ST 337 / ST 405: Bayesian Forecasting and Intervention

Dr Bärbel Finkenstädt Rand[1], Department of Statistics, University of Warwick

[1] B.F.Finkenstadt@warwick.ac.uk.

CONTENTS

CHAPTER 2
INTRODUCTION TO THE DLM

In this chapter we will focus on the Dynamic Linear Model (DLM) defining and analysing its main building blocks and deriving the updating and forecasting equations. The DLM is a *state-space model*. First we will introduce some basic concepts. In this course all the examples we consider will have a state vector θ_t, an absolutely continuous random vector of dimension p at each time t. The observations $(Y_t)_{t\geq 1}$ are a sequence of random vectors of dimension m and will usually be Gaussian (and hence absolutely continuous), although they can sometimes be discrete or mixed. In this section we will assume we are in a closed system so that we can only learn about the future through past observations. In our notation we will use $x^t = \{x_1, \ldots, x_t\}$ throughout for any quantity x while π will generally denote a pdf as specified by its arguments. We may also use the notation $x_{t_1:t_2}$ to denote the set of values from time index t_1 to t_2, that is $x_{t_1:t_2} = \{x_{t_1}, x_{t_1+1}, ..., x_{t_2}\}$ for $t_1 \leq t_2$.

Consider a time series $(Y_t)_{t\geq 1}$. Specifying the joint finite-dimensional distribution of $(Y_1, ..., Y_t)$ for any $t \geq 1$ is a challenging task as in time series applications the assumption of independence is generally not satisfied. Arguably the simplest form of dependence assumes Markovian dependence. We say that a process $(Y_t)_{t\geq 1}$ is a *Markov chain* if, for any $t > 1$

$$\pi(y_t \mid y_{1:t-1}) = \pi(y_t \mid y_{t-1}).$$

This means that all the information about Y_t is contained in the previous value Y_{t-1} alone and that Y_t and Y^{t-2} are conditionally independent given y_{t-1}. Hence, for a Markov chain, the joint distribution has a simple form

$$\pi(y_{1:t}) = \pi(y_1)\prod_{j=2}^{t}\pi(y_j \mid y_{j-1}).$$

In the DLM we will not assume a Markovian structure for the observations but for the unobserved states, *i.e.* we assume there is an unobserved Markov chain θ_t and that the observable Y_t is some imprecise measurement of θ_t. Formally the state-space model is defined on the basis of the following assumptions:

(A1) θ_t is a Markov Chain.

(A2) Conditionally on θ_t, the Y_t's are independent and Y_t depends on θ_t only.

A *state space model* consists of an $\mathbb{R}^m$ valued time series $(Y_t)_{t\geq 1}$ and a sequence of $\mathbb{R}^p$ valued states $(\theta_t)_{t\geq 0}$ satisfying (A1) and (A2). The information flow in a state space model is represented in the conditional independence graph seen in chapter 1 which is a special case of a directed acyclic graph (DAG). State-space models where the states are discrete-valued random variables are usually called hidden Markov models.

2.1 Joint distribution of observations and states

Our first task shall be to write down the joint density $\pi(\theta_{0:t}, y_{1:t})$ of all observations and states. In a state space model we have because of (A1) and (A2) for any $t > 0$

$$\pi(\theta_{0:t}, y_{1:t}) = \pi(\theta_0)\prod_{j=1}^{t}\pi\left(\theta_j \mid \theta_{j-1}\right)\pi\left(y_j \mid \theta_j\right) . \tag{2.1}$$

Equation (2.1) tells us that in a closed system, the joint density of states and observations is completely specified as a simple product of

- the prior density $\pi(\theta_0)$,
- the *observation densities* $\pi\left(y_j \mid \theta_j\right)$
- and the *state transition densities* $\pi\left(\theta_j \mid \theta_{j-1}\right)$.

From (2.1) one can derive by conditioning or marginalization any other distribution of interest.

Later we will be interested in the distribution of the state vector θ_t, conditional on what we know up to time t (in a closed system this is either y^{t-1} or y^t). Clearly it is possible, in principle, to calculate this distribution directly from (2.1), whatever the observation and transition densities are, using the familiar operations of *marginalisation* and *conditionning*. We will see that for the Gaussian setup this is quite straightforward.

It is often the case that the time series distribution (2.1) can be further simplified because neither the observation densities (the distribution that relates the observation y_t with its state) nor the transition densities (which drives the transition of states) depend on the time index t. This is the case for, so-called, constant DLM's.

2.2 A basic DLM

Dynamic linear models (DLM) are state space models where the relationship between consecutive states is linear. Moreover, in a *constant* DLM the densities do not depend on time. Furthermore, if we assume Gaussianity of all implied distributions then a simple example is given by the *Random Walk plus Noise*

$$Y_t \mid \theta_t \sim \mathrm{N}\left(y_t \mid \theta_t, V_t\right)$$

for the observations,

$$\theta_t \mid \theta_{t-1} \sim \mathrm{N}\left(\theta_t \mid \theta_{t-1}, W_t\right)$$

for the transition between states, and

$$\theta_0 \mid D_0 \sim \mathrm{N}\left(\theta_0 \mid m_0, C_0\right)$$

for the prior. The representation above is often called the *c*onditional representation. Defining zero-mean random quantities $\{v_t, \omega_t \; : \; t = 1, 2, \ldots\}$, which are all independent of each other, the model can be written in an equivalent additive form, the *a*dditive representation:

$$\begin{array}{rll}
\text{observation eqn.} & Y_t = \theta_t + v_t & v_t \sim \mathrm{N}\left(v_t \mid 0, V_t\right) \text{ ind.} \\
\text{system eqn.} & \theta_t = \theta_{t-1} + \omega_t & \omega_t \sim \mathrm{N}\left(\omega_t \mid 0, W_t\right) \text{ ind.} \\
\text{prior} & \theta_0 \mid D_0 \sim \mathrm{N}\left(\theta_0 \mid m_0, C_0\right) &
\end{array}$$

At each time point, Y_t is a noisy observation (measured with error) of a true signal or level. The magnitude of the measurement error is determined by the *observational variance*, V_t. The true level (the state at time t) cannot be observed directly and is modelled like a random walk. So, the level at time t is expected to be centered at the same level as time $t-1$ but is randomly perturbed; the magnitude of the perturbation is determined by the variance W_t. This simple Gaussian setup describes a host of processes, particularly those related with stock control and sales of a steady selling product. Therefore, this model is often referred to as the (Gaussian) steady model, or the first order polynomial model.

Note: above we have assumed that all error terms $\{v_t, \omega_t\}$ are mutually independent. Denoting by $x \amalg y|z$ that random quantities x and y are independent conditionally on the random variable z, this assumption on the error terms is equivalent to

$$\nu_t \amalg \omega^t, \nu^{t-1} \text{ and } \omega_t \amalg \omega^{t-1}, \nu^{t-1}.$$

Bearing in mind that ω^t can be written as a function of the entire past of all θs, i.e. θ^t, and ν^{t-1} is a function of θ^{t-1}, y^{t-1} this assumption itself can be shown to be exactly equivalent to

$$\nu_t \amalg \theta^t, y^{t-1} \text{ and } \omega_t \amalg \theta^{t-1}, y^{t-1}.$$

Remark that the latter will e.g. imply that $\omega_t \amalg \theta^{t-1} | y^{t-1}$ and also that $\omega_t \amalg \theta_{t-1} | y^{t-1}$.

Example 1.
In order to illustrate the relationship established in the equations above, suppose that $\theta_0 = 25$ was drawn from a prior on θ_0. Also, assume that $V_t = V = 1$, and that *a*) $W_t = W = 0.05$ and *b*) $W_t = W = 0.5$. Figure 2.1 depicts one simulated realisation of both processes.

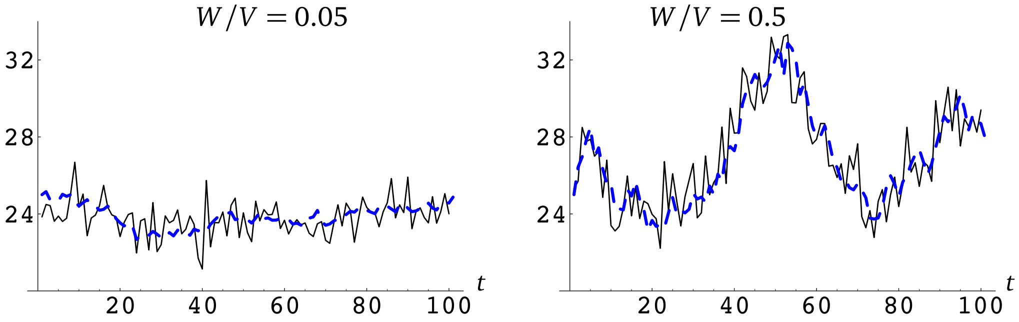

Figure 2.1. *Two example series. The series on the left has $W/V = 0.05$ and the series on the right has $W/V = 0.5$. The level of the series, θ_t, is drawn with the broken line.*

Note that if W is large (respect to V) it is more probable that the observed series presents larger fluctuations than when the relation W/V is small. The larger W, the larger the fluctuations in the underlying level θ_t. $\triangleleft$

2.3 Definition of a DLM

An important class of state space models is given by the Gaussian linear state space models, also called Gaussian DLM. A Gaussian DLM is characterised by a quadruple $\{F_t, G_t, V_t, W_t\}$, $t = 1, 2, \ldots$, such that

$$\begin{aligned} &\text{observation equation} && Y_t = F_t\,\theta_t + \nu_t && \text{with } \nu_t \sim \mathrm{N}(\nu_t \mid 0, V_t) \text{ ind.} \\ &\text{system equation} && \theta_t = G_t\,\theta_{t-1} + \omega_t && \text{with } \omega_t \sim \mathrm{N}(\omega_t \mid 0, W_t) \text{ ind.} \end{aligned}$$

combined with some prior information $\theta_0 \sim \mathrm{N}\left(\theta_0 \mid m_0, C_0\right)$, the initial state of the system. G_t and F_t are known matrices of order p x p and m x p, respectively, and $(v_t)_{t\geq 1}$ and $(\omega_t)_{t\geq 1}$ are two independent sequences of independent Gaussian random vectors with mean zero and known variance matrices $(V_t)_{t\geq 1}$ and $(W_t)_{t\geq 1}$, respectively. G_t is relating the components of states at time t with the same components at time $t-1$ and is often called the *transition matrix* of the DLM. The error sequences $\boldsymbol{\omega}_t$ and $\boldsymbol{v}_t$ are internally and mutually independent. At any time t

F_t is the *design matrix* linking observable Y_t to its state;

θ_t is the state –or system– vector;

$\mu_t = F_t\theta_t$ is the mean response –level.

v_t is the observational error.

G_t is the evolution, system, transfer or state matrix;

ω_t is the system – or evolution– error.

There are some particular cases which may be of interest:

TSDLM If $F_t = F$ and $G_t = G$ for all t, then $\left\{F, G, V_t, W_t\right\}$ is known as a Time Series DLM.

CDLM A TSDLM whose observation and evolution variance matrices are constant over time is known as a constant DLM.

Univariate When $m = 1$, *i.e.* $y_t \in \mathbb{R}$, then $\left\{F_t, G_t, v_t, W_t\right\}$ is a univariate DLM.

One can show that a Gaussian DLM satisfies (A1) and (A2) with conditional densities

$$Y_t \mid \theta_t \sim \mathrm{N}\left(y_t \mid F_t\theta_t, V_t\right) \text{ ind.}$$
$$\theta_t \mid \theta_{t-1} \sim \mathrm{N}\left(\theta_t \mid G_t\theta_{t-1}, W_t\right) \text{ ind.}$$

In the sequel we will often use DLM to refer to a Gaussian DLM. Typically, the distributions involved are assumed to be Normal, but this is not necessary.

The quadruple $\{F, G, V, W\}_t$, together with the parameters of the initial density of the state vector, specify the joint (often Gaussian) distribution of (Y^t, θ^t) for any $t \geq 1$. It is often the case that, as time advances, the prior on θ_0 has little impact on the forecasts of later observations, so we typically think of a DLM as being specified through $\{F, G, V, W\}_t$.

Note that the Random Walk plus noise model (normal steady model, first order polynomial model)

$$\begin{array}{rll} \text{observation eqn.} & Y_t = \theta_t + \nu_t & \nu_t \sim \mathrm{N}\left(\nu_t \mid 0, V_t\right) \text{ ind.} \\ \text{system eqn.} & \theta_t = \theta_{t-1} + \omega_t & \omega_t \sim \mathrm{N}\left(\omega_t \mid 0, W_t\right) \text{ ind.} \\ \text{prior} & \theta_0 \mid D_0 \sim \mathrm{N}\left(\theta_0 \mid m_0, C_0\right) & \end{array}$$

is a Gaussian DLM characterized by the quadruple $\{F_t, G_t, V_t, W_t\} = \{1, 1, V_t, W_t\}$. Also $m = p = 1$, *i.e.* both variables Y_t and θ_t are univariate.

Knowledge The knowledge (information) available at any time t is denoted by D_t.

Initial knowledge The information about the initial state of the system, θ_0 is captured by a distribution which has mean m_0 and variance C_0. We will denote it by D_0.

Closed DLM A closed DLM has $D_t = \{y_t, D_{t-1}\}$.

Forecast The one-step forecast distribution at time $t-1$, given D_{t-1} is the distribution of $Y_t \mid D_{t-1}$ where $f_t = \mathrm{E}[Y_t \mid D_{t-1}]$ and $Q_t = \mathrm{Var}[Y_t \mid D_{t-1}]$.

Forecast error The one-step forecast error is defined as $e_t = y_t - f_t$, so that $\mathrm{E}[e_t \mid D_{t-1}] = 0$ and $\mathrm{Var}[e_t \mid D_{t-1}] = Q_t$.

Remark 2.1. *In a closed system we have $D_t = \{D_{t-1}, y_t\}$ and hence $D_t = \{D_0, y_1, ..., y_t\} = \{D_0, y^t\}$. Usually we will simply write y^t for notational convenience and write D_t as the conditionning argument if we want to make dependence on initial information D_0 explicit.*

The great flexibility of state space models is the major reason for their extensive use in a large range of applied problems. We start by deriving some general results about state space models and how to update them sequentially in order to be able to forecast. We then go on to show that when the distributions involved are Gaussian, the algebra of the updating and forecast equations is straightforward. At the moment we will not consider issues about the model building and start by assuming that our model is given, *i.e.* we assume that the densities $\pi(y_t \mid \theta_t)$ and $\pi(\theta_t \mid \theta_{t-1})$ have been specified.

For a given state space model, and based on the information available, two main tasks are

1. to make inference on unobserved states, *i.e.* estimate the state vector. To do this we compute the conditional density $\pi(\theta_s \mid y^t)$. On can distinguish between

   - filtering: $s = t$. To solve the filtering problem we compute the conditional density $\pi(\theta_t \mid y^t)$ called *filtering density;*

- *state prediction: $s > t$;*
- *smoothing: $s < t$ (retrospective analysis).*

2. to predict or forecast future observations and states.

2.4 Updating (filtering) in state space models

The concept of filtering is usually associated with data arriving sequentially in time. An example would be a financial application where one has to estimate, day by day, the term structure of interest rates, updating the current estimate as new data comes in. In a Gaussian DLM, the so-called *Kalman Filter* provides formulae for updating the current inference on the state vector as a new observation y_t becomes available for passing from any previous filtering density $\pi(\theta_{t-1} \mid y^{t-1})$ to the next filtering density $\pi(\theta_t \mid y^t)$. We start by solving the filtering problem. It will serve as the starting point to solving any forecasting or prediction problem. We will not consider the problem of smoothing here. The following proposition formulates and proves the updating or filtering rules for a general state space model.

Proposition 1: (Updating (filtering) in state space models)
For a general state space model defined by assumptions (A1) & (A2) above the following statements hold

(i) The 1-step-ahead predictive density for the states $\pi(\theta_t \mid y^{t-1})$ can be computed from the filtering density $\pi(\theta_{t-1} \mid y^{t-1})$ according to

$$\pi(\theta_t \mid y^{t-1}) = \int \pi(\theta_t \mid \theta_{t-1})\pi(\theta_{t-1} \mid y^{t-1})d\theta_{t-1}.$$

(ii) The 1-step-ahead predictive density for the observations $\pi(y_t \mid y^{t-1})$ can be computed from the predictive density of states as

$$\pi(y_t \mid y^{t-1}) = \int \pi(y_t \mid \theta_t)\pi(\theta_t \mid y^{t-1})d\theta_t.$$

(iii) The filtering density can be computed from the above densities via

$$\pi(\theta_t \mid y^t) = \frac{\pi(y_t \mid \theta_t)\pi(\theta_t \mid y^{t-1})}{\pi(y_t \mid y^{t-1})}.$$

Proof: See lecture. The proof relies heavily on assumptions (A1) and (A2).

2.4.1 Updating in a Gaussian DLM

Gaussian DLMs are an important case where standard results about multivariate Gaussian distributions can be used implying that all marginal and conditional distributions are Gaussian and thus are fully specified by their means and variance matrices. The filtering equations stated in Proposition 1 for Gaussian DLMs are solved by the famous *Kalman Filter* given as follows:

Proposition 2: (Kalman Filter)

Consider the Gaussian DLM characterised by the quadruple $\{F_t, G_t, V_t, W_t\}$, $t = 1, 2, \ldots$ with prior information $\theta_0 \sim \mathrm{N}\left(\theta_0 \mid m_0, C_0\right)$. Suppose $\theta_{t-1} \mid y^{t-1} \sim \mathrm{N}\left(\theta_{t-1} \mid m_{t-1}, C_{t-1}\right)$. Then

(i) The 1-step-ahead predictive density for the states is Gaussian $\theta_t \mid y^{t-1} \sim \mathrm{N}\left(\theta_t \mid a_t, R_t\right)$ with parameters

$$\begin{aligned} a_t &= \mathrm{E}[\theta_t \mid y^{t-1}] = G_t m_{t-1} \\ R_t &= \mathrm{Var}[\theta_t \mid y^{t-1}] = G_t C_{t-1} G_t' + W_t \end{aligned}$$

(ii) The 1-step-ahead predictive density of $y_t \mid y^{t-1}$ is Gaussian with parameters

$$\begin{aligned} f_t &= \mathrm{E}[Y_t \mid y^{t-1}] = F_t a_t \\ Q_t &= \mathrm{Var}[Y_t \mid y^{t-1}] = F_t R_t F_t' + V_t \end{aligned}$$

(iii) The filtering distribution of $\theta_t \mid y^t$ is Gaussian with parameters

$$\begin{aligned} m_t &= \mathrm{E}[\theta_t \mid y^t] = a_t + R_t F_t' Q_t^{-1} e_t \\ C_t &= \mathrm{Var}[\theta_t \mid y^t] = R_t - R_t F_t' Q_t^{-1} F_t R_t \end{aligned}$$

where $e_t = y_t - f_t$ is the forecast error.

Proof: See lecture.

The Kalman filter in proposition 2 allows us to compute predictive and filtering distributions starting from $\theta_0 \sim \mathrm{N}\left(\theta_0 \mid m_0, C_0\right)$, then computing $\pi(\theta_1 \mid y_1)$ and continuing recursively as new data becomes available.

Note that in step (iii) the expression of m_t has an intuitive estimation-correction form in that the new filtered mean is equal to the prediction mean a_t plus a correction term which depends on how much the new observation differs from its prediction

$$m_t = a_t + R_t F_t' Q_t^{-1}(y_t - F_t a_t) = a_t + K_t e_t.$$

The weight of the correction term is given by the *adaptive coefficient* $K_t = R_t F_t' Q_t^{-1}$ which depends on the observation variance V_t (through Q_t) and on $R_t = \text{Var}[\theta_t \mid y^{t-1}]$ which is a function of the variance of the transition density W_t.

All matrices in the Kalman filter are easily computable. In R the Kalman filter is performed in the package 'dlm' by the function 'dlmFilter'.

A brief note on smoothing (retrospective updating): Suppose you are given data $y^T = \{y_1, ..., y_T\}$ then it may be of interest to reconstruct the behaviour of the system retrospectively, *i.e.* reconstruct the unobserved values of the states $\{\theta_1, ..., \theta_T\}$. This is achieved through a backward recursive algorithm to compute $\pi(\theta_t \mid y^T)$ for any $t < T$ starting from the filtering distribution $\pi(\theta_T \mid y^T)$ and iterating backwards. Again, in the Gaussian DLM the distributions involved are all Gaussian and the algorithm is provided by the so-called 'Kalman smoother' which we will not consider in detail here.

Example 2.
Consider the constant steady model

$$\begin{aligned} Y_t &= \theta_t + \nu_t & \nu_t &\sim \mathrm{N}\left(\nu_t \mid 0, V\right) \text{ ind.} \\ \theta_t &= \theta_{t-1} + \omega_t & \omega_t &\sim \mathrm{N}\left(\omega_t \mid 0, W\right) \text{ ind.} \end{aligned} \qquad \triangleleft$$

which is a Gaussian DLM with $m = p = 1$ and $F_t = G_t = 1$. Suppose V and W are known constants. Let $\theta_{t-1} \mid y^{t-1} \sim \mathrm{N}\left(\theta_{t-1} \mid m_{t-1}, C_{t-1}\right)$. Applying the three steps of the Kalman filter gives:

(i) The 1-step-ahead predictive density for the states is Gaussian $\theta_t \mid y^{t-1} \sim \mathrm{N}\left(\theta_t \mid a_t, R_t\right)$ with

$$\begin{aligned} a_t &= G_t m_{t-1} = m_{t-1} \\ R_t &= G_t C_{t-1} G_t' + W_t = C_{t-1} + W \end{aligned}$$

Hence $\theta_t \mid y^{t-1} \sim \mathrm{N}\left(\theta_t \mid m_{t-1}, C_{t-1} + W\right)$

(ii) $y_t \mid y^{t-1} \sim \mathrm{N}\left(y_t \mid f_t, Q_t\right)$ with parameters

$$\begin{aligned} f_t &= F_t a_t = m_{t-1} \\ Q_t &= F_t R_t F_t' + V_t = R_t + V = C_{t-1} + W + V \end{aligned}$$

Hence $y_t \mid y^{t-1} \sim \mathrm{N}\left(y_t \mid m_{t-1}, C_{t-1} + W + V\right)$,*i.e.* the 1-step-ahead prediction for the observation is the same as for the state with variance equal to the one of the state prediction plus observational variance.

(iii) The filtering distribution is $\theta_t \mid y^{t-1} \sim \mathrm{N}\left(\theta_t \mid m_t, C_t\right)$ where

$$\begin{aligned} m_t &= a_t + R_t F_t' Q_t^{-1} e_t = m_{t-1} + K_t e_t \\ C_t &= R_t - R_t F_t' Q_t^{-1} F_t R_t = V K_t \end{aligned}$$

where $e_t = y_t - f_t$ and the adaptive coefficient is

$$K_t = R_t F_t' Q_t^{-1} = \frac{R_t}{Q_t} = \frac{C_{t-1} + W}{C_{t-1} + W + V}$$

In particular, note that in step (iii) we have

$$\begin{aligned} m_t &= m_{t-1} + K_t (y_t - m_{t-1}) \\ &= (1 - K_t) m_{t-1} + K_t y_t \end{aligned}$$

and since $0 < K_t < 1$ the mean of the filtering distribution is a weighted average between the previous mean m_{t-1} and the new observation y_t. The weights are determined by the adaptive coefficient which depends on the ratio $\frac{W}{V}$. If $V \to 0$, then $\frac{W}{V}$ becomes large and $K_t \to 1$, and hence in the case of very small observational noise the 1-step ahead prediction m_t is given by the most recent observation . On the other hand if V is large then $\frac{W}{V}$ is small and y_t receives a smaller weight. In the constant steady model we also have, from step (iii), for the precision

$$C_t^{-1} = (C_{t-1} + W)^{-1} + V^{-1}$$

and one can derive the following simple updating rule for the adaptive coefficient

$$K_t^{-1} = (K_{t-1} + \frac{W}{V})^{-1} + 1$$

2.4.2 Discounting for DLM state evolution variances

Discounting is a central and practically important method to handle models with unknown variance W_t. Recall from step (i) of the Kalman filter

$$R_t = \mathrm{Var}[\theta_t \mid y^{t-1}] = G_t C_{t-1} G_t' + W_t = P_t + W_t.$$

The matrix $P_t = G_t C_{t-1} G_t'$ can be seen as the prior variance in a DLM with no evolution error, *i.e.* in a model where $W_t = 0$ where the state vector is stable and requires no stochastic variation. This is not a realistic assumption, however it can be assumed that $R_t = \frac{P_t}{\delta}$ for $0 < \delta \leq 1$, and so the prior variance at time t is that of a model without stochastic variation times a correction factor which inflates such variance. Combining $R_t = \frac{P_t}{\delta}$ with the recursion above for R_t we have that

$$W_t = \frac{1-\delta}{\delta} P_t .$$

Low values of the discount factor δ are consistent with high variability in the θ_t sequence, while high values $\delta > 0.9$ are typically more relevant in practice. (Note: if one performs parameter estimation, then δ can be estimated along with other parameters).

In the constant steady model we have

$$C_t^{-1} = \left[C_{t-1} + W\right]^{-1} + V^{-1} ,$$

and using $C_t = V K_t$ we get in equilibrium (i.e. in the limit as $t \to \infty$)

$$K^{-1} = [K + s]^{-1} + 1 ,$$

thus

$$s = \frac{1}{K^{-1} - 1} - K = \frac{K}{1-K} - K .$$

Also,

$$W = s V = \frac{C}{1-K} - C = C \frac{K}{1-K}$$

i.e. W is a fixed proportion of C, the state variance. So, once we update one state to the next, we have in the limit

$$R = C + W = \frac{C}{1-K} = C \frac{1}{\delta} ,$$

where δ is the *discount factor* ($0 < \delta \leq 1$). Thus, once we update to the next state (and before new information is available), we increase a bit the uncertainty by a constant factor.

Generally, we choose

$$W_t = C_{t-1} \frac{1-\delta}{\delta} ,$$

as a ‘reasonable’ choice for W_t with a certain discount rate δ, (usually $\delta \in (0.8, 1)$).

Example 3.
Using the formulae above, for

$$\begin{array}{llll} \delta = 1 & \text{we get} & W_t = 0 & R_t = C_{t-1} \\ \delta = 0.8 & \text{we get} & W_t = \frac{1}{4} C_{t-1} & R_t = \frac{5}{4} C_{t-1} . \end{array}$$

◁

Example 4.
In this example we will use the CANDY data set[1] which comprises sales data from 01/76 to 12/81 (i.e. $T = 72$ data points). This series is depicted in Figure 2.2 where we can see that it is not very stable, but may vary in a *seasonal* manner; however let us not introduce this concept yet.

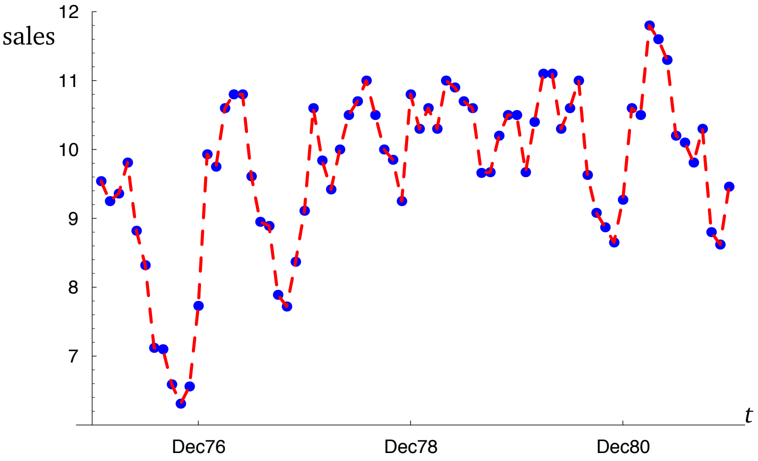

Figure 2.2. *Time plot of* SALES *series from data set* CANDY.

We will fit the simple steady model

$$\begin{aligned} Y_t &= \theta_t + \nu_t & \nu_t &\sim \mathrm{N}\left(\nu_t \mid 0, V_t\right) \\ \text{sales} &= \text{level} + \text{obs. error} & & \\ \theta_t &= \theta_{t-1} + \omega_t & \omega_t &\sim \mathrm{N}\left(\omega_t \mid 0, W_t\right) \end{aligned}$$

From the discussion above regarding the choice of W_t, we will set the discount factor to $\delta = 0.9$ and for the prior distribution we will use $(m_0, C_0) = (0, 1)$. Also, we set the observational variance $V_t = 1$.

Substituting this initial value and using the recursive formulae given above we can produce the one-step ahead forecast, $f_t = \mathrm{E}[Y_t \mid y^{t-1}]$, for $t = 1, \ldots, T$. This series is plotted in Figure 2.3, together with the 90% HPD. There we can see how, as more data becomes available, the uncertainty of our forecasts decreases (the HPD intervals are narrower), but not too fast, given that the level, θ_t, also varies over time.

It can be argued that this forecast function performs badly, mainly because seasonal variations are not included in the model; however, HPD intervals are so wide that they cover almost all the observed data (we expect them to cover 90% of the observations).

Figure 2.4 plots the estimated level (trend) values, $\mathrm{E}[\theta_t \mid y^{t-1}]$, for this series, together with the 90% HPD intervals. First, note that given our model specification, $\mathrm{E}[\theta_t \mid y^{t-1}] = f_t$.

[1] from Pole, West & Harrison, Applied Bayesian Forecasting and Time Series Analysis

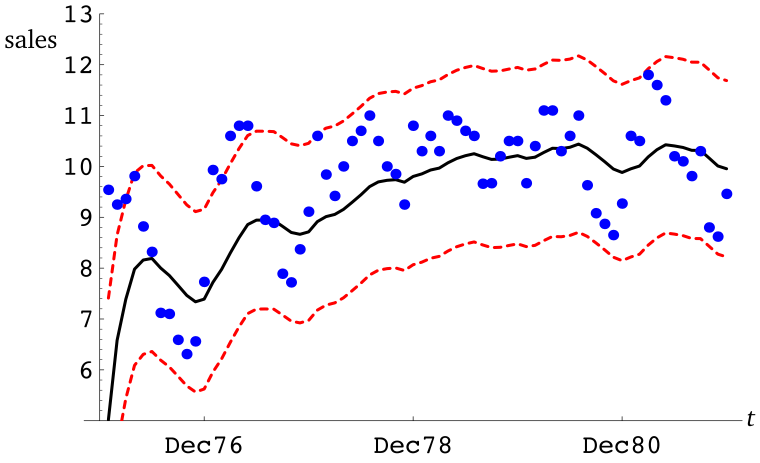

Figure 2.3. *One-step ahead forecast values (solid line) and the 90% HPD intervals (dotted lines) for the candy sales data (points).*

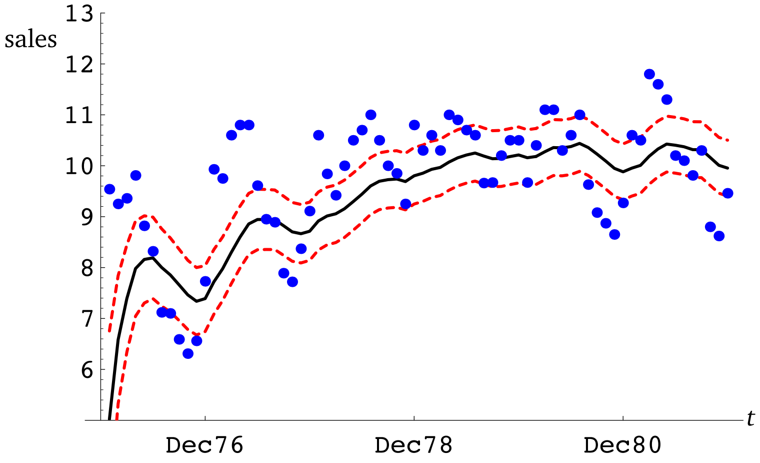

Figure 2.4. *Estimated trend values (solid line) and the 90% HPD intervals (dotted lines) for the candy sales data (points).*

It is also interesting to note that the credible intervals for the trend are narrower, since they do not involve the observational error (with variance one).

For the sake of comparison and to understand how the choice of prior and tuning parameters can affect inferences through time, let us repeat the above analysis but this time with $m_0 = 50$ (clearly a 'wild guess') and $\delta = 0.7$.

Figures 2.5 and 2.6 illustrate the main consequences of an inappropriate selection of these parameters. First, notice that the first few forecasts and estimated trends are, compared with the previous analysis, quite far away from their actual values; however, as information accumulates, both point estimations come close to each other. On the other hand, it is apparent how an ill-chosen discount factor will have a persistent effect over time; in this case, a very low δ yields too wide HPD regions, which are not very useful.

The low discount factor artificially blows up the uncertainty about the level (see Exam-

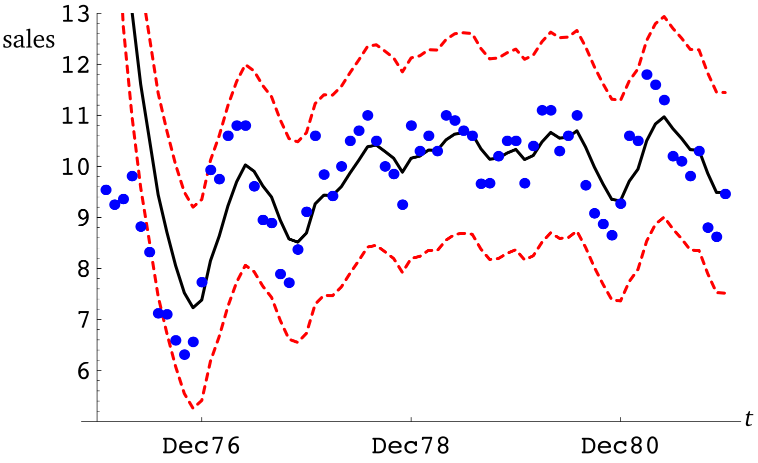

Figure 2.5. *One-step ahead forecast values (solid line) and the 90%* HPD *intervals (dotted lines) for the candy sales data (points) with $m_0 = 50$ and $\delta = 0.7$.*

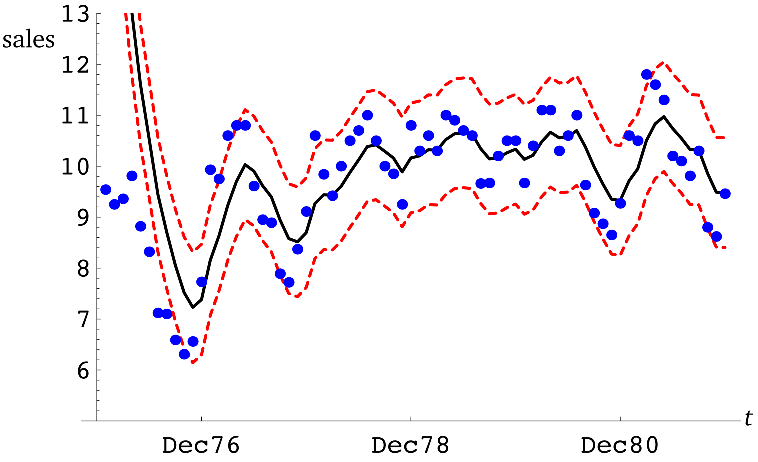

Figure 2.6. *Estimated trend values (solid line) and the 90%* HPD *intervals (dotted lines) for the candy sales data (points) with $m_0 = 50$ and $\delta = 0.7$.*

ple 3), leading to more erratic level estimates and higher variance for the level. ◁

2.5 Forecasting

Given $y^t = \{y_1, ..., y_t\}$ we might be interested in predicting future values of observations Y_{t+k}, or of states, θ_{t+k}, for $k \geq 1$. It follows naturally from the dependence structure in the state space model that prediction is obtained by recursively computing 1-step ahead forecasts. This is formulated in proposition 3 below:

Proposition 3: (Forecasting in a state space model)
For a general state space model defined by assumptions (A1) & (A2) the following is true for any $k > 0$:

(i) The k-step ahead forecast distribution for states can be computed as

$$\pi(\theta_{t+k} \mid y^t) = \int \pi(\theta_{t+k} \mid \theta_{t+k-1}) \pi(\theta_{t+k-1} \mid y^t) d\theta_{t+k-1}$$

(ii) and then, for the observations, as

$$\pi(y_{t+k} \mid y^t) = \int \pi(y_{t+k} \mid \theta_{t+k}) \pi(\theta_{t+k} \mid y^t) d\theta_{t+k}$$

Proof: see lecture.

All densities involved are Gaussian for the case of the Gaussian DLM and are fully specified by means and variances as given in proposition 4 below: **Proposition 4: (Forecasting**

in a Gaussian DLM)
For the Gaussian DLM:

(i) The k-step ahead forecast distribution for states $\pi(\theta_{t+k} \mid y^t)$ is Gaussian with parameters $a_t(k), R_t(k)$ with

$$\begin{aligned} \mathrm{E}[\theta_{t+k} \mid y^t] &= a_t(k) = G_{t+k} a_t(k-1) \\ \mathrm{Var}[\theta_{t+k} \mid y^t] &= R_t(k) = G_{t+k} R_t(k-1) G'_{t+k} + W_{t+k} \end{aligned}$$

with $a_t(0) = m(t)$ and $R_t(0) = C(t)$, *i.e.* from filtering distribution.

(ii) The k-step ahead forecast distribution for the observations, $\pi(y_{t+k} \mid y^t)$ is Gaussian with parameters $f_t(k), Q_t(k)$

$$\begin{aligned} \mathrm{E}[Y_{t+k} \mid y^t] &= f_t(k) = F_{t+k} a_t(k) \\ \mathrm{Var}[Y_{t+k} \mid y^t] &= Q_t(k) = F_{t+k} R_t(k) F'_{t+k} + V_{t+k} \end{aligned}$$

Proof outline: We know that the result holds for $k = 1$ (from proposition 2). The rest follows by mathematical induction.

2.5.1 Forecast function

The mean of the k-step ahead forecast distribution taken as a function of k is called *forecast function*. For a given value of the state vector θ_t at an arbitrary origin time t, the expected development of the series into the future up to $k \geq 1$ steps ahead is

$$\mathrm{E}[Y_{t+k} \mid y^t] = F_{t+k} G_{t+k} G_{t+k-1} \dots G_{t+1} \theta_t.$$

Hence, in general, the *forecast function* in a DLM is

$$f_t(k) = \mathrm{E}[Y_{t+k} \mid D_t] = F_{t+k} G_{t+k} G_{t+k-1} \dots G_{t+1} \mathrm{E}[\theta_t \mid D_t] \tag{2.2}$$

where, in a closed system, $\mathrm{E}[\theta_t \mid D_t]$ corresponds to the mean of the filtering distribution.

The corresponding variance of the forecast function $\mathrm{Var}[Y_{t+k} \mid D_t]$ is usually computed alongside using the recursions in steps (i) and (ii).

Note that the data only enter the predictive distributions through the filtering distribution at the time of the last observation.

in a Time series DLM, since F and G are constant, we have for a given state θ_t

$$\mathrm{E}[\theta_{t+k} \mid \theta_t] = G^k \theta_t$$

and hence the forecast function in a TSDLM simplifies to

$$f_t(k) = \mathrm{E}[Y_{t+k} \mid D_t] = F G^k \mathrm{E}[\theta_t \mid D_t] \tag{2.3}$$

2.6 Innovation Process and Model Checking

Recall that for the DLM we can compute the 1-step ahead forecasts $f_t = \mathrm{E}[Y_t \mid y^{t-1}]$ and the forecast error is defined as

$$e_t = Y_t - \mathrm{E}[Y_t \mid y^{t-1}] = Y_t - f_t,$$

which can also be expressed in terms of the 1-step ahead estimation errors as

$$e_t = Y_t - F_t a_t = F_t(\theta_t - a_t) + v_t.$$

The sequence $(e_t)_{t \geq 1}$ is called the *innovation process*. This terminology is linked to the fact that we can view Y_t as being the sum of predictable f_t and an innovative term e_t. The properties of $(e_t)_{t \geq 1}$ are stated in the following proposition

Proposition 5: Properties of Innovation process

Let $(e_t)_{t\geq 1}$ be the sequence of forecast errors in a Gaussian DLM. Then $(e_t)_{t\geq 1}$ has the following important properties

(i) $\mathrm{E}[e_t] = 0$;

(ii) e_t is uncorrelated with e_s for any $s < t$;

(iii) $(e_t)_{t\geq 1}$ is a Gaussian process.

Proof: see lecture When the observations are univariate then the standardized innovations $e_t^* = e_t/\sqrt{Q_t}$ behave like a random sample from an iid standard normal distribution. In practice this property can be exploited for model checking and various tests, all readily available in R. The most useful procedures are to check for normality via a QQ-plot and to check for uncorrelatedness by inspecting the empirical autocorrelation function, namely a plot of the estimated correlation between e_t and $e_{t-\tau}$ against $\tau = 0, 1, 2, ...$ One should also plot the sequence of standardized innovations to check for outliers or any other patterns.

2.7 Unknown Parameters

So far we have assumed that F_t, G_t, V_t, W_t are known which allowed us to more easily study the basic properties of the DLM. This is not always the case in practice. Let us suppose that the model matrices depend on a vector of unknown parameters ψ which are assumed to be constant over time. We start by giving both the general frequentist and Bayesian approach to dealing with this.

2.7.1 Maximum Likelihood estimation for a DLM

In the classical frequentist framework ψ is estimated by the method of maximum likelihood (ML). The usual approach is to first find the ML estimator (MLE) $\hat{\psi}$ which is then plugged into the filtering and forecasting equations. To set up the data likelihood for the DLM we use the form

$$\pi(y_1, ..., y_n \mid \psi) = \prod_{t=1}^{n} \pi(y_t \mid y^{t-1}, \psi).$$

We know that for the Gaussian DLM $\pi(y_t \mid y^{t-1}, \psi)$ are Gaussian with mean f_t and variance Q_t (see proposition 2). Hence our log likelihood function is

$$\log L(\psi \mid y) = -\frac{1}{2}\sum_{t=1}^{n} \log |Q_t| - \frac{1}{2}\sum_{t=1}^{n} (y_t - f_t)' Q_t^{-1} (y_t - f_t)$$

which can be maximized to obtain the MLE

$$\hat{\psi} = \text{argmax}_{\psi} \log L(\psi \mid y)$$

- Note that Q_t and e_t are computed as part of the Kalman filter in proposition 2 and hence ML estimation is also based on the Kalman filtering methodology.
- $\hat{\psi}$ is asymptotically normally distributed with mean ψ. The variance of the MLE is obtained by computing the inverse Hessian of the log likelihood function evaluated at the MLE. This follows from the general asymptotic properties of MLEs.
- In practice one can use numerical optimization routines (such as 'optim' in R). Moreover, R provides a function 'dlmMLE' as part of the R package 'dlm'. This package also provides an environment for Bayesian inference.
- Once the MLE $\hat{\psi}$ has been obtained, it is usually plugged into the filtering and forecasting equations provided by the Kalman methodology. As the latter is based on given known matrices the uncertainty about estimating parameters will NOT be taken into account.

2.7.2 Bayesian Inference

The Bayesian methodology offers a more consistent formulation as it does properly take into account uncertainty about ψ. The price for this comes in terms of computational effort. Let prior knowledge about ψ be specified by $\pi(\psi)$ and assume that assumptions (A1) & (A2) hold conditionally on ψ. Then, for $t \geq 1$ we assume that (compare with equation (2.1)

$$\pi(\theta^t, y^t, \psi) \sim \pi(\theta_0 \mid \psi)\pi(\psi) \prod_{j=1}^{t} \pi\left(\theta_j \mid \theta_{j-1}, \psi\right) \pi\left(y_j \mid \theta_j, \psi\right) \tag{2.4}$$

Given data y^t inference on the unknown state at time s, θ_s, and on ψ is done by computing

$$\pi(\theta_s, \psi \mid y^t) = \pi(\theta_s \mid \psi, y^t)\pi(\psi \mid y^t)$$

where, as before, we distinguish between the cases $s = t$ (filtering), $s > t$ (prediction) and $s < t$ (smoothing). In particular, the filtering density is given by the integral

$$\pi(\theta_t \mid y^t) = \int \pi(\theta_t \mid \psi, y^t)\pi(\psi \mid y^t) d\psi.$$

Hence, the recursions introduced in proposition 2 can also be used but need to be averaged with respect to the posterior of ψ, $\pi(\psi \mid y^t)$. While the Kalman filter provides a closed form methodology for recursively updating the filtering density when the parameters are known, such closed forms do rarely exist for the case that ψ is unknown, and we will consider a special case where a conjugate Bayesian analysis is possible below. In other cases one can use stochastic simulation methods, in particular Markov chain Monte Carlo (MCMC) and sequential Monte Carlo techniques (SMC). The latter are particularly focusing on efficient online updating through sequential analysis.

2.7.3 Conjugate Bayesian Analysis in the DLM with unknown variance

Consider the Gaussian DLM with known system matrices F_t and G_t. Suppose V_t and W_t are unknown but have a known common scaling factor σ^2 so we can write

$$V_t = \sigma^2 \tilde{V}_t, \quad W_t = \sigma^2 \tilde{W}_t \quad \text{with } \sigma^2 \text{ unknown.}$$

Also $C_0 = \sigma^2 \tilde{C}_0$. Hence $\tilde{V}_t, \tilde{W}_t, \tilde{C}_t$ are known along with F_t and G_t. Let, for convenience, $\phi = 1/\sigma^2$. Then our DLM is

$$\begin{aligned} Y_t \mid \theta_t, \phi &\sim \mathrm{N}\left(y_t \mid F_t\theta_t, \phi^{-1}\tilde{V}_t\right) \\ \theta_t \mid \theta_{t-1}, \phi &\sim \mathrm{N}\left(\theta_t \mid G_t\theta_{t-1}, \phi^{-1}\tilde{V}_t\right) \end{aligned}$$

As prior for (ϕ, θ_0) one can find that a conjugate choice is the Normal-Gamma prior where

$$\begin{aligned} \phi &\sim \mathrm{Ga}\left(\phi \mid \alpha_0, \beta_0\right) \\ \theta_0 \mid \phi &\sim \mathrm{N}\left(\theta_0 \mid m_0, \phi^{-1}\tilde{C}_0\right) \end{aligned}$$

We shall write $(\phi, \theta_0) \sim NG(m_0, \tilde{C}_0, \alpha_0, \beta_0)$ for short. For this model we have the following updating recursions:

Proposition 6: (Conjugate inference, unknown variance matrices)
Consider the Gaussian DLM described above, if

$$(\theta_{t-1}, \phi) \mid y^{t-1} \sim NG(m_{t-1}, \tilde{C}_{t-1}, \alpha_{t-1}, \beta_{t-1}), \quad t \geq 1,$$

then

(i) The 1-step-ahead predictive density of $(\theta_t, \phi) \mid y^{t-1}$ is $NG(a_t, \tilde{R}_t, \alpha_{t-1}, \beta_{t-1})$ where

$$\begin{aligned} a_t &= G_t m_{t-1} \\ \tilde{R}_t &= G_t \tilde{C}_{t-1} G_t' + \tilde{W}_t \end{aligned}$$

(ii) The 1-step-ahead predictive density of $y_t \mid y^{t-1}$ is Student-t with parameters $(f_t, \tilde{Q}_t \beta_{t-1}/\alpha_{t-1}, 2\alpha_{t-1})$, where

$$\begin{aligned} f_t &= F_t a_t \\ \tilde{Q}_t &= F_t \tilde{R}_t F_t' + \tilde{V}_t \end{aligned}$$

(iii) The filtering density of $(\theta_t, \phi) \mid y^t$ is $NG(m_t, \tilde{C}_t, \alpha_t, \beta_t)$, where

$$\begin{aligned} m_t &= a_t + \tilde{R}_t F_t' \tilde{Q}_t^{-1}(y_t - f_t) \\ \tilde{C}_t &= \tilde{R}_t - \tilde{R}_t F_t' \tilde{Q}_t^{-1} F_t \tilde{R}_t \\ \alpha_t &= \alpha_{t-1} + \frac{m_t}{2} \\ \beta_t &= \beta_{t-1} + \frac{1}{2}(y_t - f_t)' \tilde{Q}_t^{-1}(y_t - f_t). \end{aligned}$$

Proof: See lecture for an outline of proof.

Compare this with the Normal-Gamma model in the univariate case treated in Chapter 1.

The result of step (iii) applies in exactly the same way to Bayesian linear regression with unknown variance using Normal-Gamma prior for regression coefficients and variance.

By the properties of the Student-t distribution we have that the 1-step ahead forecast of Y_t given y^{t-1} is $\mathrm{E}[Y_t \mid y^{t-1}] = f_t$ with variance matrix $\mathrm{Var}[Y_t \mid y^{t-1}] = \tilde{Q}_t \beta_{t-1}/(\alpha_{t-1} - 1)$.

From (iii) it follows that the marginal filtering density of $\sigma^2 = \phi^{-1}$ given y^t is inverse-Gamma with parameters (α_t, β_t) so that, for $\alpha_t > 2$,

$$\mathrm{E}[\sigma^2 \mid y^t] = \frac{\beta_t}{\alpha_t - 1}, \quad \mathrm{Var}[\sigma^2 \mid y^t] = \frac{\beta_t^2}{(\alpha_t - 1)^2(\alpha_t - 2)}.$$