



December 3, 2010

The Kalman Filter Explained

Tristan Fletcher

www.cs.ucl.ac.uk/staff/T.Fletcher/

1 Introduction

The aim of this document is to derive the filtering equations for the simplest Linear Dynamical System case, the Kalman Filter, outline the filter's implementation, do a similar thing for the smoothing equations and conclude with parameter learning in an LDS (calibrating the Kalman Filter).

The document is based closely on Bishop [1] and Ghahramani's [2] texts, but is more suitable for those who wish to understand every aspect of the mathematics required and see how it all comes together in a procedural sense.

2 Model Specification

The simplest form of Linear Dynamical System (LDS) models a discrete time process where a latent variable h is updated every time step by a constant linear state transition A with the addition of zero-mean Gaussian noise η^h :

$$\begin{aligned} h_t &= Ah_{t-1} + \eta_t^h \quad \text{where } \eta_t^h \sim N(0, \Sigma_H) \\ &\Rightarrow p(h_t|h_{t-1}) \sim N(Ah_{t-1}, \Sigma_H) \end{aligned} \tag{2.1}$$

This latent variable is observed through a constant linear function of the latent state B also subject to zero-mean Gaussian noise η^v :

$$\begin{aligned} v_t &= Bh_t + \eta_t^v \quad \text{where } \eta_t^v \sim N(0, \Sigma_V) \\ &\Rightarrow p(v_t|h_t) \sim N(Bh_t, \Sigma_V) \end{aligned} \tag{2.2}$$

We wish to infer the probability distribution of h_t given the observations up to that point in time $v_{1:t}$, i.e. $p(h_t|v_{1:t})$, which can be expressed recursively.

Starting with an initial distribution for our latent variable given the first observation:

$$p(h_1|v_1) \propto p(v_1|h_1)p(h_1)$$

and the assumption that h_1 has a Gaussian distribution:

$$p(h_1) \sim N(\mu_0, \sigma_0^2)$$

values for $p(h_t|v_{1:t})$ for subsequent values of t can be found by iteration:

$$\begin{aligned}
p(h_t|v_{1:t}) &= p(h_t|v_t, v_{1:t-1}) \\
&= \frac{p(h_t, v_t|v_{1:t-1})}{p(v_t|v_{1:t-1})} \\
&\propto p(h_t, v_t|v_{1:t-1}) \\
&= \int_{h_{t-1}} p(h_t, v_t|v_{1:t-1}, h_{t-1})p(h_{t-1}|v_{1:t-1}) \\
&= \int_{h_{t-1}} p(v_t|v_{1:t-1}, h_{t-1}, h_t)p(h_t|h_{t-1}, v_{1:t-1})p(h_{t-1}|v_{1:t-1}) \\
h_t \perp v_{1:t-1} | h_{t-1} &\Rightarrow \int_{h_{t-1}} p(v_t|v_{1:t-1}, h_{t-1}, h_t)p(h_t|h_{t-1})p(h_{t-1}|v_{1:t-1}) \\
v_t \perp h_{t-1} | v_{1:t-1}, h_t &\Rightarrow \int_{h_{t-1}} p(v_t|v_{1:t-1}, h_t)p(h_t|h_{t-1})p(h_{t-1}|v_{1:t-1}) \\
&= \int_{h_{t-1}} p(v_t|h_t)p(h_t|h_{t-1})p(h_{t-1}|v_{1:t-1}) \tag{2.3}
\end{aligned}$$

The fact that the distributions described in (2.1) and (2.2) are both Gaussian and the operations in (2.3) of multiplication and then integration will yield Gaussians when performed on Gaussians, means that we know $p(h_t|v_{1:t})$ will itself be a Gaussian:

$$p(h_t|v_{1:t}) \sim N(\mu_t, \sigma_t^2) \tag{2.4}$$

and the task is to derive the expressions for μ_t and σ_t^2 .

3 Deriving the state estimate variance

If we define $\Delta h \equiv h - \mathbb{E}[\hat{h}]$, where h denotes the actual value of the latent variable, \hat{h} is its estimated value, $\mathbb{E}[\hat{h}]$ is the expected value of this estimate and F as the covariance of the error estimate then:

$$\begin{aligned}
F_{t|t-1} &= \mathbb{E}[\Delta h_t \Delta h_t^T] \\
&= \mathbb{E}[(A\Delta h_{t-1} + \Delta \eta_t^h)(A\Delta h_{t-1} + \Delta \eta_t^h)^T] \\
&= \mathbb{E}[A\Delta h_{t-1}\Delta h_{t-1}^T A^T + \Delta \eta_t^h \Delta \eta_t^{hT} + \Delta \eta_t^h A^T \Delta h_{t-1}^T + A\Delta h_{t-1} \Delta \eta_t^{hT}] \\
&= A\mathbb{E}[\Delta h_{t-1}\Delta h_{t-1}^T] A^T + \mathbb{E}[\Delta \eta_t^h \Delta \eta_t^{hT}] \\
&= AF_{t-1:t-1}A^T + \Sigma_H \tag{3.1}
\end{aligned}$$

The subscript in $F_{t|t-1}$ denotes the fact that this is F 's value before an observation is made at time t (i.e. its *a priori* value) while $F_{t|t}$ would denote

a value for $F_{t|t}$ after an observation is made (its *posterior* value). This more informative notation allows the update equation in (2.1) to be expressed as follows:

$$\hat{h}_{t|t-1} = A\hat{h}_{t-1|t-1} \quad (3.2)$$

Once we have an observation (and are therefore dealing with *posterior* values), we can define ϵ_t as the difference between the observation we'd expect to see given our estimate of the latent state (its *a priori* value) and the one actually observed, i.e.:

$$\epsilon_t = v_t - B\hat{h}_{t|t-1} \quad (3.3)$$

Now that we have an observation, if we wish to add a correction to our *a priori* estimate that is proportional to the error ϵ_t we can use a coefficient κ :

$$\hat{h}_{t|t} = \hat{h}_{t|t-1} + \kappa\epsilon_t \quad (3.4)$$

This allows us to express $F_{t|t}$ recursively:

$$\begin{aligned} F_{t|t} &= Cov(h_t - \hat{h}_{t|t}) \\ &= Cov(h_t - (\hat{h}_{t|t-1} + \kappa\epsilon_t)) \\ &= Cov(h_t - (\hat{h}_{t|t-1} + \kappa(v_t - B\hat{h}_{t|t-1}))) \\ &= Cov(h_t - (\hat{h}_{t|t-1} + \kappa(Bh_t + \eta_t^v - B\hat{h}_{t|t-1}))) \\ &= Cov(h_t - \hat{h}_{t|t-1} - \kappa Bh_t - \kappa\eta_t^v + \kappa B\hat{h}_{t|t-1}) \\ &= Cov((I - \kappa B)(h_t - \hat{h}_{t|t-1}) - \kappa\eta_t^v) \\ &= Cov((I - \kappa B)(h_t - \hat{h}_{t|t-1})) + Cov(\kappa\eta_t^v) \\ &= (I - \kappa B)Cov(h_t - \hat{h}_{t|t-1})(I - \kappa B)^T + \kappa Cov(\eta_t^v)\kappa^T \\ &= (I - \kappa B)F_{t|t-1}(I - \kappa B)^T + \kappa\Sigma_V\kappa^T \\ &= (F_{t|t-1} - \kappa BF_{t|t-1})(I - \kappa B)^T + \kappa\Sigma_V\kappa^T \\ &= F_{t|t-1} - \kappa BF_{t|t-1} - F_{t|t-1}(\kappa B)^T + \kappa BF_{t|t-1}(\kappa B)^T + \kappa\Sigma_V\kappa^T \\ &= F_{t|t-1} - \kappa BF_{t|t-1} - F_{t|t-1}B^T\kappa^T + \kappa(BF_{t|t-1}B^T + \Sigma_V)\kappa^T \quad (3.5) \end{aligned}$$

If we define the innovation variance as $S_t = BF_{t|t-1}B^T + \Sigma_V$ then (3.5) becomes:

$$F_{t|t} = F_{t|t-1} - \kappa BF_{t|t-1} - F_{t|t-1}B^T\kappa^T + \kappa S_t \kappa^T \quad (3.6)$$

4 Minimizing the state estimate variance

If we wish to minimize the variance of $F_{t|t}$, we can use the mean square error measure (MSE):

$$\mathbb{E} \left[\left| h_t - \hat{h}_{t|t} \right|^2 \right] = \text{Tr}(\text{Cov}(h_t - \hat{h}_{t|t})) = \text{Tr}(F_{t|t}) \quad (4.1)$$

The only coefficient we have control over is κ , so we wish to find the κ that gives us the minimum MSE, i.e. we need to find κ such that:

$$\begin{aligned} \frac{\delta \text{Tr}(F_{t|t})}{\delta \kappa} &= 0 \\ (2.6) \Rightarrow \frac{\delta \text{Tr}(F_{t|t-1} - \kappa B F_{t|t-1} - F_{t|t-1} B^T \kappa^T + \kappa S_t \kappa^T)}{\delta \kappa} &= 0 \\ \Rightarrow \kappa &= F_{t|t-1} B^T S_t^{-1} \end{aligned} \quad (4.2)$$

This optimum value for κ in terms of minimizing MSE is known as the Kalman Gain and will be denoted K .

If we multiply by both sides of (4.2) by $S K^T$:

$$K S K^T = F_{t|t-1} B^T K^T \quad (4.3)$$

Substituting this into (3.6):

$$\begin{aligned} F_{t|t} &= F_{t|t-1} - K B F_{t|t-1} - F_{t|t-1} B^T K^T + F_{t|t-1} B^T K^T \\ &= (I - K B) F_{t|t-1} \end{aligned} \quad (4.4)$$

5 Filtered Latent State Estimation Procedure (The Kalman Filter)

The procedure for estimating the state of h_t , which when using the MSE optimal value for κ is called Kalman Filtering, proceeds as follows:

1. Choose initial values for \hat{h} and F (i.e. $\hat{h}_{0|0}$ and $F_{0|0}$).
2. Advance latent state estimate:

$$\hat{h}_{t|t-1} = A\hat{h}_{t-1|t-1}$$
3. Advance estimate covariance:

$$F_{t|t-1} = AF_{t-1|t-1}A^T + \Sigma_H$$
4. Make an observation v_t
5. Calculate innovation:

$$\epsilon_t = v_t - B\hat{h}_{t|t-1}$$
6. Calculate S_t :

$$S_t = BF_{t|t-1}B^T + \Sigma_V$$
7. Calculate K :

$$K = F_{t|t-1}B^TS_t^{-1}$$
8. Update latent state estimate:

$$\hat{h}_{t|t} = \hat{h}_{t|t-1} + K\epsilon_t$$
9. Update estimate covariance (from (4.2)):

$$F_{t|t} = (I - KB)F_{t|t-1}$$
10. Cycle through stages 2 to 9 for each time step.

Note that $\hat{h}_{t|t}$ and $F_{t|t}$ correspond to μ_t and σ_t^2 from (2.4).

6 Smoothed Latent State Estimation

The smoothed probability of the latent variable is the probability it had a value at time t after a sequence of T observations, i.e. $p(h_t|v_{1:T})$. Unlike the Kalman Filter which you can update with each observation, one has to wait until T observations have been made and then retrospectively calculate the probability the latent variable had a value at time t where $t < T$.

Commencing at the final time step in the sequence ($t = T$) and working backwards to the start ($t = 1$), $p(h_t|v_{1:T})$ can be evaluated as follows:

$$\begin{aligned}
 p(h_t|v_{1:T}) &= \int_{h_{t+1}} p(h_t|h_{t+1}, v_{1:T})p(h_{t+1}|v_{1:T}) \\
 h_t \perp v_{t+1:T} | h_{t+1} &\Rightarrow p(h_t|h_{t+1}, v_{1:T}) = p(h_t|h_{t+1}, v_{1:t}) \\
 \therefore p(h_t|v_{1:T}) &= \int_{h_{t+1}} p(h_t|h_{t+1}, v_{1:t})p(h_{t+1}|v_{1:T}) \\
 &= \frac{\int_{h_{t+1}} p(h_{t+1}, h_t|v_{1:t})p(h_{t+1}|v_{1:T})}{p(h_{t+1}|v_{1:t})} \\
 &\propto \int_{h_{t+1}} p(h_{t+1}, h_t|v_{1:t})p(h_{t+1}|v_{1:T}) \\
 &= \int_{h_{t+1}} p(h_{t+1}|h_t, v_{1:t})p(h_t|v_{1:t})p(h_{t+1}|v_{1:T}) \\
 h_{t+1} \perp v_{1:t} | h_t &\Rightarrow \int_{h_{t+1}} p(h_{t+1}|h_t)p(h_t|v_{1:t})p(h_{t+1}|v_{1:T}) \tag{6.1}
 \end{aligned}$$

As before, we know that $p(h_t|v_{1:T})$ will be a Gaussian and we will need to establish it's mean and variance at each t , i.e. in a similar manner to (2.4):

$$p(h_t|v_{1:T}) \sim N(h_t^s, F_t^s) \quad (6.2)$$

Using the filtered values calculated in the previous section for $\hat{h}_{t|t}$ and $F_{t|t}$ for each time step, the procedure for estimating the smoothed parameters h_t^s and F_t^s works backwards from the last time step in the sequence, i.e. at $t = T$ as follows:

1. Set h_T^s and F_T^s to $\hat{h}_{T|T}$ and $F_{T|T}$ from steps 8 and 9 in section 5.
2. Calculate A_t^s :

$$A_t^s = (AF_{t|t})^T (AF_{t|t}A^T + \Sigma_H)^{-1}$$
3. Calculate S_t^s :

$$S_t^s = F_{t|t} - A_t^s A F_{t|t}$$
4. Calculate the smoothed latent variable estimate h_t^s :

$$h_t^s = A_t^s h_{t+1}^s + \hat{h}_{t|t} - A_t^s A \hat{h}_{t|t}$$
5. Calculate the smoothed estimate covariance F_t^s :

$$F_t^s = \frac{1}{2} [(A_t^s F_{t+1}^s A^T + S_t^s) + (A_t^s F_{t+1}^s A^T + S_t^s)^T]$$
6. Calculate the smoothed cross-variance X_t^s :

$$X_t^s = A_t^s F_t^s + h_t^s \hat{h}_{t|t}^T$$
7. Cycle through stages 2 to 6 for each time step backwards through the sequence from $t = T$ to $t = 1$.

7 Expectation Maximization (Calibrating the Kalman Filter)

The procedures outlined in the previous sections are fine if we assume that we know the value in the parameter set $\theta = \{\mu_0, \sigma_0^2, A, \Sigma_H, B, \Sigma_V\}$ but in order to learn these values, we will need to perform the Expectation Maximization algorithm.

The joint probability of T time steps of the latent and observable variables is:

$$p(h_{1:T}, v_{1:T}) = p(h_1) \prod_{t=2}^T p(h_t | h_{t-1}) \prod_{t=1}^T p(v_t | h_t) \quad (7.1)$$

Making the dependence on the parameters explicit, the likelihood of the model given the parameter set θ is:

$$p(h_{1:T}, v_{1:T} | \theta) = p(h_1 | \mu_0, \sigma_0^2) \prod_{t=2}^T p(h_t | h_{t-1}, A, \Sigma_H) \prod_{t=1}^T p(v_t | h_t, B, \Sigma_V) \quad (7.2)$$

Taking logs gives us the model's log likelihood:

$$\ln p(h_{1:T}, v_{1:T} | \theta) = \ln p(h_1 | \mu_0, \sigma_0^2) + \sum_{t=2}^T \ln p(h_t | h_{t-1}, A, \Sigma_H) + \sum_{t=1}^T \ln p(v_t | h_t, B, \Sigma_V) \quad (7.3)$$

We will deal with each of the three components of (7.3) in turn. Using V to represent the set of observations up to and including time t (i.e. $v_{1:t}$), H for $h_{1:T}$, θ^{old} to represent our parameter values before an iteration of the EM loop, the superscript n to represent the value of a parameter after an iteration of the loop, c to represent terms that are not dependent on μ_0 or σ_0^2 , λ to represent $(\sigma_0^2)^{-1}$ and $Q = \mathbb{E}_{H|\theta^{old}} [\ln p(H, V | \theta)]$ we will first find the expected value for $p(h_1 | \mu_0, \sigma_0^2)$:

$$\begin{aligned} Q &= -\frac{1}{2} \ln |\sigma_0^2| - \mathbb{E}_{H|\theta^{old}} \left[\frac{1}{2} (h_1 - \mu_0)^T \lambda (h_1 - \mu_0) \right] + c \\ &= -\frac{1}{2} \ln |\sigma_0^2| - \frac{1}{2} \mathbb{E}_{H|\theta^{old}} [h_1^T \lambda h_1 - h_1^T \lambda \mu_0 - \mu_0^T \lambda h_1 + \mu_0^T \lambda \mu_0] + c \\ &= \frac{1}{2} \left(\ln |\lambda| - \text{Tr} \left[\lambda \mathbb{E}_{H|\theta^{old}} [h_1 h_1^T - h_1 \mu_0^T - \mu_0 h_1^T + \mu_0 \mu_0^T] \right] \right) + c \end{aligned} \quad (7.4)$$

In order to find the μ_0 which maximizes the expected log likelihood described in (7.4), we will differentiate it wrt μ_0 and set the differential to zero:

$$\begin{aligned}\frac{\delta Q}{\delta \mu_0} &= 2\lambda\mu_0 - 2\lambda\mathbb{E}[h_1] = 0 \\ \Rightarrow \mu_0^n &= \mathbb{E}[h_1]\end{aligned}\tag{7.5}$$

Proceeding in a similar manner to establish the maximal λ :

$$\begin{aligned}\frac{\delta Q}{\delta \lambda} &= \frac{1}{2} (\sigma_0^2 - \mathbb{E}[h_1 h_1^T] - \mathbb{E}[h_1] \mu_0^T - \mu_0 \mathbb{E}[h_1^T] + \mu_0 \mu_0^T) = 0 \\ \sigma_0^{2n} &= \mathbb{E}[h_1 h_1^T] - \mathbb{E}[h_1] \mathbb{E}[h_1^T]\end{aligned}\tag{7.6}$$

In order to optimize for A and Σ_H we will substitute for $p(h_t|h_{t-1}, A, \Sigma_H)$ in (7.3) giving:

$$Q = -\frac{T-1}{2} \ln |\Sigma_H| - \mathbb{E}_{H|\theta^{old}} \left[\frac{1}{2} \sum_{t=2}^T (h_t - Ah_{t-1})^T \Sigma_H^{-1} (h_t - Ah_{t-1}) \right] + c\tag{7.7}$$

Maximizing with respect to these parameters then gives:

$$A^n = \left(\sum_{t=2}^T \mathbb{E}[h_t h_{t-1}^T] \right) \left(\sum_{t=2}^T \mathbb{E}[h_t h_t^T] \right)^{-1}\tag{7.8}$$

$$\Sigma_H^n = \frac{1}{T-1} \sum_{t=2}^T \left\{ \mathbb{E}[h_t h_{t-1}^T] - A^n \mathbb{E}[h_{t-1} h_t^T] - \mathbb{E}[h_t h_{t-1}^T] A^n + A^n \mathbb{E}[h_{t-1} h_{t-1}^T] A^n \right\}\tag{7.9}$$

In order to determine values for B and Σ_V we substitute for $p(v_t|h_t, B, \Sigma_V)$ in (7.3) to give:

$$Q = -\frac{T}{2} \ln |\Sigma_V| - \mathbb{E}_{H|\theta^{old}} \left[\frac{1}{2} \sum_{t=2}^T (v_t - Bh_t)^T \Sigma_V^{-1} (v_t - Bh_t) \right] + c\tag{7.10}$$

Maximizing this with respect to B and Σ_V gives:

$$B^n = \left(\sum_{t=1}^T v_t \mathbb{E}[h_t^T] \right) \left(\sum_{t=1}^T \mathbb{E}[h_t h_t^T] \right)^{-1}\tag{7.11}$$

$$\Sigma_V^n = \frac{1}{T} \sum_{t=1}^T \left\{ v_t v_t^T - B^n \mathbb{E}[h_t] v_t^T - v_t \mathbb{E}[h_t^T] B^n + B^n \mathbb{E}[h_t h_t^T] B^n \right\}\tag{7.12}$$

Using the values calculated from the smoothing procedure in section 5:

$$\begin{aligned}\mathbb{E}[h_t] &= h_t^s \\ \mathbb{E}[h_t h_t^T] &= F_t^s \\ \mathbb{E}[h_t h_{t-1}^T] &= X_t^s\end{aligned}$$

We can now set out the procedure for parameter learning using Expectation Maximization:

1. Choose starting values for the parameters $\theta = \{\mu_0, \sigma_0^2, A, \Sigma_H, B, \Sigma_V\}$.
2. Using the parameter set θ , calculate the filtered statistics $\hat{h}_{t|t}$ and $F_{t|t}$ for each time step as described in section 4.
3. Using the parameter set θ , calculate the smoothed statistics h_t^s , F_t^s and X_t^s for each time step as described in section 5.

4. Update A :

$$A^n = \left[\sum_{t=1}^T h_t^s h_{t-1}^{sT} + \sum_{t=1}^T X_t^s \right] \left[\sum_{t=1}^T h_t^s h_t^{sT} + \sum_{t=1}^T F_t^s \right]^{-1}$$

5. Update Σ_H :

$$\Sigma_H^n = [T-1]^{-1} \left[\sum_{t=2}^T h_t^s h_t^{sT} + \sum_{t=2}^T F_t^s - A^n \left(\sum_{t=1}^T h_t^s h_{t-1}^{sT} + \sum_{t=1}^T X_t^s \right)^T \right]$$

6. Update B :

$$B^n = \left[\sum_{t=1}^T v_t h_t^{sT} \right] \left[\sum_{t=1}^T h_t^s h_t^{sT} + \sum_{t=1}^T F_t^s \right]^{-1}$$

7. Update Σ_V :

$$\Sigma_V^n = \left[\sum_{t=1}^T v_t v_t^T - B^n \left(\sum_{t=1}^T v_t h_t^{sT} \right)^T \right] T^{-1}$$

8. Update μ_0 :

$$\mu_0^n = h_1^s$$

9. Update σ_0^2 :

$$\sigma_0^{2n} = [F_1^s + h_1^s h_1^{sT}] [1 - \mu_0^n \mu_0^{nT}]^{-1}$$

10. Iterate steps 2 to 10 a given number of times or until the difference between parameter values from succeeding iterations is below a pre-defined threshold.

References

- [1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.
- [2] Z. Ghahramani and G. Hinton, “Parameter estimation for linear dynamical systems,” Tech. Rep., 1996.