \dots, \mathbf{S}_1$ is the same as the distribution of \mathbf{S}_t given \mathbf{S}_{t-1} alone.

7.3 The Kalman Recursions

7.3.1 Filtering, Prediction and Smoothing

In state space models the state is generally the aspect of greatest interest, but it is not usually observed directly. What *are* observed are the X_t 's. So we'd like to have methods for estimating \mathbf{S}_t from the observations. Three scenarios are:

Prediction Problem	Estimate \mathbf{S}_t from X_{t-1}, X_{t-2}, \dots
Filtering Problem	Estimate \mathbf{S}_t from X_t, X_{t-1}, \dots
Smoothing Problem	Estimate \mathbf{S}_t from X_n, X_{n-1}, \dots , where $n > t$.

A further problem, which turns out to have an answer useful for other things too, is

X-Prediction Problem Estimate X_t from X_{t-1}, X_{t-2}, \dots

7.3.2 General Approach

Note that equation (7.13) above shows that \mathbf{S}_t and X_t are linear combinations of the initial state \mathbf{S}_1 and the white noise processes \mathbf{V}_t and ϵ_t . If these are *Gaussian*, then both \mathbf{S}_t and X_t will be Gaussian too for every t and their distributions will be completely determined by their means and covariances. Thus the whole evolution of the model will be known if the means and covariances can be calculated. The Kalman recursions give a highly efficient way of computing these means and covariances by building them up successively from earlier values. The recursions lead to algorithms for the problems above and for fitting the models to data. They are an enormously powerful tool for handling a wide range of time series models.

The basis of the Kalman recursions is the following simple result about multivariate Normal distributions.

7.3.3 Conditioning in a Multivariate Normal Distribution

Let \mathbf{Z} and \mathbf{W} denote random vectors with Normal distributions

$$\mathbf{Z} \sim \mathcal{N}(\mu_{\mathbf{z}}, \Sigma_{\mathbf{zz}}) \quad \mathbf{W} \sim \mathcal{N}(\mu_{\mathbf{w}}, \Sigma_{\mathbf{ww}})$$

and with covariance matrix

$$E((\mathbf{Z} - \mu_{\mathbf{z}})(\mathbf{W} - \mu_{\mathbf{w}})') = \Sigma_{\mathbf{zw}},$$

so that the distribution of the vector obtained by stacking \mathbf{Z} on \mathbf{W} is

$$\begin{pmatrix} \mathbf{Z} \\ \mathbf{W} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{\mathbf{z}} \\ \mu_{\mathbf{w}} \end{pmatrix}, \begin{pmatrix} \Sigma_{\mathbf{zz}} & \Sigma_{\mathbf{zw}} \\ \Sigma_{\mathbf{wz}} & \Sigma_{\mathbf{ww}} \end{pmatrix}\right).$$

Then

$$\mathbf{Z}|\mathbf{W} \sim \mathcal{N}\left(\mu_{\mathbf{z}} + \Sigma_{\mathbf{zw}}\Sigma_{\mathbf{ww}}^{-1}(\mathbf{W} - \mu_{\mathbf{w}}), \Sigma_{\mathbf{zz}} - \Sigma_{\mathbf{zw}}\Sigma_{\mathbf{ww}}^{-1}\Sigma_{\mathbf{wz}}\right). \quad (7.14)$$

For a proof, write down the ratio of the probability densities of $(\mathbf{Z}', \mathbf{W}')$ and of \mathbf{W} and complete the square in the exponent term.

7.3.4 The Recursions

Suppose after data $D_{t-1} = \{X_1, \dots, X_{t-1}\}$ have been observed we know by some means that the state \mathbf{S}_{t-1} has mean m_{t-1} and covariance matrix P_{t-1} , so that

$$\mathbf{S}_{t-1}|D_{t-1} \sim \mathcal{N}(m_{t-1}, P_{t-1}).$$

The recursions are built on relating the distributions of

(a) $\mathbf{S}_t|D_{t-1}$, and

(b) $\mathbf{S}_t|\{X_t, D_{t-1}\}$

to this.

For (a), because \mathbf{S}_t is related to \mathbf{S}_{t-1} through the state equation (7.11) ($\mathbf{S}_t = F_t\mathbf{S}_{t-1} + V_t$), it follows that, given D_{t-1} , \mathbf{S}_t is also Normally distributed

$$\mathbf{S}_t|D_{t-1} \sim \mathcal{N}(m_{t|t-1}, P_{t|t-1}), \quad (7.15)$$

and its mean vector and covariance matrix are

$$m_{t|t-1} = E(\mathbf{S}_t|D_{t-1}) = E(F_t\mathbf{S}_{t-1}|D_{t-1}) + E(V_t|D_{t-1}) = F_tm_{t-1} \quad (7.16)$$

and

$$P_{t|t-1} = E((\mathbf{S}_t - m_{t|t-1})(\mathbf{S}_t - m_{t|t-1})' | D_{t-1}) = F_tP_{t-1}F_t' + Q_t. \quad (7.17)$$

For (b), from (7.15) and the observation equation (7.10), ($X_t = G_t\mathbf{S}_t + \epsilon_t$) we find

$$E(X_t|D_{t-1}) = G_tm_{t|t-1}, \quad (7.18)$$

so that

$$X_t - E(X_t|D_{t-1}) = G_t(\mathbf{S}_t - m_{t|t-1}) + \epsilon_t$$

and hence

$$\text{Var}(X_t|D_{t-1}) = G_tP_{t|t-1}G_t' + \sigma_\epsilon^2. \quad (7.19)$$

Similarly

$$\text{Cov}(X_t, \mathbf{S}_t|D_{t-1}) = G_tP_{t|t-1}, \quad \text{Cov}(\mathbf{S}_t, X_t|D_{t-1}) = P_{t|t-1}G_t'.$$

Thus, given the data D_{t-1} , \mathbf{S}_t and X_t have the joint distribution

$$\left(\begin{array}{c} \mathbf{S}_t \\ X_t \end{array} \middle| D_{t-1} \right) \sim \mathcal{N} \left(\left(\begin{array}{c} m_{t|t-1} \\ G_tm_{t|t-1} \end{array} \right), \left(\begin{array}{cc} P_{t|t-1} & P_{t|t-1}G_t' \\ G_tP_{t|t-1} & G_tP_{t|t-1}G_t' + \sigma_\epsilon^2 \end{array} \right) \right).$$

It follows from the result in §7.3.3 that the conditional distribution of \mathbf{S}_t given the new observation X_t in addition to D_{t-1} (that is, the distribution of $\mathbf{S}_t|D_t$) is

$$\mathbf{S}_t|\{X_t, D_{t-1}\} \sim \mathcal{N}(m_t, P_t),$$

where

$$m_t = m_{t|t-1} + P_{t|t-1}G_t'\text{Var}(X_t|D_{t-1})^{-1}(X_t - G_tm_{t|t-1}) \quad (7.20)$$

$$P_t = P_{t|t-1} - P_{t|t-1}G_t'\text{Var}(X_t|D_{t-1})^{-1}G_tP_{t|t-1}. \quad (7.21)$$

The three equations (7.19), (7.20) and (7.21) – called the *updating equations* – together with (7.16) and (7.17) – called the *prediction equations* – are collectively referred to as the Kalman Filter equations. Given starting values m_0 and P_0 they can be applied successively to calculate the distribution of the state vector as each new observation becomes available. At any time they give values which contain all the information needed to make optimal predictions of future values of both the state and the observations, as follows.

7.3.5 The Prediction and Filtering Problems

By the general result about minimum mean square error forecasts in §6.1, the conditional mean of \mathbf{S}_t given $D_{t-1}, m_{t|t-1}$, is the minimum mean square error estimate of the state \mathbf{S}_t given observations up to and including time $t - 1$. The covariance matrix $P_{t|t-1}$ gives the estimation error variances and covariances. Thus the *prediction equations* (7.16) and (7.17) give the means to solve the **Prediction Problem** of §7.3.1.

In the same way the conditional mean of \mathbf{S}_t given D_t, m_t , is the solution to the **Filtering Problem** of §7.3.1, and the variances and covariances of the error in estimating \mathbf{S}_t by m_t are given by P_t .

7.3.6 The X -Prediction Problem

The minimum mean square error forecast of X_t given observations up to time $t - 1$, that is, given D_{t-1} , is simply $\hat{X}_t = G_t m_{t|t-1}$, by (7.18).

The prediction error, which we will denote by e_t , is therefore

$$e_t = X_t - \hat{X}_t = X_t - G_t m_{t|t-1} = G_t(\mathbf{S}_t - m_{t|t-1}) + \epsilon_t.$$

e_t is also known as the *innovation* at time t since it consists of the new information in the observation at t .

From the updating equation (7.20) it can be seen that the innovations play a key part in the updating of the estimate of \mathbf{S}_{t-1} to \mathbf{S}_t . The further e_t is from the zero vector, the greater the correction in the estimator of \mathbf{S}_{t-1} .

The innovations have means $E(e_t) = 0$, and variances, which we will denote by ϕ_t , given by

$$\begin{aligned} \phi_t = \text{Var}(e_t) &= E(X_t - G_t m_{t|t-1})^2 \\ &= G_t P_{t|t-1} G_t' + \sigma_\epsilon^2, \end{aligned} \tag{7.22}$$

from (7.19) and (7.21). The ϕ_t can be calculated straightforwardly from the Kalman filter equations.

7.3.7 Likelihood

The likelihood function, L say, for any model is the probability density (or probability in the discrete case) of the observed data, taken as a function of the unknown parameters, θ say. For a state space model therefore, if data $X_1 = x_1, \dots, X_t = x_t$ have been observed, and if p is the joint probability density function of X_1, \dots, X_t ,

$$\begin{aligned} L(\theta : \mathbf{x}) &= p(\mathbf{x} : \theta) \\ &= p(x_1 | \theta) \cdot \prod_{s=2}^t p(x_s | D_{s-1} : \theta) \end{aligned}$$

where the density function $p(x_s | D_{s-1})$ is that of the Normal distribution (of X_s given D_{s-1}) with mean $E(X_s | D_{s-1}) = \hat{X}_s = G_s m_{s|s-1}$ and variance ϕ_s given by (7.22).

Thus

$$\begin{aligned}\log L &= \text{const} + \log p(x_1|\theta) - \frac{1}{2} \sum_{s=2}^t \log |\phi_s| - \frac{1}{2} \sum_{s=2}^t \frac{(x_s - \hat{X}_s)^2}{\phi_s} \\ &= \text{const} + \log p(x_1|\theta) - \frac{1}{2} \sum_{s=2}^t \log |\phi_s| - \frac{1}{2} \sum_{s=2}^t \frac{e_s^2}{\phi_s},\end{aligned}$$

which is easily calculated from the innovations and their variances, and $p(x_1|\theta)$ if necessary. Standard methods of numerical maximization may then be used to estimate the unknown parameters θ . This is the approach described in §5.3.1.

Summary of Ideas in Chapter 6

State space models specified by

- state variables
- observation variables

and by *state equations* describing the evolution of states, and *observation equations* describing the relationship of observations to states.

Special cases include the ARIMA models studied in earlier chapters, and models underlying the Decomposition Method of §2.3 of Chapter 2.

Various problems in relation to forecasting in state space models may be specified:

- prediction
- filtering
- smoothing
- X -prediction

Solutions are based on the fact that if variables are Gaussian, then to describe the whole evolution of the system all that's needed are the means and variances-covariances of the variables through time. These can be calculated recursively by the *Kalman recursions*.

The Kalman recursions yield immediately

- forecasts and their error variances
- efficient computation of the likelihood function, and therefore a powerful way of fitting models.