Chapter 2

Dynamic linear model

In the previous chapter, we have seen how to use Bayesian inference to learn about an unknown quantity of interest from a sequence of conditionally-independent observations. Yet, in practical situations, it is common that the quantity of interest evolves in time according to a model which can itself contain random effects. The objective in this chapter is to model these aspects in a general way and to solve the corresponding Bayesian inference problem in the Gaussian case.

2.1 State space model

Instead of a single random variable $\boldsymbol{\theta}$, we now consider a sequence of random variables $(\boldsymbol{\theta}_k)_{k\geq 0}$ in $\Theta \subseteq \mathbb{R}^d$ for some $d > 0$. The integer k is interpreted as a time index. The time step Δ between two consecutive time indices, expressed for instance in seconds, is assumed to be constant. At every time index $k \geq 0$, an observation $y_k \in \mathcal{Y} \subseteq \mathbb{R}^{d'}$, $d' > 0$, is received and we assume that the underlying random variable $\boldsymbol{y}_k$ is conditionally-independent of $\boldsymbol{y}_l$ given $\boldsymbol{\theta}_k$ for any integer $l \neq k$.

2.1.1 Markov chain and state space model

The general objective is to determine the posterior distribution of $\boldsymbol{\theta}_k$ or $\boldsymbol{\theta}_{0:k}$ given some observations $y_0, \ldots, y_k$. Considering the assumption on the observations, it would not be meaningful to assume that $\boldsymbol{\theta}_k$ is independent of $\boldsymbol{\theta}_l$ for any $l \neq k$ since the problem would collapse to a collection of static Bayesian inference problems in this case. Instead, we consider one of the simplest form of dependence as defined below.

Definition 2.1 (Markov chain)**.** The sequence of random variables $(\boldsymbol{\theta}_k)_{k\geq 0}$ forms a (discrete-time) *Markov chain* if

$$p_{\boldsymbol{\theta}_k|\boldsymbol{\theta}_{0:k-1}}(\theta \,|\, \theta_0, \ldots, \theta_{k-1}) = p_{\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}}(\theta \,|\, \theta_{k-1}) \tag{2.1}$$

for any $\theta_0, \ldots, \theta_{k-1} \in \Theta$ and any $k > 0$.

Equation (2.1) is often referred to as the *Markov property*. We therefore assume that $(\boldsymbol{\theta}_k)_{k\geq 0}$ is a Markov chain and we refer to $(\boldsymbol{\theta}_k, \boldsymbol{y}_k)_{k\geq 0}$ as a *state space model* or a *hidden Markov model*. It is common to describe the evolution of the quantity of interest and the generation of observations via a system of equations of the form

$$\begin{aligned} \boldsymbol{\theta}_k &= f_k(\boldsymbol{\theta}_{k-1}, \boldsymbol{u}_k) \\ \boldsymbol{y}_k &= h_k(\boldsymbol{\theta}_k, \boldsymbol{v}_k) \end{aligned}$$

with f_k and h_k the state and observation functions respectively and with $(\boldsymbol{u}_k)_{k>0}$ and $(\boldsymbol{v}_k)_{k\geq 0}$ sequences of zero-mean independent random variables.

2.1.2 Smoothing and filtering distributions

We introduce the following notations for the sake of simplicity

$$\pi_k(\theta) = p_{\theta_k}(\theta), \qquad q_k(\theta \,|\, \theta') = p_{\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}}(\theta \,|\, \theta'), \qquad \ell_k(y \,|\, \theta) = p_{\boldsymbol{y}_k|\boldsymbol{\theta}_k}(y \,|\, \theta).$$

With these notations and the considered assumptions, the joint distribution of $\boldsymbol{\theta}_0, \dots, \boldsymbol{\theta}_n$ given the observations $y_0, \dots, y_n$, called the *smoothing* distribution at time step n, can be simply expressed as

$$p_{\boldsymbol{\theta}_{0:n}|\boldsymbol{y}_{0:n}}(\theta_0, \dots, \theta_n \,|\, y_0, \dots, y_n) \propto \pi_0(\theta_0)\ell_0(y_0 \,|\, \theta_0) \prod_{k=1}^{n} q_k(\theta_k \,|\, \theta_{k-1})\ell_k(y_k \,|\, \theta_k).$$

This expression can be easily verified by induction. It is generally difficult to determine the smoothing distribution since the dimension of the problem can be arbitrarily large. Also, in many cases, the interest is mainly in the distribution of the quantity of interest at the current time step, that is in

$$p_{\boldsymbol{\theta}_n|\boldsymbol{y}_{0:n}}(\theta_n \,|\, y_0, \dots, y_n) = \int p_{\boldsymbol{\theta}_{0:n}|\boldsymbol{y}_{0:n}}(\theta_0, \dots, \theta_n \,|\, y_0, \dots, y_n) \,\mathrm{d}\theta_{0:n-1}. \tag{2.3}$$

Given the importance of this distribution, usually referred to as the *filtering* distribution, we also introduce a special notation for it, that is

$$\pi_{k|n}(\theta \,|\, y_0, \dots, y_n) = p_{\boldsymbol{\theta}_k|\boldsymbol{y}_{0:n}}(\theta \,|\, y_0, \dots, y_n), \qquad k, n \geq 0.$$

When looking at (2.3), it seems even more difficult to determine the filtering distribution when compared to the smoothing distribution since it requires an extra marginalisation step; however, it is easy to show that the filtering distribution can instead be computed recursively: assuming that the filtering distribution at time step $n-1$ is available, we can deduce an expression of the filtering distribution at time step n as follows

(predictive state distribution) the distribution of $\boldsymbol{\theta}_n$ given observations up to time step $n-1$ is

$$\pi_{n|n-1}(\theta \,|\, y_0, \dots, y_{n-1}) = \int q_n(\theta \,|\, \theta')\pi_{n-1|n-1}(\theta' \,|\, y_0, \dots, y_{n-1}) \,\mathrm{d}\theta'$$

(posterior distribution) the distribution of $\boldsymbol{\theta}_n$ given observations up to time step n is

$$\pi_{n|n}(\theta \,|\, y_0, \dots, y_n) = \frac{\ell_n(y_n \,|\, \theta)\pi_{n|n-1}(\theta \,|\, y_0, \dots, y_{n-1})}{\int \ell_n(y_n \,|\, \theta')\pi_{n|n-1}(\theta' \,|\, y_0, \dots, y_{n-1}) \,\mathrm{d}\theta'} \tag{2.4}$$

These equations are fundamental to address the (discrete-time) *filtering problem*. Note that the denominator in the right-hand side of (2.4) is the predictive distribution $p_{\boldsymbol{y}_n|\boldsymbol{y}_{0:n-1}}(\cdot \,|\, y_0, \dots, y_{n-1})$ of the observation, or *forecasting distribution*, at time step n evaluated at y_n.

2.2 Dynamic linear model

The filtering problem can be significantly simplified by considering the case where the state and observation functions take the form

$$f_k(\theta, u) = F_k\theta + u \qquad \text{and} \qquad h_k(\theta, v) = H_k\theta + v$$

where F_k is a $d \times d$ matrix and H_k is a $d' \times d$ matrix, respectively called the *transition* matrix and the *observation* matrix. The state and observation equations become

$$\begin{aligned} \boldsymbol{\theta}_k &= F_k\boldsymbol{\theta}_{k-1} + \boldsymbol{u}_k \\ \boldsymbol{y}_k &= H_k\boldsymbol{\theta}_k + \boldsymbol{v}_k \end{aligned}$$

and this case will be referred to as a *dynamic linear model* (DLM). We will often consider the case where $F_k = F$, $H_k = H$, $p_{\boldsymbol{u}_k}(\cdot) = p_{\boldsymbol{u}}(\cdot)$ and $p_{\boldsymbol{v}_k}(\cdot) = p_{\boldsymbol{v}}(\cdot)$ for any $k \geq 0$, which will be called a *constant DLM*.

An example of DLM is given in Figure 2.1 for a state of the form $\boldsymbol{\theta}_k = (\boldsymbol{x}_k, \dot{\boldsymbol{x}}_k)^\intercal$ with $\boldsymbol{x}_k$ and $\dot{\boldsymbol{x}}_k$ the position and velocity of an object at time step k. Only the position is observed at each time step, so that the velocity is *hidden*. The specific model considered here will be detailed in Section 2.2.3.

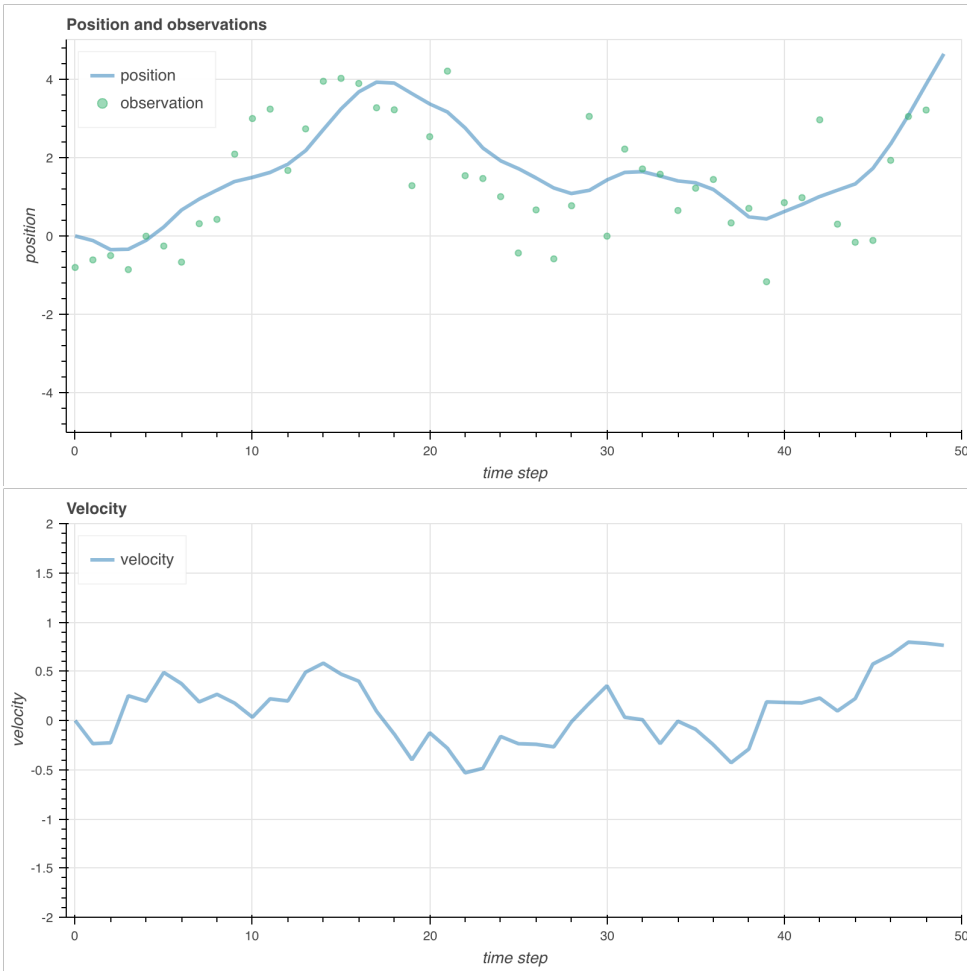


Figure 2.1: The two components and observations of a 2-dimensional DLM.

2.2.1 Mean and variance of the predictive distributions

Since the transition noise $\boldsymbol{u}_k$ and the observation noise $\boldsymbol{v}_k$ are assumed to have a mean equal to zero, it is easy to compute the mean and variance corresponding to the predictive state distribution at time step k as

$$\begin{aligned} m_k &\doteq \mathbb{E}(\boldsymbol{\theta}_k \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) \\ &= \mathbb{E}(F_k \boldsymbol{\theta}_{k-1} + \boldsymbol{u}_k \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) \\ &= F_k \mathbb{E}(\boldsymbol{\theta}_{k-1} \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) \\ &= F_k \hat{m}_{k-1} \end{aligned}$$

with $\hat{m}_{k-1} \doteq \mathbb{E}(\boldsymbol{\theta}_{k-1} \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1})$ and

$$\begin{aligned} P_k &\doteq \operatorname{var}(\boldsymbol{\theta}_k \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) \\ &= \mathbb{E}((\boldsymbol{\theta}_k - m_k)(\boldsymbol{\theta}_k - m_k)^\intercal \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) \\ &= \mathbb{E}\big((F_k \boldsymbol{\theta}_{k-1} - F_k \hat{m}_{k-1})(F_k \boldsymbol{\theta}_{k-1} - F_k \hat{m}_{k-1})^\intercal \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}\big) + \operatorname{var}(\boldsymbol{u}_k) \\ &= F_k \operatorname{var}(\boldsymbol{\theta}_{k-1} \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) F_k^\intercal + \operatorname{var}(\boldsymbol{u}_k) \\ &= F_k \hat{P}_{k-1} F_k^\intercal + \operatorname{var}(\boldsymbol{u}_k), \end{aligned}$$

with $\hat{P}_{k-1} \doteq \operatorname{var}(\boldsymbol{\theta}_{k-1} \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1})$. Following the same approach, we can also compute the mean and variance of the forecasting distribution $p_{\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1}}(\cdot \,|\, \cdot)$ as

$$\begin{aligned} \mathbb{E}(\boldsymbol{y}_k \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) &= H_k m_k \\ \operatorname{var}(\boldsymbol{y}_k \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) &= H_k P_k H_k^\intercal + \operatorname{var}(\boldsymbol{v}_k). \end{aligned}$$

These calculations reveal that DLMs are well-suited to Gaussian assumptions, as will be demonstrated in the next section.

Note that the mean and variance of the state and observation can be predicted several steps ahead using the same approach. For instance

$$\mathbb{E}(\boldsymbol{\theta}_n \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) = F_n F_{n-1} \dots F_k \mathbb{E}(\boldsymbol{\theta}_{k-1} \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1})$$

for any $n \geq k$, and

$$\mathbb{E}(\boldsymbol{y}_n \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) = H_n F_n F_{n-1} \dots F_k \mathbb{E}(\boldsymbol{\theta}_{k-1} \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}).$$

The corresponding expressions for the variance are equally easy to derive but take a more complex form.

2.2.2 Kalman filter

In many applications, the distribution of the transition noise $\boldsymbol{u}_k$ and of the observation noise $\boldsymbol{v}_k$ are adequately modelled as (multivariate) Gaussian, that is

$$p_{\boldsymbol{u}_k}(u) = \mathrm{N}(u; 0, U_k) \qquad \text{and} \qquad p_{\boldsymbol{v}_k}(v) = \mathrm{N}(v; 0, V_k)$$

for some (positive-definite) matrices U_k and V_k and for any $k \geq 0$. If the initial distribution $\pi_0(\cdot)$ is also Gaussian with mean m_0 and variance P_0, then the corresponding model is referred to as a *Gaussian DLM*.

Theorem 2.1 (Kalman filter)**.** *Consider a Gaussian DLM, the predictive distribution at any time step $k > 0$ is also Gaussian and verifies $\pi_{k|k-1}(\theta \,|\, y_0, \dots, y_{k-1}) = \mathrm{N}(\theta; m_k, P_k)$ with*

$$\begin{cases} m_k = F_k \hat{m}_{k-1} \\ P_k = F_k \hat{P}_{k-1} F_k^\intercal + U_k, \end{cases} \qquad \text{(prediction)}$$

and the posterior distribution at any time step $k \geq 0$ is also Gaussian and verifies $\pi_{k|k}(\theta \,|\, y_0, \dots, y_k) = \mathrm{N}(\theta; \hat{m}_k, \hat{P}_k)$ with

$$\begin{cases} \hat{m}_k = m_k + K_k z_k \\ \hat{P}_k = (I_d - K_k H_k) P_k \end{cases} \qquad \text{(update)}$$

where I_d is the identity matrix of size d and where

$$z_k = y_k - H_k m_k \qquad \text{(innovation)}$$
$$K_k = P_k H_k^\intercal S_k^{-1} \qquad \text{(optimal Kalman gain)}$$
$$S_k = H_k P_k H_k^\intercal + V_k \qquad \text{(covariance of the innovation)}$$

The Kalman filter defines a recursive algorithm to compute the mean and variance of the predictive and filtering distributions at different time steps. We have already proved the formula for the mean and variance of the predictive distribution in Section 2.2.1, however, the corresponding proof for the posterior distribution is more complex and will be dealt with later.

In R, the Kalman filter is available as the function `dlmFilter` in the package `dlm`.

2.2.3 A concrete example

We consider the scenario where the state of an object moving on a line must be inferred from the noisy observations given by a sensor. The random variable $\boldsymbol{\theta}_k$ models the state of the object at time step k, which is assumed to be of the form $\boldsymbol{\theta}_k = (\boldsymbol{x}_k, \dot{\boldsymbol{x}}_k)^\intercal \in \Theta = \mathbb{R}^2$, with $\boldsymbol{x}_k$ the position of the object and $\dot{\boldsymbol{x}}_k$ its velocity. We consider a constant Gaussian DLM (that is $U_k = U$ and $V_k = V$ for any $k \geq 0$) and assume that the object moves according to a nearly-constant velocity model described by

$$F = \begin{pmatrix} 1 & \Delta \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad U = \sigma'^2 \begin{pmatrix} \Delta^4/4 & \Delta^3/2 \\ \Delta^3/2 & \Delta^2 \end{pmatrix}$$

for some $\sigma' > 0$, where Δ is the time step. Additionally, we assume that the sensor simply observes the position, that is

$$H = \begin{pmatrix} 1 & 0 \end{pmatrix} \qquad \text{and} \qquad V = \sigma^2$$

for some $\sigma > 0$. An example of the application of this model on simulated data is shown in Figure 2.2.

2.2.4 Proof of the Kalman-filter update

There exist several ways to prove the update formula of the Kalman filter. The most obvious one is to make the explicit calculations of the multiplication and integration between the Gaussian probability distributions in Bayes' rule. Although this is doable by using some usual algebraic manipulations, it is more interesting to use statistical arguments to recover the desired result.

For this purpose, and dropping the time steps in the notation where there is no ambiguity, we will first assume that our updated estimate $\hat{\boldsymbol{\theta}}$ is a weighted average between the predicted estimate $\breve{\boldsymbol{\theta}} \doteq \mathbb{E}(\boldsymbol{\theta}_k \,|\, \boldsymbol{y}_{0:k-1})$ and the observation $\boldsymbol{y}_k$, that is

$$\hat{\boldsymbol{\theta}} \doteq K' \breve{\boldsymbol{\theta}} + K \boldsymbol{y}_k$$

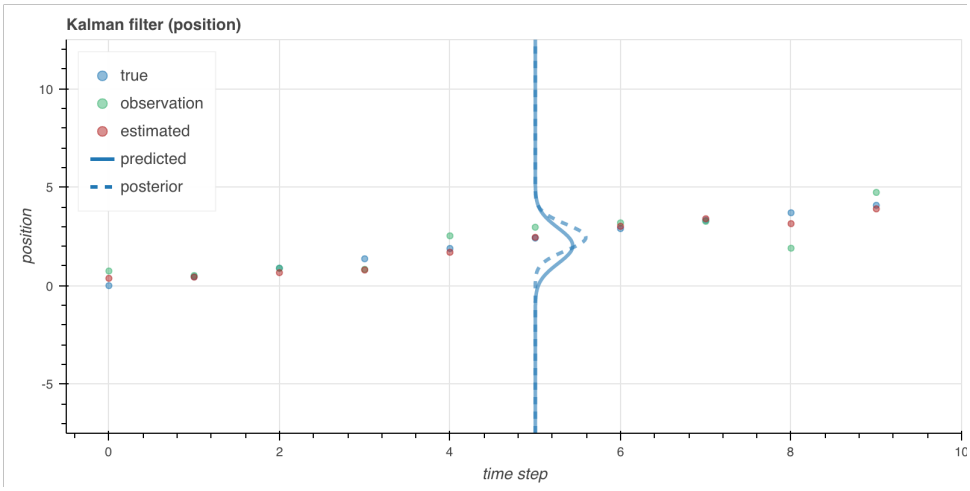
with K a $d \times d'$ matrix and K' a $d \times d$ matrix to be determined.

Unbiased estimator It is easy to verify that the prediction step preserves unbiasedness, so assuming that our predicted estimate $\breve{\boldsymbol{\theta}}$ is unbiased, that is $\mathbb{E}(\breve{\boldsymbol{\theta}}) = \mathbb{E}(\boldsymbol{\theta}_k)$ holds, we want to make sure that $\hat{\boldsymbol{\theta}}$ is also unbiased:

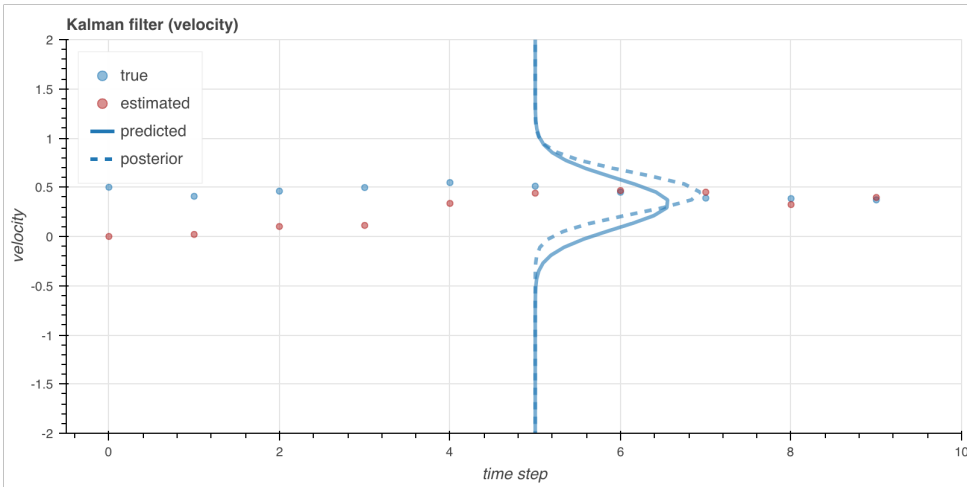
$$\begin{aligned} \mathbb{E}(\hat{\boldsymbol{\theta}}) &= \mathbb{E}(K' \breve{\boldsymbol{\theta}} + K \boldsymbol{y}_k) \\ &= \mathbb{E}(K' \breve{\boldsymbol{\theta}} + KH \boldsymbol{\theta}_k + K \boldsymbol{v}_k) \\ &= K' \mathbb{E}(\breve{\boldsymbol{\theta}}) + KH \mathbb{E}(\breve{\boldsymbol{\theta}}) \\ &= (K' + KH) \mathbb{E}(\breve{\boldsymbol{\theta}}) \end{aligned}$$

which indicates that we need to choose $K' = I_d - KH$ to ensure that the update step also preserves unbiasedness. This yields

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (I_d - KH) \breve{\boldsymbol{\theta}} + K \boldsymbol{y}_k \\ &= \breve{\boldsymbol{\theta}} + K(\boldsymbol{y}_k - H \breve{\boldsymbol{\theta}}). \end{aligned}$$



(a) Observations, true and estimated position as well as predicted and posterior distribution at a given time step.



(b) True and estimated velocity as well as predicted and posterior distribution at a given time step.

Figure 2.2: Kalman filtering results for a nearly-constant velocity model.

We can therefore verify the update equation for the mean by taking the expectation $\mathbb{E}(\cdot \mid \boldsymbol{y}_{0:k} = y_{0:k})$. A formula for the covariance can also be deduced as follows

$$\operatorname{var}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_k \mid \boldsymbol{y}_{0:k}) = (I_d - KH) \operatorname{var}(\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}_k \mid \boldsymbol{y}_{0:k-1})(I_d - KH)^\intercal + K \operatorname{var}(\boldsymbol{v}_k) K^\intercal.$$

This does not lead directly to the update equation for the covariance and yields instead

$$\hat{P} = (I_d - KH)P(I_d - KH)^\intercal + KVK^\intercal. \tag{2.12}$$

This formula is actually valid for any gain K, including the above-defined optimal Kalman gain. It remains to show in which sense this gain is optimal.

Optimal Kalman gain We want our estimator $\hat{\boldsymbol{\theta}}$ to lead to a mean-square error that is as small as possible and K is the only design matrix left in (2.12). Thus, we choose as a gain

$$K^* = \operatorname*{argmin}_K \mathbb{E}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_k\|^2 \mid \boldsymbol{y}_{0:k})$$

where $\|\cdot\|$ is the Euclidean norm. It holds that

$$\mathbb{E}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_k\|^2 \mid \boldsymbol{y}_{0:k}) = \operatorname{tr} \operatorname{var}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_k \mid \boldsymbol{y}_{0:k})$$

with tr the trace operator, so we can minimise the trace of the variance instead. A useful result when optimising a trace is

$$\frac{\partial}{\partial A} \operatorname{tr}(ABA^\intercal) = 2AB$$

which holds for any matrix A and any symmetric matrix B of appropriate dimensions. To find the minimum of the trace of $\hat{P}$, we differentiate the trace of its expression given in (2.12) with respect to K as follows

$$\begin{aligned}\frac{\partial}{\partial K} \operatorname{tr}(\hat{P}) &= \frac{\partial}{\partial K} \operatorname{tr}((I_d - KH)P(I_d - KH)^\intercal) + \frac{\partial}{\partial K} \operatorname{tr}(KVK^\intercal) \\ &= -2(I_d - KH)PH^\intercal + 2KV,\end{aligned}$$

so that the minimum is found at

$$K = PH^\intercal(HPH^\intercal + V)^{-1},$$

which is indeed the expression of the Kalman gain in Theorem 2.1.

2.2.5 Model checking

When dealing with real data, one way to verify that the considered model is appropriate is to look at the distribution of the innovation. The innovation z_k introduced in Theorem 2.1 is not random because the observation y_k at time step k has been assumed fixed. In general, if we consider the random variable $\boldsymbol{y}_k$ instead, the innovation also becomes a random variable $\boldsymbol{z}_k$ defined as

$$\boldsymbol{z}_k = \boldsymbol{y}_k - \mathbb{E}(\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1}) = \boldsymbol{y}_k - H_k \breve{\boldsymbol{\theta}}_k.$$

The conditional distribution of $\boldsymbol{z}_k$ given $\boldsymbol{y}_{0:k-1} = y_{0:k-1}$ follows easily from the one of $\boldsymbol{y}_k$ as $\mathrm{N}(\cdot; 0, S_k)$. It can also be proved that the *innovation process* $(\boldsymbol{z}_k)_{k \geq 0}$ is uncorrelated, that is $\operatorname{cov}(\boldsymbol{z}_{k-\delta}, \boldsymbol{z}_k) = 0$ for any $\delta > 0$.

One can therefore verify the suitability of the model by checking that the innovation $\boldsymbol{z}_k$ is indeed distributed according to $\mathrm{N}(\cdot; 0, S_k)$ using, e.g., Q-Q plots, and that the innovation process is uncorrelated using, e.g., autocorrelation functions.

2.3 Unknown parameters

It is often the case in practice that at least some of the parameters of the model are now fully known in advance and must be estimated in parallel. The most usual situation is where U and/or V have a known form but have unknown coefficients like σ or σ' in Section 2.2.3. More generally, we denote ψ the unknown parameter of the model in a given space Ψ, which could affect any or all of the matrices F_k, H_k, U_k or V_k. We consider two ways of solving this parameter estimation problem: the maximum likelihood estimation (MLE) approach and the Bayesian approach.

2.3.1 Maximum likelihood estimation

Consider the likelihood $L_n(\psi; y_0, \dots, y_n) \doteq p^{\psi}_{\boldsymbol{y}_{0:n}}(y_0, \dots, y_n)$ at the value ψ of the unknown parameter given the observations $y_0, \dots, y_n$. In the context of MLE, the logarithm of the likelihood is usually considered as

$$\begin{aligned}\log L_n(\psi; y_0, \dots, y_n) &= \log p^{\psi}_{\boldsymbol{y}_{0:n}}(y_0, \dots, y_n) \\ &= \log p^{\psi}_{\boldsymbol{y}_0}(y_0) + \sum_{k=1}^{n} \log p^{\psi}_{\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1}}(y_k \,|\, y_0, \dots, y_{k-1}).\end{aligned}$$

In the case of a Gaussian DLM, this expression simplifies to

$$\log L_n(\psi; y_0, \dots, y_n) = -\frac{1}{2}\sum_{k=0}^{n} \log |\tilde{P}^{\psi}_k| - \frac{1}{2}\sum_{k=0}^{n} (y_k - \tilde{m}^{\psi}_k)^{\intercal} (\tilde{P}^{\psi}_k)^{-1} (y_k - \tilde{m}^{\psi}_k)$$

with $\tilde{m}^{\psi}_k$ and $\tilde{P}^{\psi}_k$ the mean and variance of the predicted observation at time step k when the unknown parameter is $\psi \in \Psi$. The MLE $\hat{\psi}_n$ at time step n can then be expressed as

$$\hat{\psi}_n = \operatorname*{argmax}_{\psi \in \Psi} \log L_n(\psi; y_0, \dots, y_n).$$

Under various assumptions, the MLE can be shown to be *consistent*, that is, denoting ψ^* the true value of the parameter, it holds that

$$\hat{\psi}_n \xrightarrow{n \to \infty} \psi^*.$$

Note that formally, this is a form of *convergence in probability*. Additionally, it can be proved that the MLE is *asymptotically normal*, that is

$$\sqrt{n}(\hat{\psi}_n - \psi^*) \xrightarrow{n \to \infty} \mathrm{N}(0, I(\psi^*)^{-1})$$

where the convergence is *in distribution* and where $I(\psi)$ is the *Fisher information matrix* such that

$$[I(\psi)]_{i,j} = \lim_{n \to \infty} \frac{1}{n+1} \mathbb{E}^{\psi}\left(\left(\frac{\partial}{\partial \psi_i} \log p^{\psi}_{\boldsymbol{y}_{0:n}}(y_{0:n}) \right) \left(\frac{\partial}{\partial \psi_j} \log p^{\psi}_{\boldsymbol{y}_{0:n}}(y_{0:n}) \right) \right)$$

with ψ defined as the column vector $(\psi_1, \dots, \psi_N)^{\intercal}$. If ψ is a scalar then the Fisher information is also a scalar defined as

$$I(\psi) = \lim_{n \to \infty} \frac{1}{n+1} \mathbb{E}^{\psi}\left(\left(\frac{\partial}{\partial \psi} \log p^{\psi}_{\boldsymbol{y}_{0:n}}(y_{0:n}) \right)^2 \right).$$

One important limitation with the MLE approach is that the uncertainty associated with $\hat{\psi}_n$ is not directly represented and hence, the Kalman filter will use this value of the parameter as the true one.

In order to solve the optimisation problem associated with the MLE, one can use existing numerical optimisation routines such as `optim` in R. There is also a function `dlmMLE` that is more specific for the type of problems presented in this section; this function is also part of the `dlm` package in R.

2.3.2 Bayesian approach

As usual in the context of Bayesian inference, we now consider that the unknown parameter is a random variable $\boldsymbol{\psi}$ and we define some prior distribution $p_{\boldsymbol{\psi}}(\cdot)$ on Ψ in order to model the available information about $\boldsymbol{\psi}$. We assume that the conditional distribution $p_{\boldsymbol{\theta}_k | \boldsymbol{\theta}_{0:k-1}, \boldsymbol{\psi}}(\cdot \,|\, \cdot)$ has the Markov property, that is

$$p_{\boldsymbol{\theta}_k | \boldsymbol{\theta}_{0:k-1}, \boldsymbol{\psi}}(\theta_k \,|\, \theta_0, \dots, \theta_{k-1}, \psi) = p_{\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}, \boldsymbol{\psi}}(\theta_k \,|\, \theta_{k-1}, \psi).$$

In a similar way, we assume that the observation $\boldsymbol{y}_k$ is conditionally independent of $\boldsymbol{y}_l$ given $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ for any $l \neq k$. It follows that the joint distribution of the states, the observations and the parameter is

$$p_{\boldsymbol{\theta}_{0:n}, \boldsymbol{y}_{0:n}, \boldsymbol{\psi}}(\theta_0, \dots, \theta_n, y_0, \dots, y_n, \psi) = p_{\boldsymbol{\psi}}(\psi) \bar{\pi}_0(\theta_0 \,|\, \psi) \bar{\ell}_0(y_0 \,|\, \theta_0, \psi) \prod_{k=1}^{n} \bar{q}_k(\theta_k \,|\, \theta_{k-1}, \psi) \bar{\ell}_k(y_k \,|\, \theta_k, \psi),$$

where $\bar{\pi}_0(\cdot \,|\, \psi)$, $\bar{\ell}_k(\cdot \,|\, \theta_k, \psi)$ and $\bar{q}_k(\cdot \,|\, \theta_{k-1}, \psi)$ stand respectively for the distributions $p_{\boldsymbol{\theta}_0|\boldsymbol{\psi}}(\cdot \,|\, \psi)$, $p_{\boldsymbol{y}_k|\boldsymbol{\theta}_k,\boldsymbol{\psi}}(\cdot \,|\, \theta_k, \psi)$ and $p_{\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1},\boldsymbol{\psi}}(\cdot \,|\, \theta_{k-1}, \psi)$. The filtering distribution can then be recovered by integrating out the parameter as

$$\begin{aligned}\pi_{k|n}(\theta_k \,|\, y_0, \dots, y_n) &= \int p_{\boldsymbol{\theta}_k,\boldsymbol{\psi}|\boldsymbol{y}_{0:n}}(\theta_k, \psi \,|\, y_0, \dots, y_n) \,\mathrm{d}\psi \\ &= \int p_{\boldsymbol{\theta}_k|\boldsymbol{\psi},\boldsymbol{y}_{0:n}}(\theta_k \,|\, \psi, y_0, \dots, y_n) p_{\boldsymbol{\psi}|\boldsymbol{y}_{0:n}}(\psi \,|\, y_0, \dots, y_n) \,\mathrm{d}\psi.\end{aligned}$$

The first term in the integral is given by the Kalman filter for a fixed ψ, however, it is not usually the case that this integral can be computed analytically (although an example where it is will be detailed in the next section). One then has to resort to stochastic simulation algorithms such as Markov chain Monte Carlo (MCMC) or sequential Monte Carlo (SMC).

2.3.3 Extension of the Kalman filter to unknown variance

It is often the case in practice that the variance of the evolution noise and/or of the observation noise are not known a priori and must be learned alongside the parameter of interest. For the sake of simplicity, we assume that the form of the matrices U_k and V_k is known except for a common coefficient that does not depend on time, that is, for any $k \geq 0$,

$$U_k = \sigma^2 U'_k \qquad \text{and} \qquad V_k = \sigma^2 V'_k$$

for some $\sigma > 0$ and some matrices U'_k and V'_k of appropriate dimensions. We also assume that the prior variance is of the form $P_0 = \sigma^2 P'_0$. The corresponding model can then be expressed via the evolution and observation equations

$$\begin{aligned}\boldsymbol{\theta}_k &= F_k \boldsymbol{\theta}_{k-1} + \sigma \boldsymbol{u}'_k \\ \boldsymbol{y}_k &= H_k \boldsymbol{\theta}_k + \sigma \boldsymbol{v}'_k\end{aligned}$$

with $\boldsymbol{u}'_k \sim \mathrm{N}(\cdot; 0, U'_k)$ and $\boldsymbol{v}'_k \sim \mathrm{N}(\cdot; 0, V'_k)$. To represent the fact that σ is unknown, we introduce a random variable $\boldsymbol{\tau}$ corresponding to the precision induced by σ. The prior distribution of $\boldsymbol{\tau}$ is defined as a gamma distribution with parameters α_0 and β_0. To sum up, we have a priori

$$\boldsymbol{\tau} \sim \mathrm{Ga}(\cdot; \alpha_0, \beta_0) \qquad \text{and} \qquad \boldsymbol{\theta}_0 \,|\, \boldsymbol{\tau} \sim \mathrm{N}(\cdot; \mu_0, \boldsymbol{\tau}^{-1} P'_0).$$

We want to show that if the posterior distributions of $\boldsymbol{\tau}$ and $\boldsymbol{\theta} \,|\, \boldsymbol{\tau}$ take this form at time step $k-1$, then they also take this form at time step k. We therefore assume that

$$\boldsymbol{\tau} \,|\, \boldsymbol{y}_{0:k-1} \sim \mathrm{Ga}(\cdot; \alpha_{k-1}, \beta_{k-1}) \qquad \text{and} \qquad \boldsymbol{\theta}_{k-1} \,|\, \boldsymbol{\tau}, \boldsymbol{y}_{0:k-1} \sim \mathrm{N}(\cdot; \hat{m}_{k-1}, \boldsymbol{\tau}^{-1} \hat{P}'_{k-1})$$

for some parameters $\alpha_{k-1} > 0$, $\beta_{k-1} > 0$, μ_{k-1} and P'_{k-1}. Since the prediction step only applies to $\boldsymbol{\theta}_{k-1}$, it has no influence on α_{k-1} and β_{k-1} and we simply find that

$$\begin{aligned}m_k &= F_k \hat{m}_{k-1} \\ P'_k &= F_k \hat{P}'_{k-1} F_k^\intercal + U'_k.\end{aligned}$$

Note that assuming that the prior variance is $\boldsymbol{\tau}^{-1} P'_0$ made the prediction equation for the variance simpler by enabling the factorisation of the precision. Similarly, the forecasting distribution at time step k given $\boldsymbol{\tau}$ is easily deduced to be

$$p_{\boldsymbol{y}_k|\boldsymbol{\tau},\boldsymbol{y}_{0:k-1}}(y_k \,|\, \tau, y_0, \dots, y_{k-1}) = \mathrm{N}\big(y_k; H_k m_k, \boldsymbol{\tau}^{-1} S'_k\big)$$

with $S'_k = H_k P'_k H_k^\intercal + V'_k$). Applying the Kalman update equation yields

$$\begin{aligned}\hat{m}_k &= m_k + K_k z_k \\ \hat{P}'_k &= (I_d - K_k H_k) P'_k\end{aligned}$$

where the innovation and the Kalman gain are expressed as

$$\begin{aligned} z_k &= y_k - H_k m_k \\ K_k &= P_k H_k^\intercal S_k^{-1} = (\boldsymbol{\tau}^{-1} P_k') H_k^\intercal (\boldsymbol{\tau}^{-1} S_k')^{-1} = P_k' H_k^\intercal S_k'^{-1}. \end{aligned}$$

At the moment, we have determined the posterior distribution of $\boldsymbol{\theta}_k$ given $\boldsymbol{\tau}$. However, it is Bayes' rule for the joint $(\boldsymbol{\theta}_k, \boldsymbol{\tau})$ that we need to obtain. Reusing the notations of Section 2.3.2 with τ instead of ψ, these two distributions can be related as follows:

$$\begin{aligned} p_{\boldsymbol{\theta}_k, \boldsymbol{\tau} | \boldsymbol{y}_{0:k}}(\theta_k, \tau \,|\, y_{0:k}) &= \frac{\bar{\ell}_k(y_k \,|\, \theta_k, \tau) p_{\boldsymbol{\theta}_k, \boldsymbol{\tau} | \boldsymbol{y}_{0:k-1}}(\theta_k, \tau \,|\, y_{0:k-1})}{p_{\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1}}(y_k \,|\, y_{0:k-1})} \\ &= \frac{\bar{\ell}_k(y_k \,|\, \theta_k, \tau) p_{\boldsymbol{\theta}_k | \boldsymbol{\tau}, \boldsymbol{y}_{0:k-1}}(\theta_k \,|\, \tau, y_{0:k-1})}{p_{\boldsymbol{y}_k | \boldsymbol{\tau}, \boldsymbol{y}_{0:k-1}}(y_k \,|\, \tau, y_{0:k-1})} \times \frac{p_{\boldsymbol{y}_k | \boldsymbol{\tau}, \boldsymbol{y}_{0:k-1}}(y_k \,|\, \tau, y_{0:k-1}) p_{\boldsymbol{\tau} | \boldsymbol{y}_{0:k-1}}(\tau \,|\, y_{0:k-1})}{p_{\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1}}(y_k \,|\, y_{0:k-1})} \end{aligned}$$

The first term indeed corresponds to the case of the Kalman filter update for a fixed τ. The distributions in the numerator of the second term are the marginal likelihood $p_{\boldsymbol{y}_k | \boldsymbol{\tau}, \boldsymbol{y}_{0:k-1}}(\cdot \,|\, \cdot)$ and the posterior distribution of $\boldsymbol{\tau}$ at time step $k-1$, which can be combined as

$$p_{\boldsymbol{y}_k | \boldsymbol{\tau}, \boldsymbol{y}_{0:k-1}}(y_k \,|\, \tau, y_{0:k-1}) \mathrm{Ga}(\tau; \alpha_{k-1}, \beta_{k-1}) \propto \tau^{d'/2} \exp\Big(-\frac{\tau}{2} z_k^\intercal S_k'^{-1} z_k \Big) \tau^{\alpha_{k-1}-1} \exp(-\beta_{k-1}\tau)$$

so that

$$\alpha_k = \alpha_{k-1} + \frac{d'}{2} \qquad \text{and} \qquad \beta_k = \beta_{k-1} + \frac{1}{2} z_k^\intercal S_k'^{-1} z_k.$$

Noting that β_k is the only term depending on y_k, one can also compute the marginal likelihood $p_{\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1}}(y_k \,|\, y_{0:k-1})$ by integration as follows

$$\begin{aligned} p_{\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1}}(y_k \,|\, y_{0:k-1}) &= \int p_{\boldsymbol{y}_k | \boldsymbol{\tau}, \boldsymbol{y}_{0:k-1}}(y_k \,|\, \tau, y_{0:k-1}) \mathrm{Ga}(\tau; \alpha_{k-1}, \beta_{k-1}) \,\mathrm{d}\tau \\ &\propto \int \tau^{\alpha_k - 1} \exp(-\beta_k \tau) \,\mathrm{d}\tau \propto \beta_k^{-\alpha_k}. \end{aligned}$$

Based on the expressions of α_k and β_k we find

$$p_{\boldsymbol{y}_k | \boldsymbol{y}_{0:k-1}}(y_k \,|\, y_{0:k-1}) \propto \left(1 + \frac{1}{2\alpha_{k-1}} (y_k - H_k m_k)^\intercal \Big(\frac{\beta_{k-1}}{\alpha_{k-1}} S_k' \Big)^{-1} (y_k - H_k m_k) \right)^{-\frac{2\alpha_{k-1} + d'}{2}}$$

which is multivariate t-distribution with $2\alpha_{k-1}$ degrees of freedom with location $H_k m_k$ and shape matrix $\frac{\beta_{k-1}}{\alpha_{k-1}} S_k'$. The statistics of such a distribution yield

$$\begin{aligned} \mathbb{E}(\boldsymbol{y}_k \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) &= H_k m_k, \qquad \text{if } 2\alpha_{k-1} > 1 \\ \mathrm{var}(\boldsymbol{y}_k \,|\, \boldsymbol{y}_{0:k-1} = y_{0:k-1}) &= \frac{\beta_{k-1}}{\alpha_{k-1} - 1} S_k', \qquad \text{if } 2\alpha_{k-1} > 2, \end{aligned}$$

the mean and the variance being undefined otherwise.

Since $\boldsymbol{\tau}$ follows a gamma distribution, the random variance $\boldsymbol{\tau}^{-1}$ follows an *inverse-gamma distribution* with the same parameters, that is

$$\boldsymbol{\tau}^{-1} \,|\, \boldsymbol{y}_{0:k} = y_{0:k} \sim \text{Inv-Ga}(\cdot; \alpha_k, \beta_k)$$

with

$$\text{Inv-Ga}(x; \alpha_k, \beta_k) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} x^{-\alpha_k - 1} \exp(-\beta_k / x).$$

The mean of such a random variable is defined only when $\alpha_k > 1$ as $\beta_k/(\alpha_k - 1)$ and the variance is defined only when $\alpha_k > 2$ as $\beta_k^2/((\alpha_k - 1)^2(\alpha_k - 2))$.