Kalman Filters: Derivation

Burak Bayramlı

July 10, 2011

Compiling an exhaustive list of fields where Kalman Filters (KF) are in use would be a formiddable task: The list could start with navigation, radar tracking, satellite orbit computation [9], continue with sonar ranging, stock price prediction, and many other fields where there is a dynamical system present. When the *Eagle* module of the Apollo 11 mission landed on the moon, it is widely known that it did so with the help of a Kalman Filter. KF was first proposed in a seminal paper by [7], later Kalman and Bucy [8] - the method's first application setting was space tracking.

Kalman Filters, or in general, SSM methods try to model dynamic data using two-level graphic model. In the first level, there are hidden states, one per time step, which are random variables distributed as Gaussians. An outside observer (modeler) does not see these states directly, instead sees them through an "observed" variable whose outputs are assumed to be distorted by another Gaussian distribution, refered to as white noise. In each time step, it is assumed the hidden variables evolve linearly, in linear algebra this means the hidden state is multiplied by a constant matrix A. Mathematically,

$$x_{t+1} \quad = \quad Ax_t + v_t$$

where

- $v_0, v_1, ...$ is white Gaussian noise with $\mathbf{E}v_t = 0$ and $cov\ v_t = Q$
- $x_t \sim N(\mu_t, \Sigma_t)$.

In order to model the output variables, we multiply hidden states with another constant matrix and add another Gaussian white noise. All together our SSM looks as follows:

$$x_{t+1} = Ax_t + v_t \tag{1}$$

$$y_t = Cx_t + w_t \tag{2}$$

where $w_0, w_1, ...$ are white Gaussian noise with $\mathbf{E}w_t = 0$ and $cov\ \ w_t = R$. Random variables w_t and v_t are independent of x_t and y_t, and from eachother, this assumption is put forth for simplicity. The dimensions of each variable are:

- $x(t) \in \mathbb{R}^n$ is the state
- $y(t) \in \mathbb{R}^p$ is the observed output
- $v(t) \in \mathbb{R}^n$ is process noise
- $w(t) \in \mathbb{R}^p$ is measurement noise

Kalman Filters carry the same conditional independence properties that characterized HMMs - given a state for time t, the future is conditionally independent from the past, hence KF also carries the first-order Markovian property just like HMMs. Structurally there are many similarities between two models as well and its methods for inference, as we shall soon see, the main difference being that HMM uses multinomial distribution for hidden states whereas KF uses Gaussians.

Inference with Kalman Filters means estimating the posterior probability given an observed output sequence. There are two tasks here, filtering and smoothing. Filtering is formulized by $p(x_t|y_0, ..., y_t)$ which tries to calculate posterior probability for hidden states using measurements up to time t. In smoothing, we will calculate the same probability using measurements taken after time t, which gives us $p(x_t|y_{t+1}, ..., y_T)$. We also hope to formulize recursive formulas for both filtering and smoothing.

Both filtering and smoothing have direct consequences in terms of real-life applications. KFs' natural roots are embedded in space tracking, the method started as an "online" algorithm trying to track down targets using noisy measurements, as each new measurement is received. In this sense, KF can be seen as an estimator trying to guess a "hidden" position at time t using measurements up to time t. Smoothing, on the other hand, can be thought of a global "corrector" - once we have better knowledge on hand, through smoothing we can go back to old estimations and correct previous online estimates using backward recursion.

1 Filtering

Following the notation and the derivation steps used by Bishop and Jordan [6], let's first denote

- $\hat{x}_t^t \triangleq \mathbf{E}[x_t|y_0, ..., y_t]$
- $P_t^t \triangleq \mathbf{E}[(x_t - \hat{x}_{t|t})(x_t - \hat{x}_{t|t})'|y_0, ..., y_t]$

Our first goal is going to be trying to calculate $\hat{x}_{t+1}^{t+1}$ and P_{t+1}^{t+1} using a new measurement y_{t+1}. For this, we need to "reverse the arrow" so to speak, in other words formulize the distribution where x_t has conditional dependence on y_t instead of other way around, which is what the base equations state right now.

If we take expectation of both sides of (1)

$$\mathbf{E}x_{t+1} = \hat{x}_{t+1} = A\mu_t = A\hat{x}_t$$

Taking covariance of both sides of (1) and denoting P_t as $cov\, x(t)$

$$P_{t+1} = AP_tA' + Q$$

This transition is called a "time update". It will allow us to propagate the normal distribution at time t to time t_1. The formulas that contain given y's use the same technique;

$$\hat{x}_{t+1}^t = Ax_t^t$$
$$P_{t+1}^t = AP_t^tA' + Q$$

$$\begin{aligned} y_{t+1} &= Cx_{t+1} + w_t \\ \mathbf{E}[y_{t+1}|y_0, ..., y_t] &= \mathbf{E}[Cx_{t+1} + w_t|y_0, ..., y_t] \\ \hat{y}_{t+1}^t &= C\hat{x}_{t+1} \end{aligned}$$

Similarly for covariance

$$E[(y_{t+1} - \hat{y}_{t+1}^t)(y_{t+1} - \hat{y}_{t+1}^t)'|y_0, ..., y_t] = CP_{t+1}^tC' + R$$

Now the harder task of reversing the arrow: If our goal is formulizing $p(x_t|y_t)$, then we need to derive the joint distribution between both variables. Addition of Gaussians is another Gaussian, we can surmize the joint probability $p(x_t, y_t)$ will be one huge Gaussian.

In order to define the joint normal distribution of x_t and y_t, we need to write the joint mean and covariance of this big Gaussian. Writing down the mean will be easy, covariance will be little harder to derive. We can use a trick [10] where we write $y_t = Cx_t + w_t$ as $z = Hu$ where

$$z = \begin{bmatrix} x_t \\ y_t \end{bmatrix}, H = \begin{bmatrix} I & 0 \\ C & I \end{bmatrix} u = \begin{bmatrix} x_t \\ w_t \end{bmatrix},$$

Now we need to take the covariance of the simpler equation:

$$\begin{aligned} cov(z) &= Hcov(u)H' \\ cov(u) &= \begin{bmatrix} P_t & 0 \\ 0 & R \end{bmatrix} \end{aligned}$$

The full product is:

$$\begin{bmatrix} I & 0 \\ C & I \end{bmatrix} \begin{bmatrix} P_t & 0 \\ 0 & R \end{bmatrix} \begin{bmatrix} I & C' \\ 0 & I \end{bmatrix}$$

which results in

$$\left[\begin{array}{cc} P_t & P_tC' \\ CP_t & CP_tC'+R \end{array}\right]$$

We can also re-write this for conditional and with the mean;

$$\left[\begin{array}{c} \hat{x}^t_t \\ C\hat{x}^t_t \end{array}\right] \; and \; \left[\begin{array}{cc} P^t_t & P^t_tC' \\ CP^t_t & CP^t_tC'+R \end{array}\right] \tag{3}$$

Same for the joint distribution for x_{t+1}, y_{t+1}.

$$\left[\begin{array}{c} \hat{x}^t_{t+1} \\ C\hat{x}^t_{t+1} \end{array}\right] \; and \; \left[\begin{array}{cc} P^t_{t+1} & P^t_{t+1}C' \\ CP^t_{t+1} & CP^t_{t+1}C'+R \end{array}\right] \tag{4}$$

Now in order to get statements for mean and variance for x^{t+1}_{t+1}, we need to understand partitioned Gaussians. This will guide us when we use portions of (4) and deriving the final equation [6].

An n dimensional Gaussian distribution can be partitioned into p and q dimensional sub-distributions where $n = p + q$. Hence we can say,

$$\mu = \left[\begin{array}{c} \mu_1 \\ \mu_2 \end{array}\right] \Sigma = \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array}\right] \tag{5}$$

$$\begin{aligned} p(x|\mu,\Sigma) \quad &= \quad \frac{1}{(2\pi)^{(p+q)/2}|\Sigma|^{1/2}} \\ &\quad\; exp\Big\{-\frac{1}{2}\left(\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array}\right)'\Big\}\left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array}\right]^{-1}\left(\begin{array}{c} x_1-\mu_1 \\ x_2-\mu_2 \end{array}\right)\Big\} \end{aligned}$$

After much algebra, we can obtain equation for $p(x_1|x_2)$. From this, we can get *conditioned* μ and Σ, as

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) \qquad (6) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

Now plugging elements from (6) into (4) according to placement in (5) we can formulize $\hat{x}^{t+1}_{t+1}$ and P^{t+1}_{t+1}.

$$\begin{aligned} \hat{x}^{t+1}_{t+1} &= \hat{x}^t_{t+1} + P^t_{t+1}C'(CP^t_{t+1}C' + \Sigma_w)^{-1} \\ &\quad\; (y_{t+1} - C\hat{x}^t_{t+1}) \\ P^{t+1}_{t+1} &= P^t_{t+1} - P^t_{t+1}C'(CP^t_{t+1}C'+R)^{-1}CP^t_{t+1} \end{aligned}$$

If we declare $K_t \triangleq P^t_{t+1}C'(CP^t_{t+1}C' + \Sigma_w)^{-1}$ then,

$$\begin{aligned} \hat{x}^{t+1}_{t+1} &= \hat{x}^t_{t+1} + K_t(y_{t+1} - C\hat{x}^t_{t+1}) \qquad (7) \\ P^{t+1}_{t+1} &= P^t_{t+1} - K_tCP^t_{t+1} \qquad (8) \end{aligned}$$

References

[1] Rabiner L. R. , *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of IEEE vol. 77, no. 2, pp. 257-286, 1989.

[2] Roweis S. and Z. Ghahramani, *A Unifying Review of Linear Gaussian Models* , Neural Computation 11(2):305–345, 1999.

[3] Ghahramani Z., H. E. Hinton, *Parameter Estimation for Linear Dynamical Systems* , Technical Report CRG-TR-96-2 [`ftp://ftp.cs.toronto.edu/pub/zoubin/tr96-2.ps.gz`], Department of Computer Science, University of Toronto, 1996.

[4] Ghahramani Z., H. E. Hinton, *Switching State Space Models*, Technical Report CRG-TR-96-3, Dept. Comp. Sci., Univ. Toronto, 1996.

[5] Shumway R., H. S. Stoffer *Time series analysis and its applications 2nd Edition*, New York, Springer, (Springer texts in statistics), 2000.

[6] Jordan M. I. , C. Bishop *An Introduction to Graphical Models*, Not yet published, 2000.

[7] Kalman R. E., *A New Approach to Linear Filtering and Prediction Problems*, Transactions of the ASME-Journal of Basic Engineering, 82 (Series D): 35-45, 1960.

[8] Kalman, R.E. and R.S. Bucy, *New results in filtering and prediction theory*, Trans. ASME J. Basic Eng., 83, 95-108, 1961.

[9] Welling, M., *The Kalman Filter - Lecture Tutorial*, California Institute of Technology, 2008.

[10] Lall, S., *Modern Control 2 Lecture Notes*, Stanford University, 2006.

[11] Tsay, R. S., *Time Series and Forecasting: Brief History and Future Research,*, Journal of the American Statistical Association, Vol. 95, No. 450., pp. 638-643., Jun 2000.

[12] Shi S. and A. S. Weigend., *Taking Time Seriously: Hidden Markov Experts Applied to Financial Engineering*. In Proceedings of the IEEE/IAFE 1997 Conference on Computational Intelligence for Financial Engineering , pages 244-252. IEEE, 1997.

[13] Mandelbrot B. B., *The variation of certain speculative prices, Journal of Business*, XXXVI (1963), pp. 392-417, 1963.