1

The Kalman Filter in Robotics

This chapter provides a gentle introduction to the Kalman filter, a numerical method used to track the evolution of a dynamical process whose state is unobservable but which can be infered from a sequence of noisy measurements. The derivation of the Kalman filter update equations can be greatly simplified by exploiting a few key mathematical properties of Gaussian distributions.

The Kalman filter is a two step-process: First a forecast is made of the linear evolution of the noisy dynamical system, then a measurement of the state is combined with the forecast, in order to produce the final probabilistic estimate of the system's state. As we show, there is a symmetrical version of the Kalman filter update equations, which is easier to derive, and the standard version (using the so-called Kalman gain) which requires a few more algebraic manipulations. There are some numerical differences between both versions which will be explained. At the end we provide references to the literature on the subject.

1.1 Motivation

The Kalman filter is widely used in statistics, aeronautics, engineering, economics, etc., whenever we have a dynamical process whose state can be described by a state variable x (an n-dimensional real vector) evolving in discrete time steps

$$x_1, x_2, \ldots, x_t.$$

We assume that the true value of the state variable is not directly observable: it has to be detected indirectly using a measuring device. The measurements produce a corresponding sequence of values

$$z_1, z_2, \ldots, z_t$$

which are affected by some noise. Our task is to derive from the sequence of measurements the best probabilistic estimate of the true evolution of the

dynamical system. An example could be a robot moving in horizontal plane limited by walls. The position and velocity of the robot at time t is represented by the vector $x_t = (u_t, v_t, \dot{u}_t, \dot{v}_t)^{\mathrm{T}}$, which we infer iteratively from a measurement $z_t = (p_t, q_t)$. The measurement could consist of the distances to the two nearest walls in two orthogonal directions, as determined by a laser range finder. Our objective is to track the probability distribution of the sequence of states $x_1, \ldots, x_t$ from the given sequence of measurements $z_1, \ldots, z_t$, possibly from a sequence of commands $u_1, \ldots, u_t$, and from a given estimate of the initial position and velocity.

The Kalman filter exploits a simplification called the "Markov assumption". When tracking a robot, for example, we can assume that the probability distribution of the state of the robot at time t depends only on the state of the robot at time $t-1$. That is we assume that:

$$p(x_t|x_{t-1}, x_{t-2}, \ldots, x_1) = p(x_t|x_{t-1})$$

It does not matter how the robot reached state x_{t-1}, its future evolution is only governed by this state. Of course, we could extend the state vector x_t so that it includes past experience by encoding the sequence $x_1, \ldots, x_{t-1}$ in one extra number (think of x_t as a list of values encrypted into a single number with a certain precision, for example). The variable x_t would then contain a complete description of its past. However, this is seldom done in practice, and, in many cases, it is pointless. The Markov assumption deals with "memoryless" processes – in our example, a robot which does not keep track of its previous locations, and decides on its next move based purely on its current position and velocity. The best example of a memoryless process is a flying ball whose future path depends only on its current state.

These simple examples set the stage for the introduction of the Kalman filter.

First we consider state evolution. We are interested in dealing with linear processes, that is, given the n-dimensional state variable x_{t-1}, the next state is determined by the expression

$$x_t = Px_{t-1} + Cu_t + q \tag{1.1}$$

where P (for process) is an $n \times n$ matrix prescribing the system's evolution, u_t is an ℓ-dimensional external command, and C (for control) an $n \times \ell$ matrix which describes the effect of the command on the system's state. Finally, q is the noise in the evolution process. Consider, for example, an object in parabolic flight described by its 3D coordinates and velocities (x_t is then of dimension six). The "command" u_t would be the three dimensional gravity vector, and q would be noise from atmospheric turbulence. The matrices P and C encode linearized and discretized versions of Newton's laws. Notice that Eq. 1.1 implicitly embodies the Markov assumption (since u_t is here a mere constant).

We now consider the measurement of the system's state. The k-dimensional measurement z_t is linearly proportional to x_t, that is,

$$z_t = Hx_t + r \tag{1.2}$$

where H is a $k \times n$ matrix and r is the measurement's noise.

We assume that q and r are normally distributed stochastic variables, i.e. Gaussian noise. Were it not for these two sources of randomicity, the system would be deterministic. Given x_0 we could compute exactly the sequence of future states and measurements. But because of the noise, the system becomes probabilistic and we have to operate with the probability distribution of the state's evolution.

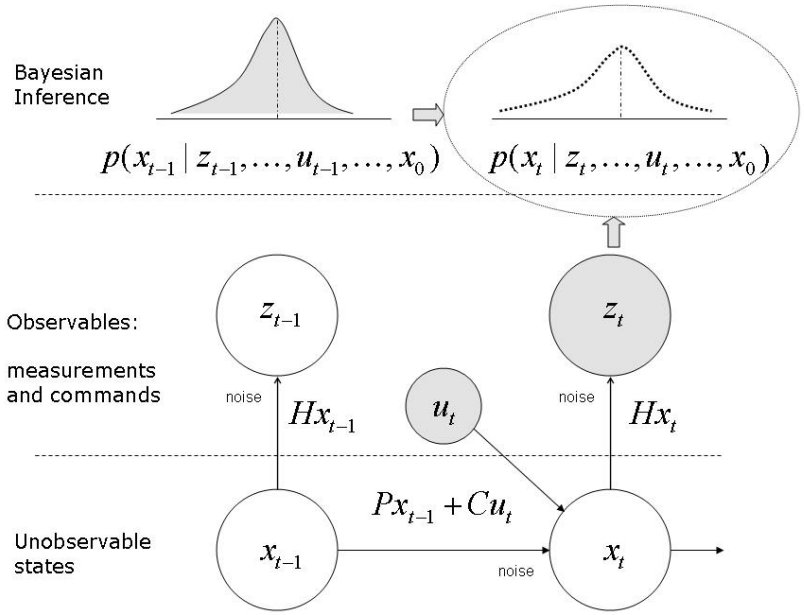

Fig. 1.1. State evolution and measurements of a dynamical system. The states are unobservable, only the sequence of measurements and the commands can be used to infer the probability distribution of the system's state. From the distribution of x_{t-1} (given the sequence of measurements $z_1, \ldots, z_{t-1}$, the commands $u_1, \ldots, u_t$, and x_0), and from the measurement z_t, we want to compute the distribution of x_t (given the observables until t, and the initial state)

Fig. 1.1 shows a diagram of the postulated evolution dynamics. At $t-1$ we have an estimate of the state x_{t-1}, given by a probability distribution

centered around $\hat{x}_{t-1}$. The states x_{t-1} and x_t cannot be observed directly. The transition from state x_{t-1} to state x_t is linear. We can only observe the measurements and the commands. Our task is to compute the probability distribution of x_t from the command u_t, the probability distribution of x_{t-1}, and of the measurement z_t. That is: our initial data in each iteration is $p(x_{t-1}|z_{t-1}, \ldots, z_1, u_{t-1}, \ldots, u_1, x_0)$ and z_t. From this we want to derive $p(x_t|z_t, \ldots, z_1, u_t, \ldots, u_1, x_0)$, and then continue with the next iteration. In the rest of this chapter we do not include the sequence $u_t, \ldots, u-1$ in the conditional probability term, since this sequence is non-stochastic and $p(y|d) = p(y)$, given any constant d and stochastic variable y. Just keep that in mind.

In what follows, we hold the matrices P, C, and H constant over time. We also hold constant the parameters of the Gaussian distributions of q and r. However, any of these constants could change at each step t and we would refer to them as P_t, C_t, and H_t. This additional complexity does not affect the Kalman filter equations at all, since all constants operate from step $t-1$ to step t. By omitting the subindex it is easier to understand the notation and the algebraic steps which follow. At the end, we can rewrite the Kalman filter equations to include such variable dynamical parameters.

1.2 Properties of Gaussian distributions

The Kalman filter is a method for managing and updating Gaussian distributions. We have to consider them first.

Many interesting problems can be described using a normal or Gaussian distribution. This is the kind of probability distribution we obtain when we compute histograms of the height of a population of male persons, of the grades obtained by many pupils, or of average temperatures in the same season. Gaussian distributions seem to be ubiquituous, mostly in complex processes whose outcome can be affected by many variables. For example, a person's size is determined by the combined action of many genes and factors in the environment. Their total effect over a population is to wash-out the individual probability distributions of each contributing factor and produce a Gaussian outcome.

The Central Limit Theorem of probability theory stands out as the main mathematical factor which makes Gaussian distributions so prevalent in the real world. The Theorem states that regardless of the distribution of a stochastic variable X (when its variance is finite), the distribution of the sum of repeated samplings, that is, of the new stocastic variable $X_1 + X_2 + \ldots + X_N$, converges to a Gaussian when N goes to infinity. Under some conditions on the individual distributions of the X_i, they don't even need to have the same distribution. This explains why the world is Gaussian, and it comes as no surprise that the Kalman Filter uses such distributions to model a dynamical system's evolution.

1.2.1 Some definitions

In the one-dimensional case, a Gaussian distribution is defined by the following formula:

$$p(x) = (2\pi)^{-1/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

where μ is the mean value of x and σ its standard deviation. In the n-dimensional case, we have the formula:

$$p(x) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right)$$

where Σ is a symmetric positive definite matrix (the covariance matrix of the Gaussian distribution), μ is the n-dimensional expected value of x, and $|\Sigma|$ is the determinant of the matrix Σ. A stochastic variable x with the above Gaussian distribution is also said to be normal, abbreviated $N(\mu, \Sigma)$

We just mentioned it, but it can be shown that for a normal distribution $N(\mu, \Sigma)$ the expected value of x is μ, that is,

$$\mu = \int xp(x)dx,$$

and also that Σ is the expected value of the exterior product of each $x - \mu$ with itself, that is,

$$\Sigma = \int (x-\mu)(x-\mu)^{\mathrm{T}}p(x)dx,$$

where the integrals run over the domain of definition. The latter integral expression is also used to compute the covariance matrix of distributions other than Gaussian. It provides the parameters for the best Gaussian approximation to the distribution being examined.

1.2.2 Useful properties

By definition, the covariance matrix is symmetric and we always assume that it is invertible. This means that the Gaussian in n-dimensional space "fills" all dimensions and is not confined to a hyperplane of lower dimension. Otherwise, the inverse of the covariance matrix cannot be used as a factor in the exponent of the Gaussian. Remember also, that the product of symmetric matrices is commutative, a fact that simplifies the algebraic considerations wich follow.

Gaussian functions have many interesting properties. We need three of them for the derivation of the Kalman filter.

Property 1: If a stochastic variable x has distribution $N(\mu, A)$, the stochastic variable $Bx + c$ has distribution $N(A\mu + c, BAB^{\mathrm{T}})$, where B is an $n \times n$ matrix and c a constant (excluding such degenerate cases in which BAB^{T} vanishes).

Property 2: The normalized product of two Gaussian distributions $N(a, A)$ and $N(b, B)$ is the Gaussian $N((A+B)^{-1}(Ba+Ab), (A^{-1}+B^{-1})^{-1})$.

Property 3: The normalized convolution of two Gaussians $N(a, A)$ and $N(b, B)$ is the Gaussian $N(a+b, A+B)$.

The three properties can be visualized. Property 1 just tells us that if we linearly transform state space, the mean of the distribution is transformed in the same form, and the covariance matrix transforms quadratically, as BAB^{T}. The second property looks complicated, but it only states that if we multiply Gaussians, their quadratic exponents add. The new quadratic exponent, after factorizing to complete the square, has the form given by the parameters of the new Gaussian in Property 2. And the third property tells us that if have an stochastic variable with a normal distribution, and we add Gaussian noise,for example, the result is also normally distributed.

In order to streamline the discussion of the Kalman filter and make it less obscure the proof of these properties and other important identities is postponed to the mathematical appendix at the end of the chapter. Let us first capture the essence of the Kalman filter and we can go over some algebraic details at the end.

1.3 The Kalman filter equations

We will state the Kalman filter equations before proving them. Remember that the dynamical system being considered evolves according to

$$x_t = Px_{t-1} + Cu_t + q \tag{1.3}$$

and the measurement is such that

$$z_t = Hx_t + r \tag{1.4}$$

with the previously explained meanings of the variables. Let us assume that q has the Gaussian distribution $N(0, Q)$, and r the Gaussian distribution $N(0, R)$.

When we are computing at iteration t, the system state is summarized by the normal distribution $N(\hat{x}_{t-1}, \Sigma_{t-1})$.

We apply the Kalman filter iteratively, each iteration comprising two steps:

- In the *forecast step*, we compute $\bar{x}_t$, a provisional estimate of the expected system state based on the distribution $N(\hat{x}_{t-1}, \Sigma_{t-1})$ obtained in the previous iteration. The bar above the x variable signals that the measurement has not yet been incorporated in the final result. We also compute the provisional covariance matrix $\bar{\Sigma}_t$ of the probability distribution of the forecast.

- In the *information update* step, we combine the forecast with the information obtained from the measurement z_t, weighting each term according to its dispersion, in order to produce an optimal estimate. The result is the Gaussian probability distribution $N(\hat{x}_t, \Sigma_t)$ of x_t.

This iterative process is bootstraped at $t = 1$ with $N(\hat{x}_0, \Sigma_0)$, that is, the probability distribution of the initial state x_0.

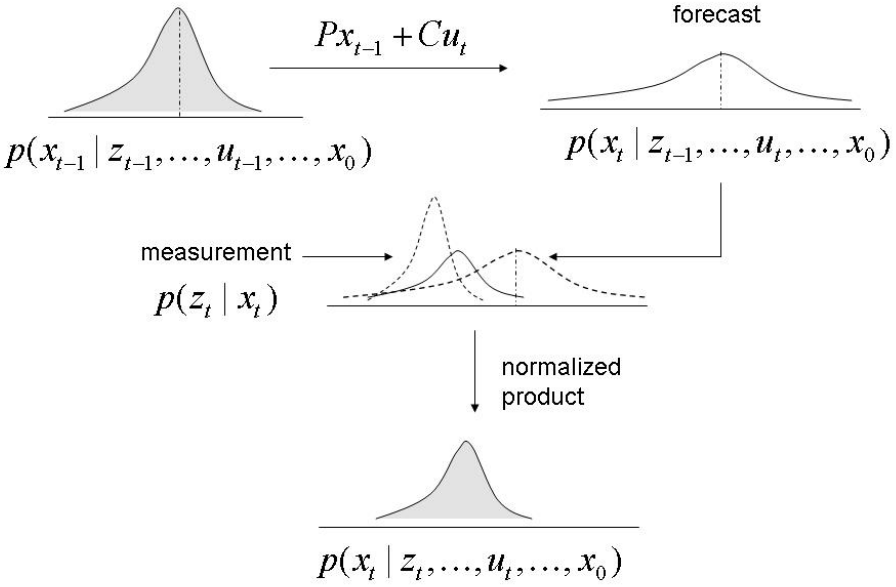

Fig. 1.2. The Kalman filter in a nutshell

Fig. 1.2 shows in a diagram the two steps of the Kalman filter computation using a simple one-dimensional example. This could be a robot moving along a line, for example. We start with the Gaussian describing $p(x_{t-1}|z_{t-1}, \ldots, z_1, x_0)$. Linear evolution, plus noise, produce the forecast, described by a displaced Gaussian with more variance in our example. The measurement provides us with another Gaussian estimate of the possible position of the robot. If forecast and measurement are independent, we just multiply

their probability distributions, obtaining the probability distribution of the state x_t given the sequence of observables until t. Since the measurement and $Px_{t-1} + Cu_t + q$ have Gaussian distributions, the probability of the state x_t follows also a Gaussian distribution (by Property 2 of Gaussians).

The Kalman filter equations are the following:

$$\begin{array}{ll} \text{Forecast} & \left| \begin{array}{l} \bar{x}_t = P\hat{x}_{t-1} + Cu_t \\ \bar{\Sigma}_t = P\Sigma_{t-1}P^{\mathrm{T}} + Q \end{array} \right. \\ \\ \text{Information Update} & \left| \begin{array}{l} \Sigma_t = (\bar{\Sigma}_t^{-1} + H^{\mathrm{T}}R^{-1}H)^{-1} \\ \hat{x}_t = \Sigma_t(\bar{\Sigma}_t^{-1}\bar{x}_t + HR^{-1}z_t) \end{array} \right. \end{array}$$

Cognoscenti of the standard form of the Kalman filter equations might not immediately recognize the third and fourth equations. Expressed in this form, the filter shows that the information fusion step computes just a weighted average of the forecast $\bar{x}_t$ with the information from the measurement z_t, as shown further down. In the next section, we transform this symmetric form of the Kalman filter to its standard form.

1.3.1 The forecast step

The forecast step is the simplest to deal with. According to Property 1, if x_{t-1} has a normal distribution $N(\hat{x}_{t-1}, \Sigma_{t-1})$, then a linear transformation of x_{t-1} as in Eq. 1.3 has a normal distribution $N(P\hat{x}_{t-1} + Bu_t, P\Sigma_{t-1}P^{\mathrm{T}})$. But remember that the estochastic process has additive noise q. Given a stochastic variable y with Gaussian distribution, if Gaussian noise q is added, then the probability of y becomes

$$p(y) = \int p(y-q)p(q)dq$$

This is a convolution of two Gaussians. In our case the probability of $Px_{t-1} + Bu_t$ is governed by the Gaussian $N(P\hat{x}_{t-1} + Bu_t, P\Sigma_{t-1}P^{\mathrm{T}})$ and the probability of q by the Gaussian $N(0, Q)$. The probability of the forecast is the convolution of the two Gaussians, so that by Property 3, the distribution of the result is the Gaussian $N(P\hat{x}_{t-1} + Bu_t, P\Sigma_{t-1}P^{\mathrm{T}} + Q)$. From this we immediately obtain the two first equations in the forecast step of the Kalman filter. We abbreviate the distribution of x_t (given the distribution of the observables until $t-1$) as $N(\bar{x}_t, \bar{\Sigma}_t)$

1.3.2 The information fusion step

Now a measurement z_t is taken. Its distribution is the Gaussian $N(Hx_t, R)$, that is

$$p(z_t|x_t) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(z_t - Hx_t)^{\mathrm{T}}R^{-1}(z_t - Hx_t)\right)$$

We can obtain now the distribution of $p(x_t|z_t, \ldots, z_1, x_0)$ applying the Bayes formula:

$$p(x_t|z_t, \ldots, z_1, x_0) = \alpha p(z_t|x_t)p(x_t|z_{t-1}, \ldots, z_1, x_0)$$

where α is a normalization factor. The distribution of the first factor on the right is $N(Hx_t, R)$ and the distribution of $p(x_t|z_{t-1}, \ldots, z_1, x_0)$ is $N(\bar{x}_t, \bar{\Sigma}_t)$. We can apply the above formula because we consider the measurement independent of the forecast.

The exponents of the two Gaussians in the product above add up to a quadratic in x_t of the form:

$$-\frac{1}{2}\left[x_t^{\mathrm{T}}(\bar{\Sigma}_t^{-1} + H^{\mathrm{T}}R^{-1}H)x_t - 2x_t^{\mathrm{T}}(\bar{\Sigma}_t^{-1}\bar{x}_t + H^{\mathrm{T}}R^{-1}z_t) + ...\right]$$

which can be rewritten in the form

$$-\frac{1}{2}\left[(x_t - \hat{x}_t)^{\mathrm{T}}\Sigma_t^{-1}(x_t - \hat{x}_t)\right] + \gamma$$

with γ a constant and

$$\Sigma_t = (\bar{\Sigma}_t^{-1} + H^{\mathrm{T}}R^{-1}H)^{-1}$$

and mean value

$$\hat{x}_t = \Sigma_t(\bar{\Sigma}_t^{-1}\bar{x}_t + H^{\mathrm{T}}R^{-1}z_t)$$

as the interested reader can verify by direct substitution. The constant γ in the exponent is absorbed as a scale factor for the exponential function. After the multiplication the new Gaussian is normalized so that it covers an area equal to 1 (and so continues being a probability distribution). Both expressions for Σ_t and $\hat{x}_t$correspond directly to the two equations in the information update step of the Kalman filter, as we had to prove.

It is interesting to note that in this symmetrical form of the Kalman filter both forecast and measurement are handled in equal terms. If there is a direct correspondence between measurement and state, that is if H is the identity matrix, the information update step reduces to

$$\Sigma_t = (\bar{\Sigma}_t^{-1} + R^{-1})^{-1},$$

which corresponds to a generalized harmonic mean of the covariances $\bar{\Sigma}_t$ and R, while

$$\hat{x}_t = \Sigma_t(\bar{\Sigma}_t^{-1}\bar{x}_t + R^{-1}z_t)$$

is a harmonic weighted average of forecast and measurement. Curiously, it is as if the function of the matrix H in the Kalman filter equations is to make the covariance matrices $\bar{\Sigma}_t$ and R comensurable, since they have different dimensions.

1.4 Standard form of the Kalman filter equations

In the standard form of the Kalman filter equations, the forecast step is written just as in the previous section. The information update step is written differently, introducing an auxiliary variable K_t, called the Kalman gain.

$$\text{Information Update} \quad \left| \begin{array}{l} K_t = \bar{\Sigma}_t H^{\mathrm{T}} (H \bar{\Sigma}_t H^{\mathrm{T}} + R)^{-1} \\ \Sigma_t = (I - K_t H) \bar{\Sigma}_t \\ \hat{x}_t \ = \bar{x}_t + K_t (z_t - H \bar{x}_t) \end{array} \right.$$

What we have to show is that these expresions are equivalent to the symmetric form of the Kalman filter equations proved in the previous section. There is one important difference, though. Whereas the symmetric form of the equations require three matrix inversions, the standard form requires just one. Also, the covariance matrix in state space is of dimension $n \times n$, whereas the matrices in measurement space are of dimension $k \times k$. Usually k is smaller than n. The standard form is therefore numerically more efficient.

To show the equivalence, consider first the updated covariance matrix Σ_t. We can plug-in the definition of K_t in the first equation into the second equation above to obtain

$$\Sigma_t = \bar{\Sigma}_t - \bar{\Sigma}_t H^{\mathrm{T}} (H \bar{\Sigma}_t H^{\mathrm{T}} + R)^{-1} H \bar{\Sigma}_t$$

We want to show that this expression is identical to the one derived before for the symmetrical Kalman filter:

$$\Sigma_t = (\bar{\Sigma}_t^{-1} + H^{\mathrm{T}} R^{-1} H)^{-1}$$

In order to prove this, we start with the expression to be proved

$$(\bar{\Sigma}_t^{-1} + H^{\mathrm{T}} R^{-1} H)^{-1} = \bar{\Sigma}_t - \bar{\Sigma}_t H^{\mathrm{T}} (H \bar{\Sigma}_t H^{\mathrm{T}} + R)^{-1} H \bar{\Sigma}_t \tag{1.5}$$

and derive equivalent expressions through algebraic manipulation. We multiply each side of the expression above by $(\bar{\Sigma}_t^{-1} + H^{\mathrm{T}} R^{-1} H)$ to obtain the first of the equivalent expressions which follow:

$$\begin{aligned} I &= \bar{\Sigma}_t (\bar{\Sigma}_t^{-1} + H^{\mathrm{T}} R^{-1} H) - \bar{\Sigma}_t V^{\mathrm{T}} (R + H \bar{\Sigma}_t H^{\mathrm{T}})^{-1} H \bar{\Sigma}_t (\bar{\Sigma}_t^{-1} + H^{\mathrm{T}} R^{-1} H) \\ I &= (I + \bar{\Sigma}_t H^{\mathrm{T}} R^{-1} H) - \bar{\Sigma}_t H^{\mathrm{T}} (R + H \bar{\Sigma}_t H^{\mathrm{T}})^{-1} (R + H \bar{\Sigma}_t H^{\mathrm{T}}) R^{-1} H \\ 0 &= \bar{\Sigma}_t H^{\mathrm{T}} R^{-1} H - \bar{\Sigma}_t H^{\mathrm{T}} R^{-1} H \end{aligned}$$

Since all formulas are equivalent, and the last one is obviously true, all of the formulas are true. Alternatively, one can start from the last expression, go backwards and prove that both definitions of Σ_t are equivalent. Eq. 1.5 is known as the matrix inversion lemma, which we could have used directly.

Now consider the last update equation in the standard Kalman filter. It can be rewritten as

$$\hat{x}_t = (I - K_t H) \bar{x}_t + K_t z_t \tag{1.6}$$

We know from the second Kalman equation of the information update step (just proved) that

$$\Sigma_t = (I - K_t H)\bar{\Sigma}_t.$$

Therefore

$$\Sigma_t \bar{\Sigma}_t^{-1} = (I - K_t H)$$

and from this we can transform Eq. 1.6 into

$$\hat{x}_t = \Sigma_t \bar{\Sigma}_t^{-1} \bar{x}_t + K_t z_t$$

Compare this expression with the corresponding one in the symmetric form of the Kalman filter: $\hat{x}_t = \Sigma_t(\bar{\Sigma}_t^{-1}\bar{x}_t + HR^{-1}z_t)$. Comparing both expressions we see that we only have to prove $K_t = \Sigma_t H^{\mathrm{T}} R^{-1}$. Let us define $U = \Sigma_t H^{\mathrm{T}} R^{-1}$. Multiplying both sides of this expression by the same factor we obtain

$$\begin{aligned} U(H\bar{\Sigma}_t H^{\mathrm{T}} + R) &= \Sigma_t H^{\mathrm{T}} R^{-1}(H\bar{\Sigma}_t H^{\mathrm{T}} + R) \\ &= \Sigma_t(H^{\mathrm{T}} R^{-1} H\bar{\Sigma}_t H^{\mathrm{T}} + H^{\mathrm{T}}) \\ &= \Sigma_t(H^{\mathrm{T}} R^{-1} H + \bar{\Sigma}_t^{-1})\bar{\Sigma}_t H^{\mathrm{T}} \\ &= \bar{\Sigma}_t H^{\mathrm{T}} \end{aligned}$$

since $(H^{\mathrm{T}} R^{-1} H + \bar{\Sigma}_t^{-1}) = \Sigma_t^{-1}$. Therefore

$$U = \bar{\Sigma}_t H^{\mathrm{T}}(H\bar{\Sigma}_t H^{\mathrm{T}} + R)^{-1}$$

and since this is the definition of K_t, we conclude that $K_t = U = \Sigma_t H^{\mathrm{T}} R^{-1}$, as we wanted to show.

This concludes our proof of the equivalence of both versions of the Kalman filter equations. For the benefit of the reader, we state them here again, allowing for evolution parameters which change from step to step. That is, we write instead of P the matrix P_t, instead of C the matrix C_t, and so on. This is the form typically used in many textbooks.

Kalman Filter

Forecast

$$\left|\begin{array}{l} \bar{x}_t = P_t \hat{x}_{t-1} + C_t u_t \\ \bar{\Sigma}_t = P_t \Sigma_{t-1} P_t^{\mathrm{T}} + Q_t \end{array}\right.$$

Information Update

$$\left|\begin{array}{l} K_t = \bar{\Sigma}_t H_t^{\mathrm{T}}(H_t \bar{\Sigma}_t H_t^{\mathrm{T}} + R_t)^{-1} \\ \Sigma_t = (I - K_t H_t)\bar{\Sigma}_t \\ \hat{x}_t = \bar{x}_t + K_t(z_t - H_t \bar{x}_t) \end{array}\right.$$

1.5 Discussion and numerical considerations

1.6 Mathematical Appendix

This section provides proofs for the algebraic identities and properties of Gaussians used in the previous sections.

Proposition 1.1. *Property 1. If a stochastic variable x has distribution $N(\mu, A)$, the stochastic variable $Bx + c$ has distribution $N(A\mu + c, B^{\mathrm{T}}AB)$, where B is an $n \times n$ matrix and c a constant (excluding degenerate cases for which $B^{\mathrm{T}}AB = 0$).*

Proof. Let us denote by μ^e and A^e the expected value and covariance of the stochastic variable $Bx + c$.

$$\mu^e = \int (Bx + c)p(x)dx = B \int xp(x)dx + c = B\mu + c$$

and

$$\begin{aligned} A^e &= \int (Bx + c - \mu^e)(Bx + c - \mu^e)^{\mathrm{T}} p(x)dx \\ &= \int (Bx - B\mu)(Bx - B\mu)^{\mathrm{T}} p(x)dx \\ &= B \left(\int (x - \mu)(x - \mu)^{\mathrm{T}} p(x)dx \right) B^{\mathrm{T}} \\ &= BAB^{\mathrm{T}} \end{aligned}$$

This proves that the distribution of $Bx + c$ is $N(A\mu + c, B^{\mathrm{T}}AB)$.

Proposition 1.2 (Property 2). *The normalized product of two Gaussian distributions $N(a, A)$ and $N(b, B)$ is Gaussian, i.e. $N((A + B)^{-1}(Ba + Ab), (A^{-1} + B^{-1})^{-1})$*

Proof: If we multiply two Gaussians, we have to add the exponents of each one (disregarding the constant normalization terms). The result of adding two quadratic functions is a quadratic, and from this we see immediately that the product of two Gaussians is Gaussian. The sum of the exponents of the two Gaussians given above can be written as

$$-1/2((x-c)^{\mathrm{T}}C^{-1}(x-c)) = -1/2((x-a)^{\mathrm{T}}A^{-1}(x-a) + (x-b)^{\mathrm{T}}B^{-1}(x-b)) \tag{1.7}$$

where the matrix C and the mean value c can be found factorizing the right hand side so as to "complete the square".

Given any matrix expression of the form $x^{\mathrm{T}}Px + 2x^{\mathrm{T}}Qz + r$, where x and z are n-dimensional vectors and P and Q invertible symmetric matrices of appropriate dimension, while r is any constant, it can be factorized as $(x - P^{-1}Q)^{\mathrm{T}}P(x - P^{-1}Q) + r'$. Here r' is a new constant whose exact value is unimportant since as part of the exponent of e in the Gaussian, it represents just a scaling factor.

In Eq. 1.7 we can simplify the sum of exponents to

$$-1/2((x-c)^{\mathrm{T}}C^{-1}(x-c)) = -1/2(x\mathrm{T}(A^{-1} + B^{-1})x + 2x\mathrm{T}(A^{-1} + B^{-1}b) + r)$$

absorbing any constants into r and making using the symmetry of A and B. From this, it is to see immediately that the matrix C^{-1}, the factor in the quadratic term, is given by the sums of the factors in the quadratic terms in the right hand side, that is, $C^{-1} = (A^{-1} + B^{-1})$ and the mean c is given by

$$c = (A^{-1} + B^{-1})^{-1}(A^{-1}a + B^{-1}b)$$

Equivalently, using the fact that the product of symmetric matrices is commutative, we can multiply the first factor in the right hand side of the equation above by $(AB)^{-1}$ and the second factor by (AB) obtaining:

$$c = (B + A)^{-1}(Ba + Ab)$$

□

Proposition 1.3 (Property 3). *The convolution of two Gaussians $N(a, A)$ and $N(b, B)$ is the Gaussian $N(a + b, A + B)$.*

Proof. We apply the convolution theorem for Fourier transforms. We take as given that the Fourier transform of a Gaussian $N(a, A)$ is the Gaussian $N(iA^{-1}a, A^{-1})$, where $i = \sqrt{-1}$. That is

$$F(N(a, A)) = N(iA^{-1}a, A^{-1})$$
$$F(N(b, B)) = N(iB^{-1}b, B^{-1})$$

According to the convolution theorem, the original convolution is represented in frequency space by the product of the Fourier transforms. Applying Property 2, the the multiplication of twwo Gaussians gives us the new Gaussian

$$N(i(A + B)^{-1}(AA^{-1}a + BB^{-1}b), (A + B)^{-1})$$

which we can abbreviate as $N(iC(a+b), C)$, where $C = (A+B)^{-1}$. The inverse Fourier transform of a Gaussian $N(d, D)$ is the Gaussian $N(-iD^{-1}d, D^{-1})$. Inverting the Fourier transform of the product of Gaussians we obtain the result of the original convolution:

$$F^{-1}(N(iC(a + b), C) = N(a + b, A + B)$$

which is $N(a + b, A + B)$ as stated above.

1.7 History and references