# Chapter 7

# State Space Models and the Kalman Filter

## 7.1 State Space Models

We say that an $I\!R^d$-valued stochastic sequence $X = (X_n : n \geq 0)$ is a *state space model* if it evolves according to a recursion of the form

$$X_{k+1} = F_k X_k + W_k + u_k$$

for $k \geq 0$, where $(F_k : k \geq 0)$ is a (deterministic) sequence of $d \times d$ matrices, $(u_k : k \geq 0)$ is a (deterministic) sequence of $d \times 1$ vectors, and $(W_k : k \geq 0)$ is a sequence of independent $I\!R^d$-valued random vectors for which $\mathrm{E}[W_k] = 0$ and $\mathrm{E}\left[\|W_k\|^2\right] < \infty$.
Note that such state space models arise naturally in the context of $p$'th order autoregressive sequences (as a consequence of representing such $p$'th order autoregressions as vector-valued first-order autoregressions). State space models form a flexible class of stochastic processes that are analytically and statistically tractable, and are broad enough to be of value in applied areas ranging from signal processing to control systems.

## 7.2 Partially Observed State Space Models

In some applications settings, one is unable to directly observe the state of the system. Instead, one observes some "noise-corrupted function" of the state. For example, in the setting of the state space model, one might directly observe $(Z_k : k \geq 0)$, where,

$$Z_k = H_k X_k + V_k$$

for $k \geq 0$. Here, $(H_k : k \geq 0)$ is a (deterministic) sequence of $n \times d$ matrices, and $(V_k : k \geq 0)$ is a stochastic sequence of $n \times 1$ random vectors. Specifically, we assume that $((W_k, V_k) : k \geq 0)$ is a sequence of $I\!R^{d+m}$-valued random vectors with $\mathrm{E}[W_k] = 0$, $\mathrm{E}[V_k] = 0$, $\mathrm{E}\left[\|W_k\|^2\right] < \infty$, $\mathrm{E}\left[\|V_k\|^2\right] < \infty$, and

$$\mathrm{E}\left[\begin{pmatrix} W_k & V_k \end{pmatrix}^T \begin{pmatrix} W_k \\ V_k \end{pmatrix}\right] \triangleq \begin{pmatrix} Q_k & S_k \\ S_k^T & R_k \end{pmatrix}$$

for $k \geq 0$.
"Filtering" is concerned with trying to extract the current state from the observed history of the observation process. To be precise, the goal is to efficiently compute

$$\mathrm{E}[X_n | Z_j : 0 \leq j \leq n]$$

For $n \geq j$, put

$$\hat{X}_{n|j} = \mathrm{E}[X_n | Z_l : 0 \leq l \leq j]$$

Thus, filtering is concerned with computing $\hat{X}_{n|n}$.
On the other hand, the "prediction problem" is concerned with computing $\hat{X}_{n|j}$ for $n > j$. The $k$-step prediction problem involves computing $\hat{X}_{n+k|n}$.
Finally, "smoothing" deals with the computation of conditional expressions of the form

$$\mathrm{E}\left[X_n | Z_i : 0 \leq i \leq n+j\right]$$

for $n \geq 0$, $j \geq 1$.
If the $(W_k, V_k)$'s are Gaussian, then we know that all the above conditional expectations are affine functions of the observed variables. The formulae for the coefficients of the affine function involve (in the case of filtering) the inverse of the covariance matrix of $(Z_j : 0 \leq j \leq n)$. In principle, the complexity of re-computing the inverse covariance matrix as a function of $n$ could increase superlinearly in $n$. Furthermore, in real-time applications, one would like a recursive (in $n$) implementation in which the complexity of an update is bounded in $n$.
Obviously, the complexity of computing such an inverse covariance matrix typically increases superlinearly in $n$, and offers no opportunity for efficient recursive computation. However, the special structure of the above state space/ observation model turns out to permit efficient recursive computation. This is the key idea underlying Kalman Filtering.
Remark: In order to compute the filtered predictor $\hat{X}_{n|n}$, we must know the joint distribution of $((X_j, Z_j) : 0 \leq j \leq n)$. In many filtering applications, one can do off-line experimentation to estimate the parameters of $((X_j, Z_j) : j \geq 0)$. For example, suppose that one wishes to develop a real-time filter to estimate the state variables for a complex pump (e.g. an artificial heart). Prior to implantation of the pump, one can attach complex (and expensive) measurement devices that may be impractical to apply in the required real-time setting. This off-line data is then used to estimate the model parameters that are needed for the real-time filter computation.

## 7.3 The Innovations Sequence

The first step in deriving the Kalman Filter is to introduce the so-called "innovations" sequence. From this point henceforth, assume that the $(W_j, V_j)$'s are Gaussian random vectors.
Remark: If the $(W_j, V_j)$'s are non-Gaussian, the Kalman filter computes the best affine predictor of $X_n$ based on $(Z_j : 0 \leq j \leq n)$ (rather than the conditional expectation).
Put

$$\hat{Z}_{n|j} = \mathrm{E}\left[Z_n | Z_i : 0 \leq i \leq j\right] ,$$

$$\tilde{X}_n = X_n - \hat{X}_{n|n-1} ,$$

$$\tilde{Z}_n = Z_n - \hat{Z}_{n|n-1} .$$

**Definition 7.1:** The sequence $(\tilde{Z}_n : n \geq 0)$ is called the *innovations* sequence.

**Proposition 7.1:** The sequence $(\tilde{Z}_n : n \geq 0)$ satisfies:

(i) $\mathrm{E}\left[\tilde{Z}_n\right] = 0$ for $n \geq 0$ ;

(ii) $\mathrm{E}\left[\tilde{Z}_i \tilde{Z}_j^T\right] = 0$ for $i \neq j$

Remark: Because the $\tilde{Z}_n$'s are Gaussian, Proposition Proposition 7.1 guarantees that the $\tilde{Z}_n$'s are independent random vectors.

*Proof.* Part (i) is trivial. For part (ii), observe that if $i < j$, then

$$\begin{aligned}
\mathrm{E}\left[\tilde{Z}_i\tilde{Z}_j^T|Z_r : 0 \le r \le j-1\right] &= \tilde{Z}_i\,\mathrm{E}\left[\tilde{Z}_j^T|Z_r : 0 \le r \le j-1\right] \\
&= \tilde{Z}_i\,\mathrm{E}\left[(Z_j - \mathrm{E}\left[Z_j|Z_u : 0 \le u \le j-1\right])|Z_r : 0 \le r \le j-1\right] \\
&= \tilde{Z}_i(\mathrm{E}\left[Z_j|Z_r : 0 \le r \le j-1\right] - \mathrm{E}\left[Z_j|Z_u : 0 \le u \le j-1\right]) \\
&= 0
\end{aligned}$$

Consequently, $\mathrm{E}\left[\tilde{Z}_i\hat{Z}_j^T\right] = 0$. □

Furthermore, the information carried in $(Z_j : 0 \le j \le n)$ is exactly identical to the information carried in $(\tilde{Z}_j : 0 \le j \le n)$. To be precise, there is a $1-1$ affine bijection between $(Z_0, \ldots, Z_n)$ and $(\tilde{Z}_0, \ldots, \tilde{Z}_n)$. To see this, note that

$$\tilde{Z}_n = Z_n - \mathrm{E}\left[Z_n|Z_j : 0 \le j \le n-1\right]$$

Since $\mathrm{E}\left[Z_n|Z_j : 0 \le j \le n-1\right]$ is affine in $(Z_j : 0 \le j \le n-1)$, it follows that $\tilde{Z}_n$ is an affine function of $(Z_j : 0 \le j \le n-1)$.
On the other hand, $Z_0 = \tilde{Z}_0 + \mathrm{E}\left[Z\right]_0$. Assume that $(Z_j : 0 \le j \le n-1)$ can be expressed as an affine function of $(\tilde{Z}_j : 0 \le j \le n-1)$ But $Z_n = \tilde{Z}_n + \mathrm{E}\left[Z_n|Z_j : 0 \le j \le n-1\right]$. Since $\mathrm{E}\left[Z_n|Z_j : 0 \le j \le n-1\right]$ is an affine function of $(Z_j : 0 \le j \le n-1)$ and hence an affine function of $(\tilde{Z}_j : 0 \le j \le n-1)$, it follows that $Z_n$ is an affine function of $(\tilde{Z}_j : 0 \le j \le n)$.
Because of the existence of this $1-1$ affine bijection, it follows that

$$\mathrm{E}\left[X_n|Z_i : 0 \le i \le j\right] = \mathrm{E}\left[X_n|\tilde{Z}_i : 0 \le i \le j\right]$$

**Exercise 7.1:** Let $X$ be real-valued and let $Z$ be $S_1$-valued. Suppose that $h : S_1 \to S_2$ is a bijection. If $\mathrm{E}\left[X^2\right] < \infty$ , prove that

$$\mathrm{E}\left[X|Z\right] = \mathrm{E}\left[X|h(Z)\right]$$

Consequently,

$$\hat{X}_{n|j} = \mathrm{E}\left[X_n|\tilde{Z}_i : 0 \le i \le j\right]$$

Because of the orthogonality of the innovations sequence, the covariance matrix of $(\tilde{Z}_0, \ldots, \tilde{Z}_n)$ is block-diagonal. The recursive nature of the Kalman filter is a consequence of this block-diagonal structure.

Assume henceforth that the covariance matrices $\mathrm{E}\left[\tilde{Z}_n\tilde{Z}_n\right]^T$ are non-singular for $n \ge 0$. (Otherwise, there are, as usual, straight forward but tedious modifications of the calculations to follow that can be implemented.) We will focus on computing $\hat{X}_{n+1|n}$ rather than the true filtered estimate $\hat{X}_{n|n}$. ($\hat{X}_{n|n}$ can be computed relatively cleanly from $\hat{X}_{n+1|n}$.) Because of the block diagonal structure,

$$\begin{aligned}
\hat{X}_{n+1|n} &= \mathrm{E}\left[X_{n+1}\right] + \sum_{j=0}^{n} \mathrm{E}\left[X_{n+1}\right]\tilde{Z}_j^T(\mathrm{E}\left[\tilde{Z}_j\tilde{Z}_j^T\right])^{-1}\tilde{Z}_j \\
&= (\mathrm{E}\left[X_{n+1}\right] + \mathrm{E}\left[X_{n+1}\right]\tilde{Z}_n^T(\mathrm{E}\left[\tilde{Z}_n\tilde{Z}_n^T\right])^{-1}\tilde{Z}_n) + \\
&\quad (\mathrm{E}\left[X_{n+1}\right] + \sum_{j=0}^{n-1} \mathrm{E}\left[X_{n+1}\right]\tilde{X}_j^T(\mathrm{E}\left[\tilde{Z}_j\tilde{Z}_j^T\right])^{-1}\tilde{Z}_j) - \mathrm{E}\left[X_{n+1}\right] \\
&= \mathrm{E}\left[X_{n+1}|\tilde{Z}_n\right] + \tilde{X}_{n+1|n-1} - \mathrm{E}\left[X_{n+1}\right]
\end{aligned}$$

## 7.4 Derivation of the Kalman filter

We have already seen that the introduction of the innovations sequence permits us to dramatically simplify the computation of the one-step ahead predictor $\hat{X}_{n+1|n}$.
To simplify further, note that

$$\begin{aligned}
\hat{X}_{n+1|n-1} &= \mathrm{E}\left[X_{n+1}|Z_j : 0 \le j \le n-1\right] \\
&= \mathrm{E}\left[F_n X_n + W_n + u_n|Z_j : 0 \le j \le n-1\right] \\
&= F_n \hat{X}_{n|n-1} + u_n
\end{aligned}$$

where $\hat{X}_{n|n-1}$ is the one-step ahead predictor at the previous time step. Hence,

$$\hat{X}_{n+1|n} = F_n \hat{X}_{n|n-1} + \mathrm{E}\left[X_{n+1}\tilde{Z}_n^T\right](\mathrm{E}\left[\tilde{Z}_n\tilde{Z}_n^T\right])^{-1}\tilde{Z}_n$$

Now,

$$\begin{aligned}
\tilde{Z}_n &= Z_n - \mathrm{E}\left[Z_n|Z_j : 0 \le j \le n-1\right] \\
&= H_n X_n + V_n - \mathrm{E}\left[H_n X_n + V_n|Z_j : 0 \le j \le n-1\right] \\
&= H_n X_n + V_n - H_n X_{n|n-1} \\
&= H_n \tilde{X}_n + V_n
\end{aligned}$$

So,

$$\begin{aligned}
\mathrm{E}\left[X_{n+1}\tilde{Z}_n^T\right] &= \mathrm{E}\left[(F_n X_n + W_n + u_n)(H_n\tilde{X}_n + V_n)^T\right] \\
&= F_n\,\mathrm{E}\left[X_n\tilde{X}_n^T\right]H_n^T + \mathrm{E}\left[W_n V_n^T\right] \\
&= F_n\,\mathrm{E}\left[\tilde{X}_n\tilde{X}_n^T\right]H_n^T + S_n \\
&= F_n\Sigma_{n|n-1}H_n^T + S_n
\end{aligned}$$

where $\Sigma_{n|n-1}$ is the convenience matrix of $\tilde{X}_n$. Also,

$$\begin{aligned}
\mathrm{E}\left[\tilde{Z}_n\tilde{Z}_n^T\right] &= \mathrm{E}\left[(H_n\tilde{X}_n + V_n)(H_n\tilde{X}_n + V_n)^T\right] \\
&= H_n\Sigma_{n|n-1}H_n^T + R_n
\end{aligned}$$

Finally, $\tilde{Z}_n$ may be rewritten as

$$\tilde{Z}_n = Z_n - H_n X_{n|n-1}$$

Consequently, we obtain the recursive equation

$$\hat{X}_{n+1|n} = F_n\hat{X}_{n|n-1} + u_n + (F_n\Sigma_{n|n-1}H_n^T)(H_n\Sigma_{n|n-1}H_n^T + R_n)^{-1}(Z_n - H_n X_{n|n-1}) \tag{7.1}$$

Of course, this presumes knowledge of $\Sigma_{n|n-1}$. Note that

$$\begin{aligned}
\tilde{X}_{n+1} &= X_{n+1} - \hat{X}_{n+1|n} \\
&= F_n X_n + W_n + u_n - \hat{X}_{n+1|n} \\
&= F_n X_n + W_n + u_n - F_n\hat{X}_{n|n-1} - u_n - K_n(Z_n - H_n X_{n|n-1}) \\
&= F_n\tilde{X}_n + W_n - K_n(H_n X_n + V_n - H_n X_{n|n-1}) \\
&= (F_n - K_n H_n)\tilde{X}_n + W_n - K_n V_n
\end{aligned}$$

where

$$K_n = F_n \Sigma_{n|n-1} H_n^T (H_n \Sigma_{n|n-1} H_n^T + R_n)^{-1} \,.$$

So,

$$\Sigma_{n+1|n} = (F_n - K_n H_n)\Sigma_{n|n-1}(F_n - K_n H_n)^T + (I - K_n) \begin{pmatrix} Q_n & S_n \\ S_n^T & R_n \end{pmatrix} \begin{pmatrix} I \\ -K_n \end{pmatrix} \tag{7.2}$$

This equation recursively computes $\Sigma_{n+1|n}$ from $\Sigma_{n|n-1}$. Equations (7.1) and (7.2) together provide a full recursive implementation for computing $\hat{X}_{n+1|n}$.

Note, that in this Gaussian context, the Kalman Filter is actually computing the entire conditional distribution of $X_{n+1}$, given $(Z_j : 0 \leq j \leq n)$. In particular, $X_{n+1}$ has a Gaussian distribution with mean $\hat{X}_{n+1|n}$ and covariance matrix $\Sigma_{n|n-1}$, when conditioned on $(Z_j : 0 \leq j \leq n)$. Hence,

$$\mathrm{P}\left\{X_{n+1} \in B | Z_0, Z_1, \ldots, Z_n\right\} = \mathrm{P}\left\{N(\hat{X}_{n+1|n}, \Sigma_{n|n-1}) \in B\right\}$$

for any subset $B \in \mathbb{R}^d$

Remark: From a numerical standpoint, recursion (7.2) can give rise to matrices $\Sigma_{n|n-1}$ that are not symmetric and non-negative definite (due to numerical instabilities). In many implementations, (7.2) is re-written as a recursion involving the "square root" (i.e. a matrix $\Gamma_n$ such that $\Gamma_n \Gamma_n^T = \Sigma_{n|n-1}$) instead. This also improves the condition number of the matrices that appear in the corresponding recursive updating procedure.

Remark: The filtered quantity $\hat{X}_{n|n}$ (as opposed to the one-step predictor $\hat{X}_{n|n-1}$) can also be computed recursively. See for example, p.115 – 118 of B.D.O. Anderson and J.B. Moore (1979), Optimal Filtering, for a full discussion.