Appendix B

Least-square estimation and filtering

B.1 Linear least-variance estimators of random vectors

Let $T: \mathbb{R}^n \rightarrow \mathbb{R}^m; \ X \rightarrow Y$ be a transformation acting between two spaces of random vectors with instances in $\mathbb{R}^m$ and $\mathbb{R}^n$ (the model generating the data). We are interested in building an estimator for the random vector $X$, given measurements of instances of the random vector $Y$. An estimator is a function $T^*: \mathbb{R}^m \rightarrow \mathbb{R}^n; \ Y \rightarrow \hat{X} = T^*(Y)$, which solves an optimization problem of the form

$$
\hat{T}^* = \arg \min_{T \in T} C(X - T^*(Y))_T
$$

(B.1)

where $T$ is a suitable chosen class of functions and $C(\cdot)_T$ some cost in the $X$-space.

We concentrate on one of the simplest possible choices, which correspond to \textit{minimum variance affine} estimators

$$
T = \{A \in \mathbb{R}^{n \times m}; b \in \mathbb{R}^n \mid T^*(Y) = AY + b\} \tag{B.2}
$$

$$
C(\cdot)_T = E[\|\cdot\|^2] \tag{B.3}
$$

where the latter operator takes the expectation of the squared euclidean norm of the random vector $Y$. Therefore, we seek for

$$
(\hat{A}, \hat{b}) = \arg \min_{A,b} E[\|X - (AY + b)\|^2] \tag{B.4}
$$

We call $\mu_X = E[X]$ and $\Sigma_X = E[XX^T]$, and similarly for $Y$. First notice that if $\mu_X = \mu_Y = 0$, then $\hat{b} = 0$. Therefore, consider the centered vectors $\hat{X} = X - \mu_X$ and $\hat{Y} = Y - \mu_Y$ and the reduced problem

$$
\hat{A} = \arg \min_{\hat{A}} E[\|\hat{X} - A\hat{Y}\|^2]. \tag{B.5}
$$

Now observe that

$$
E[\|X - AY - b\|^2] = E[\|A\hat{Y} - \hat{X} + (A\mu_X + b - \mu_Y)\|^2]
$$

$$
= E[\|\hat{X} - A\hat{Y}\|^2] + \|A\mu_X + b - \mu_Y\|^2. \tag{B.6}
$$

Hence, if we assume for a moment that we have found $\hat{A}$ that solves the problem (B.5), then trivially

$$
\hat{b} = \mu_X - \hat{A}\mu_Y \tag{B.7}
$$

annihilates the second term of eq. (B.6).

Therefore, we will concentrate on the case $\mu_X = \mu_Y = 0$ without loss of generality.
B.1.1 Projections onto the range of a random vector

The set of all random variables $Z_i$ defined on the same probability space, with zero-mean $E[Z_i] = 0$ and finite variance $\Sigma_{Z_i} < \infty$ is a Hilbert space with the inner-product given by

$$\langle Z_i, Z_j \rangle_H = \Sigma_{Z_i Z_j} = E[Z_i Z_j].$$

(B.8)

In this space the notion of orthogonality corresponds to the notion of uncorrelatedness. The components of a random vector $Y$ define a subspace of such Hilbert space:

$$\mathcal{H}(Y) = \text{span}(Y_1, \ldots, Y_m)$$

(B.9)

where the span is intended over the reals.\footnote{So $\mathcal{H}(Y)$ is the space of random variables which are linear combination of $Y_i$, $i = 1, \ldots, m$.} We say that the subspace $\mathcal{H}(Y)$ is full rank if $\Sigma_Y = E[YY^T] > 0$.

The structure of a Hilbert space allows us to make use of the concept of orthogonal projection of a random variable onto the span of a random vector:

$$\hat{Z} = \text{pr}_{\mathcal{H}(Y)}(X) \iff \langle X - \hat{Z}, Z \rangle_H = 0, \quad \forall \ Z \in \mathcal{H}(Y)$$

$$\iff \langle X - \hat{Z}, Y_i \rangle_H = 0, \quad \forall \ i = 1 \ldots n$$

(B.10)

$$\Rightarrow \hat{E}[X|Y]$$

$$\Rightarrow \hat{X}(Y)$$

(B.11)

(B.12)

The notation $\hat{E}[X|Y]$ is often used for the projection of $X$ over the span of $Y$.\footnote{The resemblance with a conditional expectation is due to the fact that, in the presence of Gaussian random vectors such a projection is indeed the conditional expectation.}

B.1.2 Solution for the linear (scalar) estimator

Let $Z = AY$ be a linear estimator for the random variable $X \in \mathbb{R}$; $A \in \mathbb{R}^{n \times m}$ is a row-vector, and $Y \in \mathbb{R}^m$ an $m$-dimensional column random vector. The least-square estimate $\hat{Z}$ is given by the choice of $A$ that solves the following problem:

$$\hat{A} = \arg \min_A \| AY - X \|_H^2$$

(B.13)

where $\| \cdot \|_H^2 = E[\| \cdot \|^2]$ is the norm induced by the inner product $\langle \cdot, \cdot \rangle_H$.

**Theorem B.1.** The solution $\hat{Z} = \hat{A}Y$ to the problem (B.13) exists, is unique and corresponds to the orthogonal projection of $X$ onto the span of $Y$:

$$\hat{Z} = \text{pr}_{\mathcal{H}(Y)}(X)$$

(B.14)

The proof is an easy exercise. In the following we report an explicit construction of the best estimator $\hat{A}$. From substituting the expression of the estimator onto the definition of orthogonal projection (B.12), we get

$$0 = \langle X - \hat{A}Y, Y_i \rangle_H = E[(X - \hat{A}Y)Y_i]$$

(B.15)

which holds iff $E[XY_i] = \hat{A}E[YY_i]$, $\forall \ i = 1 \ldots n$. In a row-vector notation we write

$$E[XY^T] = \hat{A}E[YY^T]$$

$$\Sigma_{XY} = \hat{A}\Sigma_Y$$

(B.16)

which, provided that $\mathcal{H}(Y)$ is full rank, gives $\hat{A} = \Sigma_{XY}\Sigma_Y^{-1}$. 

B.1.3 Affine least-variance estimator

Suppose we want to compute the best estimator of a zero-mean random vector $X$ as a linear map of the zero-mean random vector $Y$. We just have to repeat the construction reported in the previous section for each component $X_i$ of $X$, so that the rows $\hat{A}_i$ of the matrix $\hat{A}$ are given by

$$
\hat{A}_i = \Sigma_{X,Y} \Sigma_Y^{-1} \\
\vdots \\
\hat{A}_n = \Sigma_{X_n,Y} \Sigma_Y^{-1}
$$

which eventually gives us

$$
\hat{A} = \Sigma_{XY} \Sigma_Y^{-1}.
$$

If now the vectors $X$ and $Y$ are not zero-mean, $\mu_X \neq 0$, $\mu_Y \neq 0$, we first transform it into a zero-mean problem by defining $\tilde{Y} = Y - \mu_Y$, $\tilde{X} = X - \mu_X$, then solve for the linear least-variance estimator $\hat{A} = \Sigma_{\tilde{X},\tilde{Y}} \Sigma_{\tilde{Y}}^{-1} = \Sigma_{XY} \Sigma_Y^{-1}$, and then substitute to get

$$
\hat{Z} = \mu_X + \Sigma_{XY} \Sigma_Y^{-1}(Y - \mu_Y)
$$

which is the least-variance affine estimator

$$
\hat{Z} = \hat{E}[X|Y] = \hat{A}Y + \hat{b}
$$

where

$$
\hat{A} = \Sigma_{XY} \Sigma_Y^{-1}, \\
\hat{b} = \mu_X - \Sigma_{XY} \Sigma_Y^{-1} \mu_Y.
$$

It is an easy exercise to compute the variance of the estimation error $\tilde{X} = X - \hat{Z}$:

$$
\Sigma_{\tilde{X}} = \Sigma_X - \Sigma_{XY} \Sigma_Y^{-1} \Sigma_Y X.
$$

If we interpret the variance of $X$ as the “prior uncertainty”, and the variance of $\tilde{X}$ as the “posterior uncertainty”, we may interpret the second term (which is positive semi-definite) of the above equation as a “decrease” of the uncertainty.

B.1.4 Properties and interpretations of the least-variance estimator

The variance of the estimation error in equation (B.23) is by construction the smallest that can be achieved with an affine estimator. Of course if we consider a broader class $T$ of estimators, the estimation error can be further decreased, unless the model that generates the data $T$ is itself affine:

$$
Y = T(X) = FX + W.
$$

In such a case, using the matrix inversion lemma \(^3\), it is easy to compute the expression of the optimal (affine) estimator that depends only upon $\Sigma_X$, $\Sigma_W$ and $F$:

$$
\hat{Z} = \Sigma_X F^T (F \Sigma_X F^T + \Sigma_W)^{-1} Y
$$

which achieves a variance of the estimation error equal to

$$
\Sigma_{\hat{X}} = \Sigma_X - \Sigma_X F^T (F \Sigma_X F^T + \Sigma_W)^{-1} F \Sigma_X.
$$

\(^3\)If $A, B, C, D$ are real matrices of the appropriate dimensions with $A$ and $C$ invertible, then $(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA$. 


Appendix B. Least-Square Estimation and Filtering

Projection onto an orthogonal sum of subspaces

Let \( Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \) be such that
\[
\mathcal{H}(Y) = \mathcal{H}(Y_1) \oplus \mathcal{H}(Y_2).
\] (B.27)

We may now wonder what are the conditions under which
\[
\hat{E}[X|Y] = \hat{E}[X|Y_1] + \hat{E}[X|Y_2].
\] (B.28)

After an easy calculation one can see that the above is true iff \( E[Y_1Y_2^T] = 0 \), which is to say when
\[
\mathcal{H}(Y_1) \perp \mathcal{H}(Y_2)
\] (B.29)

Change of basis

Suppose that instead of measuring the instances of a random vector \( Y \) we measure another random vector \( Z \) which is related to \( Y \) via a change of basis: \( Z = TY \mid T \in GL(m) \). If we call \( \hat{E}[X|Y] = \hat{A}Y \), then it is immediate to see that
\[
\hat{E}[X|Z] = \Sigma_XZ \Sigma_Z^{-1}Z
\]
\[
= \Sigma_XTY^T(T^{-1}\Sigma_YT^{-1})Z
\]
\[
= \Sigma_X\Sigma_Y^{-1}T^{-1}Z.
\] (B.30)

Innovations

The linear least-variance estimator involves the computation of the inverse of the output covariance matrix \( \Sigma_Y \). It may be interesting to look for changes of bases \( T \) that transform the output \( Y \) into \( Z = TY \) such that \( \Sigma_Z = I \). In such a case the optimal estimator is simply
\[
\hat{E}[X|Z] = \Sigma_XZ Z.
\] (B.31)

Let us pretend for a moment that the components of the vector \( Y \) are samples of a process taken over time: \( Y_i = y(i) \), and call \( y^t = [Y_1, \ldots, Y_t]^T \) the history of the process up to time \( t \). Each component (sample) is an element of the Hilbert space \( \mathcal{H} \), which has a well-defined notion of orthogonality, and where we can apply Gram-Schmidt procedure in order to make the “vectors” \( y(i) \) orthogonal (uncorrelated).

\[
\begin{align*}
 v_1 & \doteq y(1) \\
v_2 & \doteq y(2) - \langle y(2), e_1 \rangle e_1 \\
 \vdots & \\
v_t & \doteq y(t) - \sum_{i=1}^{t-1} \langle y(i), e_i \rangle e_i
\end{align*}
\]

The process \( \{e\} \), whose instances up to time \( t \) are collected into the vector \( e^t = [e_1, \ldots, e_t]^T \) has a number of important properties:

1. The component of \( e^t \) are orthonormal in \( \mathcal{H} \) (or equivalently \( \{e\} \) is an uncorrelated process). This holds by construction.

2. The transformation from \( y \) to \( e \) is causal, in the sense that – if we represent it as a matrix \( L_t \) such that
\[
y^t = L_t e^t
\] (B.32)
then \( L_t \in L_+ \) is lower-triangular with positive diagonal. This follows from the Gram-Schmidt procedure.
3. The process \( \{e\} \) is equivalent to \( \{y\} \) in the sense that they generate the same span

\[
\mathcal{H}(y^t) = \mathcal{H}(e^t). \tag{B.33}
\]

This property follows from the fact that \( L_t \) is non-singular.

4. If we write \( y^t = L_t e^t \) in matrix form as \( Y = LE \), then \( \Sigma_Y = LL^T \).

The meaning of the components of \( v \), and the name innovation, comes from the fact that we can interpret

\[
v_t = y(t) - \hat{E}[y(t)|y^{t-1}] \tag{B.34}
\]

as a one-step prediction error. The process \( e \) is a scaled version of \( v \) such that its variance is the identity.

We may now wonder whether each process \( \{y\} \) has an innovation, and if so, whether it is unique. The following theorem, which is known as the Cholesky factorization theorem or Spectral Factorization theorem depending upon the context, states the conditions:

**Theorem B.2.** There exists a unique vector \( E \) which is causally equivalent to \( Y \) iff there exists a unique lower-triangular matrix \( L \), called Choleski’s factor, such that \( \Sigma_Y = LL^T \).

**Remark B.1.** The Cholesky factor can be interpreted as a “whitening filter”, in the sense that it acts on the components of the vector \( Y \) in a causal fashion to make them uncorrelated.

We may consider a two-step solution to the problem of finding the least-square filter: a “whitening step”

\[
E = L^{-1}Y \tag{B.35}
\]

where \( \Sigma_E = I \), and a projection onto \( \mathcal{H}(E) \):

\[
\hat{X}(Y) = \Sigma_{XE}L^{-1}Y. \tag{B.36}
\]

### B.2 Linear least-variance estimator for stationary processes

In the previous section we have interpreted a column-vector as a collection of samples from a scalar random process, and computed the least-variance estimator by orthogonal projection. In this section we see how this plot generalizes to proper stationary processes. We consider only scalar processes for simplicity of notation, although all considerations can be extended to vector-valued processes.

Let us assume that \( \{x(t)\} \in \mathbb{R}^n \) and \( \{y(t)\} \in \mathbb{R}^m \) are (wide-sense) jointly stationary, i.e.

\[
\Sigma_{xy}(t, s) = E[x(t)y^T(s)] = \Sigma_{xy}(t - s). \tag{B.37}
\]

Again, we restrict our attention to linear estimators of \( \{x(t)\} \) given the measurements of \( \{y(s); s \leq t\} \) up to time \( t \). We denote the estimate by \( \hat{x}(t|t) \). A linear estimator is described by a convolution kernel \( h \) such that

\[
\hat{x}(t|t) = \sum_{k=-\infty}^{t} h(t, k)y(k). \tag{B.38}
\]

The design of the least-variance estimator involves finding the kernel \( h \) such that the estimation error \( \hat{x}(t) = x(t) - \hat{x}(t|t) \) has minimum variance. This is found, as in the previous sections for the
static case, by imposing that the estimation error be orthogonal to the history of the process \{y\} up to time \(t\):

\[
\langle x(t) - \hat{x}(t|t), y(s) \rangle_H = 0, \quad \forall \ s \leq t
\]

\[
E[x(t) y^T(s)] - \sum_{k=-\infty}^{t} h(t,k) E[y(k) y^T(s)] = 0, \quad \forall \ s \leq t
\]

(B.39)

which is equivalent to

\[
\Sigma_{xy}(t-s) = \sum_{k=-\infty}^{t} h(t,k) \Sigma_y(k-s).
\]

(B.40)

The above is equivalent to a linear system with an infinite number of equations, and we will assume that it has a unique solution for \(H\). Given that the processes involved are (jointly) stationary, and the convolution starts at \(-\infty\), it can be easily seen that the kernel \(h\) is time invariant: \(h(t,k) = h(t-k)\). Therefore the last equation is equivalent to

\[
\Sigma_{xy}(t) = \sum_{s=0}^{\infty} h(s) \Sigma_y(t-s), \quad \forall \ t \geq 0
\]

(B.41)

which is called Wiener-Hopf equation and is exactly equivalent to the orthogonality conditions (B.16). In fact, if we \(Z\)-transform the above equation

\[
S_{xy}(z) = H(z) S_y(z)
\]

(B.42)

we have exactly the same expression as equation (B.16), which we could try to solve as

\[
\hat{H}(z) = S_{xy}(z) S_y^{-1}(z)
\]

(B.43)

provided that the spectral density \(S_y\) is invertible. This, however, is not quite the solution we are looking for. In fact, in order to be of any use, the estimator must be causal (it must not depend upon “future” samples of the process \{y\}) and stable (it must return a bounded estimate for bounded data). We can express these conditions by requiring

- causality: \(h(t) = 0, \forall t < 0\) (or \(H(z)\) analytic at \(\infty\))
- stability: \(H(z)\) analytic in \(|z| \geq 1\) (or \(h(t)\) square-summable).

One particular case is when the spectral density of \{y\} is the identity (or equivalently \{y\} is a white noise). Then \(S_y = I\) and we could choose

\[
h(t) = \begin{cases} \\
\Sigma_{xy}(t), & t \geq 0 \\
0, & t < 0.
\end{cases}
\]

(B.44)

This suggests us to try to whiten (or orthonormalize) the measurement process \(\{y\}\) in a similar fashion to what we did in section B.1.4. Indeed we can state a theorem similar to B.2, which is known as the spectral factorization theorem:

**Theorem B.3.** There exists a process \(\{\xi\}\) such that \(\mathcal{H}(\xi^2) = \mathcal{H}(y^2)\) and \(\Sigma_\xi(t) = \Lambda \delta(t)\) iff there exists \(W(z)\) stable and causal, with \(W^{-1}(z)\) causal such that \(S_y(z) = W(z) W(z^{-1})\).
Remark B.2. In words there exists a white process \{\varepsilon\} (called the innovation) which is causally equivalent to \{y\} iff the spectral density of \( y \) has a causal, stable and minimum-phase spectral factor. If we re-scale \( W(z) \) to \( L(z) = W(z)W(\infty)^{-1} \), the innovation \{\varepsilon\} is re-normalized so that \( \Sigma_e(t) = I \delta(t) \), and is called normalized innovation.

We may at this point repeat the two-step construction of the least-variance estimator. First the “whitening step”:

\[
E(z) = L^{-1}(z)Y(z)
\]

and then the causal part of the projection:

\[
\hat{X}(Y) = \begin{cases} 
   \Sigma_{xe}(t) \ast e(t), & t \geq 0 \\
   0, & t < 0 
\end{cases}
\]

where \( \ast \) indicates the standard convolution. Equivalently, if we denote by \([S_{xe}(z)]_+\) the causal part of the \( \mathcal{Z} \)-transform of \( \Sigma_{xe}(t) \), we can write

\[
\hat{X}(Y)(z) = [S_{xe}(z)]_+ E(z).
\]

Since \( S_{xe}(z) = S_{xy}(z)L(z^{-1})^{-1} \), the final expression of our linear, least-variance estimator is (in the \( \mathcal{Z} \)-domain) \( \hat{x} = H(y) \), where the kernel \( H \) is given by

\[
H(z) = [S_{xy}(z)L^{-1}(z^{-1})]_+ L^{-1}(z).
\]

The corresponding filter is known as the Wiener filter. Again we can recover the meaning of the innovation as the one-step prediction error for the measurements: in fact, the best prediction of the process \( \{y\} \), indicated with \( \hat{y}(t|t-1) \), is defined as the projection of \( y(t) \) onto the span of \( \{y\} \) up to \( t-1 \), indicated with \( \mathcal{H}_{t-1}(y) \). Such projection is therefore defined such that

\[
y(t) = \hat{y}(t|t-1)+e(t)
\]

where \( e(t) \perp \mathcal{H}_{t-1}(y) = \mathcal{H}_{t-1}(e) \).

B.3 Linear, finite-dimensional stochastic processes

A linear, finite-dimensional stochastic process (LFDSP) is defined as the output of a linear, finite-dimensional dynamical system driven by white Gaussian noise. Let \( A(t), B(t), C(t), D(t) \) be time-varying matrices of suitable dimensions, \( \{n(t)\} \in \mathcal{N}(0, I) \mid E[n(t)n^T(s)] = I \delta(t-s) \) a white, zero-mean Gaussian noise and \( x_0 \) a random vector which is uncorrelated with \( \{n\} \): \( E[x_0n^T(t)] = 0, \forall t \). Then \( \{y(t)\} \) is a LFDSP if there exists \( \{x(t)\} \) such that

\[
\begin{cases} 
   x(t+1) = A(t)x(t) + B(t)n(t) \\
   y(t) = C(t)x(t) + D(t)n(t)
\end{cases} \quad x(t_0) = x_0
\]

We call \( \{x\} \) the state process, \( \{y\} \) the output (or measurement) process, and \( \{n\} \) the input (or driving) noise. The time-evolution of the state process can be written as the orthogonal sum of the past history (prior to the initial condition), and the present history (from the initial condition until the present time)

\[
x(t) = \Phi^x_{t_0}x_0 + \sum_{k=t_0}^{t-1} \Phi^x_{k+1}B(t)n(t) = \hat{E}[x(t)|\mathcal{H}(x^{t_0})]+\hat{E}[x(t)|x(t_0),\ldots,x(t-1)]
\]

where \( \hat{E}[x(t)|\mathcal{H}(x^{t_0})] = \hat{E}[x(t)|x(t_0),\ldots,x(t-1)] \).
where $\Phi$ denotes a fundamental set of solutions, which is the flow of the differential equation

$$\begin{cases}
\Phi(t + 1, s) = A(t)\Phi(t, s) \\
\Phi(t, t) = I.
\end{cases}$$

(B.52)

In the case of a time-invariant system $A(t) = A$, $\forall t$, then $\Phi(t, s) = A^{t-s}$.

**Remark B.3.** As a consequence of the definitions, the orthogonality between the state and the input noise propagates up to the current time:

$$n(t) \perp \mathcal{H} x(s), \; \forall s \leq t.$$  

(B.53)

Moreover, the past history up to time $s$ is always summarized by the value of the state at that time (Markov property):

$$\hat{E}[x(t)|\mathcal{H}_s(x)] = \hat{E}[x(t)|x(s)] = \Phi(t, s)x(s), \; \forall t \geq s.$$  

(B.54)

### B.4 Stationarity of LFDSF

In order to design the least-squares estimator as in the previous sections, we ask what are the conditions under which a LFDSF is stationary. The first restriction we require is that the system be time-invariant. The mean of the state process at time $t$ is given by

$$\mu_x(t) = A^{t-t_0}\mu_x(0)$$

(B.55)

while the covariance of the state-process

$$\Sigma_x(t, s) = A^{t-s}\Sigma_x(s)$$

(B.56)

evolves according to the following *Ljapunov equation*

$$\Sigma_x(s + 1) = A\Sigma_x(s)A^T + BB^T.$$  

(B.57)

The conditions for stationarity impose that $\sigma_x(t) = \text{const}$ and $\mu_x(t) = \text{const}$. It is easy to prove the following

**Theorem B.4.** Let $A$ be stable (have all eigenvalues in the unit complex circle), then $\Sigma_x(t - t_0) \to \hat{\Sigma}$, where $\hat{\Sigma} = \sum_{k=0}^{\infty} A^k BB^T A^{Tk}$ is the unique equilibrium solution of the above Ljapunov equation, and $\{x\}$ describes asymptotically a stationary process. If $x_0$ is such that $\Sigma_x(t_0) = \hat{\Sigma}$, then the process is stationary for all $t \geq t_0$.

**Remark B.4.** The condition of stability for $A$ is sufficient, but not necessary for generating a stationary process. If, however, the pair $(A, B)$ is completely controllable, so that the noise input affects all of the components of the state, then such a stability condition becomes also necessary.

### B.5 The linear Kalman filter

Suppose we are given a linear finite-dimensional process, which has a realization $(A, B, C, D)$ as in equation (B.50). While we measure the (noisy) output $y(t)$ of such a realization, we do not have access to its state $x(t)$. The Kalman filter is a dynamical model that accepts as input the output
of the process realization, and returns an estimate of its state that has the property of having the least error variance. In order to derive the expression for the filter, we write the LFDSP as follows:

\[
\begin{align*}
    x(t + 1) &= Ax(t) + v(t) & x(t_0) = x_0 \\
y(t) &= Cx(t) + w(t)
\end{align*}
\]  

(B.58)

where we have neglected the time argument in the matrices \( A(t) \) and \( C(t) \) (all considerations can be carried through for time-varying systems as well). \( v(t) = Bn(t) \) is a white, zero-mean Gaussian noise with variance \( Q \), \( w(t) = Dn(t) \), also a white, zero-mean noise, has variance \( R \), so that we could write

\[
\begin{align*}
v(t) &= \sqrt{Q}n(t) \\
w(t) &= \sqrt{R}n(t)
\end{align*}
\]

where \( n \) is a unit-variance noise. In general \( v \) and \( w \) will be correlated, and in particular we will call

\[
S(t) = E[v(t)w^T(t)].
\]

We require that the initial condition \( x_0 \) be uncorrelated from the noise processes

\[
x_0 \perp \{v\}, \{w\}, \forall t.
\]  

(B.60)

The first step is to modify the above model so that the model error \( v \) is uncorrelated from the measurement error \( w \).

**Uncorrelating the model from the measurements**

In order to uncorrelate the model error from the measurement error we can just substitute \( v \) with the complement of its projection onto the span of \( w \). Let us call

\[
\tilde{v}(t) = v(t) - \tilde{E}[v(t)|H(w)] = v(t) - \tilde{E}[v(t)|w(t)]
\]  

(B.61)

the last equivalence is due to the fact that \( w \) is a white noise. We can now use the results from section B.1 to conclude that

\[
\tilde{v}(t) = v(t) - SR^{-1}w(t)
\]  

(B.62)

and similarly for the variance matrix

\[
\tilde{Q} = Q - SR^{-1}S^T.
\]  

(B.63)

Substituting the expression of \( v(t) \) into the model (B.58) we get

\[
\begin{align*}
x(t + 1) &= Fx(t) + SR^{-1}y(t) + \tilde{v} \\
y(t) &= Cx(t) + w(t)
\end{align*}
\]  

(B.64)

where \( F = A - SR^{-1}C \). The model error \( \tilde{v} \) in the above model is uncorrelated from the measurement noise \( w \), and the cost is that we had to add an output-injection term \( SR^{-1}y(t) \).
Prediction step
Suppose at some point in time we are given a current estimate for the state \( \hat{x}(t|t) \) and a corresponding estimate of the variance of the model error \( P(t|t) = E[\hat{x}(t)\hat{x}(t)^T] \) where \( \hat{x} = x - \hat{x} \). At the initial time \( t_0 \) we can take \( \hat{x}(t_0|t_0) = x_0 \) with some bona-fide variance matrix. Then it is immediate to compute
\[
\hat{x}(t + 1|t) = F\hat{x}(t|t) + SR^{-1}y(t) + \hat{E}[\hat{v}(t)|H_t(y)]
\] (B.65)
where the last term is zero since \( \hat{v}(t) \perp x(s), \forall x \leq t \) and \( \hat{v}(t) \perp w(s), \forall s \) and therefore \( \hat{v}(t) \perp y(s), \forall s \leq t \). The estimation error is therefore
\[
\hat{x}(t + 1|t) = F\hat{x}(t|t) + \hat{v}(t)
\] (B.66)
where the sum is an orthogonal sum, and therefore it is trivial to compute the variance as
\[
P(t + 1|t) = FP(t|t)F^T + \hat{Q}.
\] (B.67)

Update step
Once a new measurement is acquired, we can take into account the new measurement. The update is defined as \( \hat{x}(t + 1|t + 1) = \hat{E}[x(t + 1)|H_{t+1}(y)] \). Now, as we have seen in section B.1.4, we can decompose the span of the measurements into the orthogonal sum
\[
H_{t+1}(y) = H_t(y) + \{e(t + 1)\}
\] (B.68)
where \( e(t + 1) \doteq y(t + 1) - \hat{E}[y(t + 1)|H_t(y)] \) is the innovation process. Therefore, we have
\[
\hat{x}(t + 1|t + 1) = \hat{E}[x(t + 1)|H_t(y)] + \hat{E}[x(t + 1)|e(t + 1)]
\] (B.69)
where the last term can be computed using the results from section B.1:
\[
\hat{x}(t + 1|t + 1) = \hat{x}(t + 1|t) + L(t + 1)e(t + 1)
\] (B.70)
where \( L(t + 1) \doteq \Sigma_{xe}(t + 1)\Sigma_e^{-1}(t + 1) \) is called the Kalman gain. Substituting the expression for the innovation we have
\[
\hat{x}(t + 1|t + 1) = \hat{x}(t + 1|t) + L(t + 1) (y(t + 1) - C\hat{x}(t + 1|t))
\] (B.71)
from which we see that the update consists in a linear correction weighted by the Kalman gain.

Computation of the gain
In order to compute the gain \( L(t + 1) \doteq \Sigma_{xe}(t + 1)\Sigma_e^{-1}(t + 1) \) we derive an alternative expression for the innovation:
\[
e(t + 1) = y(t + 1) - Cx(t + 1) + Cx(t + 1) - C\hat{x}(t + 1|t) = w(t + 1) + C\hat{x}(t + 1|t)
\] (B.72)
from which it is immediate to compute
\[
\Sigma_{xe}(t + 1) = P(t + 1|t)C^T.
\] (B.73)
Similarly we can derive the variance of the innovation \( \Lambda(t + 1) \):
\[
\Lambda(t + 1) \doteq \Sigma_e(t + 1) = CP(t + 1|t)C^T + R
\] (B.74)
and therefore the Kalman gain is
\[
L(t + 1) = P(t + 1|t)C^T\Lambda^{-1}(t + 1).
\] (B.75)
B.6. ASYMPTOTIC PROPERTIES

Variance update

From the update of the estimation error
\[ \hat{x}(t+1|t+1) = \hat{x}(t+1|t) - L(t+1)e(t+1) \] (B.76)
we can easily compute the update for the variance. We first observe that \( \hat{x}(t+1|t+1) \) is by
definition orthogonal to \( H_{t+1}(y) \), while the correction term \( L(t+1)e(t+1) \) is contained in the
history of the innovation, which is by construction equal to the history of the process \( y : H_{t+1}(y) \).
Then it is immediate to see that
\[ P(t+1|t+1) = P(t+1|t) - L(t+1)\Lambda(t+1)\Lambda^T(t+1). \] (B.77)
The above equation is not convenient for computational purposes, since it does not guarantee that
the updated variance is a symmetric matrix. An alternative form of the above that does guarantee
symmetry of the result is
\[ P(t+1|t) = \Gamma(t+1)P(t+1|t)\Gamma(t+1)^T + L(t+1)RL(t+1)^T \] (B.78)
where \( \Gamma(t+1) = I - L(t+1)C \). The last equation is in the form of a discrete Riccati equation
(DRE).

Predictor equations

It is possible to combine the two steps above and derive a single model for the one-step predictor.
We summarize the result as follows:
\[ \hat{x}(t+1|t) = A\hat{x}(t|t-1) + K_s(t)\left( y(t) - C\hat{x}(t|t-1) \right) \] (B.79)
\[ P(t+1|t) = FP(t|t-1)\Gamma(t)^TF^T + FL(t)RL(t)^TF^T + Q \] (B.80)
where we have defined
\[ K_s(t) = FL(t) + SR^{-1} \] (B.81)
\[ = (AP(t|t-1)C^T + S)\Lambda^{-1}(t) \] (B.82)

B.6 Asymptotic properties

If we consider time-invariant models (all matrices \( A, C, Q, S, R \) are constant in time), we can study
the asymptotic behavior of the estimator.

Remark B.5. In particular, the dynamics of the estimator depends upon \( P(t) \), the solution of
the DRE of equation (B.78). We want such a solution to converge asymptotically to a small but
non-zero value. In fact, \( P = 0 \) corresponds to a zero gain \( K = 0 \), which indicates that the filter
does take into account the measurements. In such a case we say that the filter is saturated.

We will not get into the details of the results of the asymptotic theory of Kalman filtering.
We will only report the main results, which essentially says that if the realization of the LFDSP
is minimal, then there exists a unique positive-definite fixed-point of the DRE, and the solution
converges to the fixed-point asymptotically. Furthermore the dynamics of the estimation error
is stable (even though the process may be unstable). The Kalman filter converges asymptotically
to the Wiener filter described in section B.2.
Claim B.1. If the pair $\{F, C\}$ is detectable and $(F, \sqrt{Q})$ is stabilizable, then there exists a unique $P \mid P = P^T \geq 0$ fixed point of the DRE (B.78). Furthermore $P(t) \to P$ for all positive semi-definite $P(t_0)$ and $\Gamma = \lim_{t \to \infty} \Gamma(t)$ is stable.

We recall that a $(F, C)$ being detectable means that the unobservable subspace is stable, as well as $(F, \sqrt{Q})$ being stabilizable means that the uncontrollable subspace is stable. The proof of the above claim, as well as other results on the asymptotic properties of the Kalman filter, can be found for instance in [12].