Discrete Time Kalman Filters and Smoothers

Please cite as

Movellan J. R. (2011) *Discrete Time Kalman Filters and Smoothers.* MPLab Tutorials, University of California San Diego

1 Conventions

Unless otherwise stated, capital letters are used for random variables, small letters for specific values taken by random variables, and Greek letters for fixed parameters and important functions. We leave implicit the properties of the probability space $(\Omega, \mathcal{F}, P)$ in which the random variables are defined. When the context makes it clear, we identify probability functions by their arguments: e.g., $p(x, y)$ is shorthand for the joint probability mass or joint probability density that the random variable X takes the specific value x and the random variable Y takes the value y. We use subscripted colons to indicate sequences: e.g., $X_{1:t} = (X_1 \cdots X_t)$. We use $\mathbb{E}$ for expected value operator and $\mathbb{V}$, $\mathbb{C}$ for the variance and covariance operators

$$\mathbb{C}(X,Y) = \mathbb{E}[XY'] - \mathbb{E}[X]\mathbb{E}[Y]' \tag{1}$$

$$\mathbb{V}(X) = \mathbb{C}(X,X) \tag{2}$$

- $(y_1, \cdots, y_T)$ is a fixed sequence of observations.
- $y_{t_i:t_j} = (y_{t_i}, \cdots, y_{t_j})$.
- $\mu_s^t = \mathbb{E}(X_t \mid y_{1:s})$.
- $\sigma_s^t = \mathbb{C}(X_t \mid y_{1:s})$, and $\eta_s^t = (\sigma_s^t)^{-1}$.
- $\phi(u|\mu, \sigma)$ represents a Gaussian density with mean vector μ, variance matrix σ, evaluated at u.
- $X \sim \mathcal{N}(\mu, \sigma)$ means that the random variable X is Gaussian with mean μ and covariance matrix σ.
- a' is the transpose of the matrix a.

2 Model

Consider a stochastic process $\{(X_t, Y_t) : t = 1, \cdots T\}$ where $X_t \in \mathbb{R}^{n_x}$ defined by the following equations

$$X_{t+1} = aX_t + Z_t \tag{3}$$

$$Y_t = bX_t + W_t \tag{4}$$

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1) \tag{5}$$

where a, b are known matrices. The parameters μ_1, σ_1 are known and represent the mean and covariance matrix of the initial state distribution. Z_t, and W_t are time independent multivariate Gaussian noise processes with zero mean and covariances matrices σ_z, and σ_w. Thus the model is specified by the parameters are $a, b, \mu_1, \sigma_1, \sigma_z, \sigma_w$.

3 Stochastic Filtering

The process X represents the system, dynamics, which may not be directly observable. The process Y represents the observations. Given a sequence of observations $y_{1:t}$ the goal in stochastic filtering is to compute the posterior distribution of the internal states $p(x_t \,|\, y_{1:t})$. It is easy to see that the X, Y process is jointly Gaussian, therefore the distribution of any subset of X conditioned on any subset of Y will also be Gaussian. It follows that the filtering distributions are fully specified by their means and variances.

3.1 Initialization

The initial distribution of the states are known

$$\boxed{\begin{array}{rclr} \mathbb{E}[X_1] & = & \mu_1 & (6) \\ \mathbb{C}[X_1] & = & \sigma_1 & (7) \end{array}}$$

3.2 Forward Prediction (Time Update)

Suppose X_t conditioned on $y_{1:t}$ is Gaussian with mean μ_t^t and covariance σ_t^t then, since

$$X_{t+1} = aX_t + Z_t \tag{8}$$

and since Z_t is independent of $y_{1:t}$ then

$$\boxed{\begin{array}{rclr} \mathbb{E}[X_{t+1} \mid y_{1:t}] & = & \mathbb{E}[aX_t + Z_t \mid y_{1:t}] = a\mu_t^t & (9) \\ \mathbb{V}[X_{t+1} \mid y_{1:t}] & = & \mathbb{V}[aX_t + Z_t \mid y_{1:t}] = a\sigma_t^t a' + \sigma_z & (10) \end{array}}$$

3.3 Forward Estimation Equations (Forward Correction)

We first consider the joint distribution of (X_{t+1}, Y_{t+1}) given $y_{1:t}$. We know it is Gaussian. Its mean is as follows

$$\mathbb{E}[X_{t+1} \mid y_{1:t}] = \mu_t^{t+1} \tag{11}$$

$$\mathbb{E}[Y_{t+1} \mid y_{1:t}] = \mathbb{E}[bX_{t+1} + W_{t+1} \mid y_{1:t}] = b\mu_t^{t+1} \tag{12}$$

and its covariance matrix has the following components

$$\mathbb{V}[X_{t+1} \mid y_{1:t}] = \mathbb{V}[aX_t + Z_t \mid y_{1:t}] = \sigma_t^{t+1} = a\sigma_t^t a' + \sigma_z \tag{13}$$

$$\mathbb{V}[Y_{t+1} \mid y_{1:t}] = \mathbb{V}[bX_{t+1} + W_{t+1} \mid y_{1:t}] = b\sigma_t^{t+1} b' + \sigma_w \tag{14}$$

$$\mathbb{C}[X_{t+1}, Y_{t+1} \mid y_{1:t}] = \mathbb{C}[X_{t+1}, bX_{t+1} + W_{t+1} \mid y_{1:t}] = \sigma_t^{t+1} b' \tag{15}$$

$$\mathbb{C}[Y_{t+1}, X_{t+1} \mid y_{1:t}] = \mathbb{C}[X_{t+1}, Y_{t+1} \mid y_{1:t}]'; \tag{16}$$

All we need to do now is get the distribution of X_{t+1} conditioned on $y_{1:t}$ and on $Y_{t+1} = y_{t+1}$. The distribution of a subset of Gaussian random variables given another subset of Gaussian random variables is well known (see Appendix). In our case

$$\mathbb{E}[X_{t+1} \mid y_{t+1}, y_{1:t}] = \mathbb{E}[X_{t+1} \mid y_{1:t}] + k_{t+1} \Big(y_{t+1} - \mathbb{E}[Y_{t+1} \mid y_{1:t}] \Big) \tag{17}$$

$$k_{t+1} = \mathbb{C}[X_{t+1}, Y_{t+1} \mid y_{1:t}] \Big(\mathbb{V}[Y_{t+1} \mid y_{1:t}] \Big)^{-1} \tag{18}$$

$$\mathbb{V}[X_{t+1} \mid y_{1:t+1}] = \mathbb{V}[X_{t+1} \mid y_{1:t}] - k_{t+1} \, \mathbb{V}[Y_{t+1} \mid y_{1:t}] k'_{t+1} \tag{19}$$

Equivalently

$$\boxed{\begin{aligned} \mu_{t+1}^{t+1} &= \mu_t^{t+1} + k_{t+1}(y_{1:t+1} - b\mu_t^{t+1}) && (20) \\ k_{t+1} &= \sigma_t^{t+1} b' (b \sigma_t^{t+1} b' + \sigma_w)^{-1} && (21) \\ \sigma_{t+1}^{t+1} &= \sigma_t^{t+1} - k_{t+1}(b\sigma_t^{t+1} b' + \sigma_w) k'_{t+1} && (22) \end{aligned}}$$

3.4 Summary of Filtering Equations

Forward Initialization	$\mathbb{V}(X_1) = \sigma_1$ $\mathbb{E}(X_1) = \mu_1$
Forward Prediction	$\sigma_t^{t+1} = a\sigma_t^t a' + \sigma_z$ $\mu_t^{t+1} = a\mu_t^t$
Forward Estimation/Correction	$k_{t+1} = \sigma_t^{t+1} b' (b\sigma_t^{t+1} b + \sigma_w)^{-1}$ $\mu_{t+1}^{t+1} = \mu_t^{t+1} + k_{t+1}(y_{t+1} - b\mu_t^{t+1})$ $\sigma_{t+1}^{t+1} = \sigma_t^{t+1} - k_{t+1}(b\sigma_t^{t+1} b' + \sigma_w) k'_{t+1}$

4 Probability of an observation sequence

The probability of an observation sequence $p(y_{1:t})$ tells us how well the model fits the sequence. Note

$$\log p(y_{1:t}) = \log p(y_1) + \sum_{t=1}^{T-1} \log p(y_{t+1} \mid y_{1:t}) \tag{23}$$

where each conditional distribution is Gaussian and thus can be characterized by its mean and variance matrix. For $t = 1$

$$\mathbb{E}[Y_1] = \mathbb{E}[bX_1 + W_1] = b\mathbb{E}[X_1] \tag{24}$$

$$\mathbb{V}[Y_1] = \mathbb{V}[bX_1 + W_1] = b\mathbb{V}[X_1]b' + \mathbb{V}[W_1] \tag{25}$$

Moreover, for $t > 1$

$$\mathbb{E}[Y_{t+1} \mid y_{1:t}] = \mathbb{E}[bX_{t+1} + W_{t+1} \mid y_{1:t}] = b\mathbb{E}[X_{t+1} \mid y_{1:t}] \tag{26}$$

$$\mathbb{V}[Y_{t+1} \mid y_{1:t}] = \mathbb{V}[bX_t + W_t \mid y_{1:t}] = b\mathbb{V}[X_t]b' + \mathbb{V}[W_t] \tag{27}$$

5 Smoothing

The goal in smoothing is to compute $p(x_t \mid y_{1:T})$ where $T \geq t$. This is a non-causal process in that we use future observations to estimate the present state.

We know the joint distribution of (X_t, X_{t+1}) given $y_{1:t}$ is Gaussian with the following mean and variance

$$\mathbb{E}[X_t \mid y_{1:t}] = \mu_t^t \tag{28}$$

$$\mathbb{E}[X_{t+1} \mid y_{1:t}] = a\mathbb{E}[X_t \mid y_{1:t}] = \mu_t^{t+1} \tag{29}$$

$$\mathbb{V}[X_t \mid y_{1:t}] = \sigma_t^t \tag{30}$$

$$\mathbb{V}[X_{t+1} \mid y_{1:t}] = \sigma_t^{t+1} = a\sigma_t^t a' + \sigma_z \tag{31}$$

$$\mathbb{C}[X_t, X_{t+1} \mid y_{1:t}] = \mathbb{C}[X_t, aX_t + W_t \mid y_{1:t}] = \sigma_t^t a' \tag{32}$$

$$\mathbb{C}[X_{t+1}, X_t \mid y_{1:t}] = \mathbb{C}[X_t, X_{t+1} \mid y_{1:t}]' = a\sigma_t^t \tag{33}$$

From the formula for the conditional distribution of a subset of Gaussian random variables given another subset of Gaussian random variables (see Appendix) it follows that

$$\mathbb{E}[X_t \mid x_{t+1}, y_{1:t}] = \mathbb{E}[X_t \mid y_{1:t}] + g_t \Big(x_{t+1} - \mathbb{E}[X_{t+1} \mid y_{1:t}]\Big) \tag{34}$$

$$g_t = \mathbb{C}[X_t, X_{t+1} \mid y_{1:t}] \Big(\mathbb{V}[X_{t+1} \mid y_{1:t}]\Big)^{-1} \tag{35}$$

$$\mathbb{V}[X_t \mid x_{t+1}, y_{1:t}] = \mathbb{V}[X_t \mid y_{1:t}] - g_t \, \mathbb{V}[X_{t+1} \mid y_{1:t}] \, g_t' \tag{36}$$

Note if we know X_{t+1} then all the future observations Y_{t+T} provide no further information about X_t, i.e.

$$p(x_t \mid x_{t+1}, y_{1:T}) = p(x_t \mid x_{t+1}, y_{1:t}) \tag{37}$$

Thus

$$\mathbb{E}[X_t \mid x_{t+1}, y_{1:T}] = \mathbb{E}[X_t \mid x_{t+1}, y_{1:t}] \tag{38}$$

$$\mathbb{V}[X_t \mid x_{t+1}, y_{1:T}] = \mathbb{V}[X_t \mid x_{t+1}, y_{1:t}] \tag{39}$$

The law of iterated expectations tells us tht for any three random vectors X, Y, Z

$$\mathbb{E}[X \mid z] = \mathbb{E}[\mathbb{E}[X \mid Y, z] \mid z] \tag{40}$$

In our case

$$\begin{aligned}\mathbb{E}[X_t \mid y_{1:T}] &= \mathbb{E}[\mathbb{E}[X_t \mid X_{t+1}, y_{1:T}] \mid y_{1:T}] \\ &= \mathbb{E}[X_t \mid y_{1:t}] + g_t\Big(\mathbb{E}[X_{t+1} \mid y_{1:T}] - \mathbb{E}[X_{t+1} \mid y_{1:t}]\Big)\end{aligned} \tag{41}$$

where we used the fact that $\mathbb{E}[X_t \,|\, y_{1:t}]$, k, $\mathbb{E}[X_{t+1} \,|\, y_{1:t}]$ are constants with respect to X_{t+1}. Regarding the covariance matrix. The law of iterated variance tells us that for any three random vectors X, Y, Z the following is true (see Appendix)

$$\mathbb{V}[X \mid z] = \mathbb{E}[\mathbb{V}[X \mid Y, z] \mid z] + \mathbb{V}[\mathbb{E}[X \mid Y, z] \mid z] \tag{42}$$

In our case,

$$\begin{aligned}\mathbb{E}[\mathbb{V}[X_t \mid X_{t+1}, y_{1:T}] \mid y_{1:T}] &= \mathbb{E}[\mathbb{V}[X_t \mid y_{1:t}] \mid y_{1:T}] \\ &\quad - \mathbb{E}[g_t \mathbb{V}[X_{t+1} \mid y_{1:t}] g_t' \mid y_{1:T}] \\ &= \mathbb{V}[X_t \mid y_{1:t}] - g_t \mathbb{V}[X_{t+1} \mid y_{1:T}] g_t'\end{aligned} \tag{43}$$

where we used the fact that $\mathbb{V}[X_t \,|\, y_{1:t}]$, g_t, $\mathbb{V}[X_{t+1} \,|\, y_{1:t}]$ are constants with respect to X_{t+1}. Moreover

$$\mathbb{V}[\mathbb{E}[X_t \mid X_{t+1}, y_{1:T}] \mid y_{1:T}] = \mathbb{V}[\mathbb{E}[X_t \mid X_{t+1}, y_{1:t}] \mid y_{1:T}] = g_t \mathbb{V}[X_{t+1} \mid y_{1:T}] g_t' \quad (44)$$

where we used the fact that $\mathbb{E}[X_t \,|\, y_{1:t}]$, g_t, $\mathbb{E}[X_{t+1} \,|\, y_{1:t}]$ are constants with respect to X_{t+1}. Thus

$$\begin{aligned}\mathbb{V}[X_t \mid y_{1:T}] &= \mathbb{V}[X_t \mid y_{1:t}] - g_t \mathbb{V}[X_{t+1} \mid y_{1:T}] g_t' + g_t \mathbb{V}[X_{t+1} \mid y_{1:T}] g_t' \\ &= \mathbb{V}[X_t \mid y_{1:t}] - g_t \Big(\mathbb{V}[X_{t+1} \mid y_{1:T}] - \mathbb{V}[X_{t+1} \mid y_{1:T}] \Big) g_t' \end{aligned} \quad (45)$$

Thus

$$\boxed{\begin{aligned} g_t &= \sigma_t^t a' (\sigma_t^{t+1})^{-1} && (46) \\ \mu_T^t &= \mu_t^t + g_t(\mu_T^{t+1} - \mu_t^{t+1}) && (47) \\ \sigma_T^t &= \sigma_t^t - g_t(\sigma_t^{t+1} - \sigma_T^{t+1}) g_t' && (48) \end{aligned}}$$

5.1 Initialization

μ_T^T and σ_T^T are given by the terminal condition of the filtering equation.

5.2 Summary of Smoothing Equations

Initialization	Compute μ_T^T, σ_T^T using a Kalman filter.
Backwards Update	$g_t = \sigma_t^t a' (\sigma_t^{t+1})^{-1}$ $\mu_T^t = \mu_t^t + g_t(\mu_T^{t+1} - \mu_t^{t+1})$ $\sigma_T^t = \sigma_t^t - g_t(\sigma_t^{t+1} - \sigma_T^{t+1}) g_t'$

6 Parameter Training using the EM Algorithm (under construction)

Let $\bar{\lambda}$ represent parameters of current model and λ parameters of a candidate model. The Q objective function takes the following form

$$Q(\bar{\lambda}, \lambda) = \int p(x_{1:T} \mid y_{1:T}, \bar{\lambda}) \log p(x_{1:T}, y_{1:T} \mid \lambda) \quad (49)$$

where

$$\log p(x_{1:T}, y_{1:T} \mid \lambda) = \log p(x_1 \mid \lambda) + \sum_{t=1}^{T-1} \log p(x_{t+1} \mid x_t, \lambda) + \sum_{t=1}^{T} \log p(y_t \mid x_t, \lambda) \quad (50)$$

6.1 Gradient with respect to a

Considering

$$\frac{\partial (ax-b)' c (ax-b)}{\partial a} = (c + c')(ax - b) x' \quad (51)$$

It follows that

$$\frac{\partial \log p(x_{t+1} \mid x_t, \lambda)}{\partial a} = -\frac{1}{2}\frac{\partial (x_{t+1} - ax_t)' \eta_z (x_{t+1} - ax_t)}{\partial a} = -\eta_z (x_{t+1} - ax_t) x_t' \tag{52}$$

Thus

$$\frac{\partial \log p(y_{1:T}, x_{1:T} \mid \lambda)}{\partial a} = \frac{\eta_z}{2}\sum_{t=1}^{T-1} (x_{t+1} - ax_t) x_t' \tag{53}$$

6.2 Gradient with respect to b

Considering

$$\frac{\partial \log p(y_t \mid x_t, \lambda)}{\partial b} = -\frac{1}{2}\frac{\partial (y_t - bx_t)' \eta_w (y_t - bx_t)}{\partial b} = -\eta_w (y_t - bx_t) x_t' \tag{54}$$

it follows that

$$\frac{\partial \log p(y_{1:T}, x_{1:T} \mid \lambda)}{\partial b} = \frac{\eta_w}{2}\sum_{t=1}^{T} (y_t - bx_t) x_t' \tag{55}$$

6.3 Gradient with respect to η_z

Considering

$$\frac{\partial \log |a|}{\partial a} = (a')^{-1} \tag{56}$$

$$\frac{\partial x' a x}{\partial a} = x x' \tag{57}$$

it follows that

$$2\frac{\partial \log p(x_{t+1} \mid x_t, \lambda)}{\partial \eta_z} = \frac{\partial \log |\eta_z|}{\partial \eta_z} - \frac{\partial (x_{t+1} - ax_t)' \eta_z (x_{t+1} - x_t)}{\partial \eta_z} \tag{58}$$

$$= \sigma_z - (x_{t+1} - ax_t)(x_{t+1} - ax_t)' \tag{59}$$

Thus

$$\frac{\partial \log p(y_{1:T}, x_{1:T} \mid \lambda)}{\partial \eta_z} = (T-1)\frac{\sigma_z}{2} - \frac{1}{2}\sum_{t=1}^{T-1} (x_{t+1} - ax_t)(x_{t+1} - ax_t)' \tag{60}$$

Note

6.4 Gradient with respect to κ_z

Sometimes we are interested in estimating the overall level of noise in the state dynamics. We can model this as

$$\eta_z = \kappa_z \alpha_z \tag{61}$$

where α_z is fixed matrix and $\kappa_z \in \mathbb{R}$ is adaptive. Note

$$2\frac{\partial \log p(x_{t+1} \mid x_t, \lambda)}{\partial \kappa_z} = \frac{\partial \log |\kappa_z \alpha_z|}{\partial \kappa_z} - \frac{\partial (x_{t+1} - ax_t)' \alpha_z \kappa_z (x_{t+1} - x_t)}{\partial \eta_z} \tag{62}$$

$$= \frac{n_x}{\kappa_z} - (x_{t+1} - ax_t)' \alpha_z (x_{t+1} - ax_t) \tag{63}$$

Thus

$$2\frac{\partial \log p(y_{1:T}, x_{1:T} \mid \lambda)}{\partial \kappa_z} = \frac{(T-1)n_x}{\kappa_z} - \sum_{t=1}^{T-1}(x_{t+1} - ax_t)'\alpha_z(x_{t+1} - ax_t) \tag{64}$$

Note

$$(x_{t+1} - ax_t)'\alpha_z(x_{t+1} - ax_t) = x'_{t+1}\alpha_z x_{t+1} + x'_t a'\alpha_z a x_t - 2x'_{t+1}\alpha_z a x_t \tag{65}$$

$$= \mathrm{Trace}[x_{t+1}x'_{t+1}\alpha_z] + \mathrm{Trace}[x_t x'_t a'\alpha_z a] - 2\mathrm{Trace}[x_t x'_{t+1}\alpha_z a] \tag{66}$$

6.5 Gradient with respect to η_w

Using the same argument as in the previous section we get

$$2\frac{\partial \log p(y_{t+1} \mid x_t, \lambda)}{\partial \eta_w} = \frac{\partial \log |\eta_w|}{\partial \eta_w} - \frac{\partial (y_t - bx_t)'\eta_w(y_t - bx_t)}{\partial \eta_w} \tag{67}$$

$$= \sigma_w - (y_t - bx_t)\eta_w(y_t - bx_t)' \tag{68}$$

Thus

$$\frac{\partial \log p(y_{1:T}, x_{1:T} \mid \lambda)}{\partial \eta_w} = T\frac{\sigma_w}{2} - \frac{1}{2}\sum_{t=1}^{T}(y_t - bx_t)(y_t - bx_t)' \tag{69}$$

6.6 Gradient with respect to κ_w

Sometimes we are interested in estimating the overall level of noise in the observations. We can model this as

$$\eta_w = \kappa_w \alpha_w \tag{70}$$

where α_w is fixed matrix and $\kappa_w \in \mathbb{R}$ is adaptive. Using the same argument as in the previous section we get

$$2\frac{\partial \log p(y_{t+1} \mid x_t, \lambda)}{\partial \kappa_w} = \frac{\partial \log |\kappa_w \alpha_w|}{\partial \kappa_w} - \frac{\partial (y_t - bx_t)'\kappa_w \alpha_w (y_t - bx_t)}{\partial \eta_w} \tag{71}$$

$$= \frac{n_y}{\kappa_w} - (y_t - bx_t)'\alpha_w(y_t - bx_t) \tag{72}$$

Thus

$$2\frac{\partial \log p(y_{1:T}, x_{1:T} \mid \lambda)}{\partial \kappa_w} = \frac{Tn_y}{\kappa_w} - \sum_{t=1}^{T}(y_t - bx_t)'\alpha_w(y_t - bx_t) \tag{73}$$

Note

$$(y_t - bx_t)'\alpha_w(y_t - bx_t) = y'_t\alpha_w y_t + x'_t b'\alpha_w b x_t - 2y'_t\alpha_w b x_t \tag{74}$$

$$= \mathrm{Trace}[y_t y'_t \alpha_w] + \mathrm{Trace}[x_t x'_t b'\alpha_w b] - 2\mathrm{Trace}[x_t y'_t \alpha_w b] \tag{75}$$

6.7 Taking Expectations

Note all the relevant gradients are linear functions of $x_t x'_t$, $x_t y'_t$, and $x_{t+1}x'_t$ terms. For example

$$\mathbb{E}[(Y_t - bX_t)(Y_t - bX_t)' \mid y_{1:T}] = \mathbb{E}[Y_t Y'_t \mid y_{1:T}] \\ + b\mathbb{E}[X_t X'_t \mid y_{1:T}]b - 2\mathbb{E}[Y_t X'_t \mid y_{1:T}] \mid y_{1;T}]b' \tag{76}$$

and

$$\mathbb{E}[(X_{t+1} - aX_t)(X_{t+1} - aX_t)' \mid y_{1:T}] = \mathbb{E}[X_{t+1}X_{t+1}' \mid y_{1:T}] \\ + a\mathbb{E}[X_tX_t' \mid y_{1:T}]a' - 2\mathbb{E}[X_{t+1}X_t' \mid y_{1:T}] \mid y_{1:T}]a' \tag{77}$$

Note

$$\mathbb{E}[X_tX_t' \mid y_{1:T}] = \mathbb{C}[X_t \mid y_{1:T}] + \mathbb{E}[X_t \mid y_{1:T}]\mathbb{E}[X_t \mid y_{1:T}]' \tag{78}$$

$$= \sigma_T^t + \mu_T^t(\mu_T^t)' \tag{79}$$

Moreover

$$\mathbb{E}[X_tY_t' \mid y_{1:T}] = \mathbb{E}[X_t \mid y_{1:T}]y_t' = \mu_T^t y_t' \tag{80}$$

Regarding $\mathbb{E}[X_tX_{t+1}' \mid y_{1:T}]$ using the law of iterated expectations (see Appendix) we get that

$$\mathbb{E}[X_tX_{t+1}' \mid y_{1:T}] = \mathbb{E}[\mathbb{E}[X_t \mid X_{t+1}, y_{1:T}]X_{t+1}' \mid y_{1:T}] \tag{81}$$

$$= \mathbb{E}[(\mu_t^t + g_t(X_{t+1} - \mu_t^{t+1}))X_{t+1}' \mid y_{1:T}]$$

$$= \mu_t^t\mathbb{E}[X_{t+1} \mid y_{1:T}]' + g_t\mathbb{E}[X_{t+1}X_{t+1}' - \mu_t^{t+1}X_{t+1}' \mid y_{1:T}] \tag{82}$$

$$= \mu_t^t\,(\mu_T^{t+1})' + g_t(\sigma_T^{t+1} + \mu_T^{t+1}(\mu_T^{t+1})' - \mu_t^{t+1}(\mu_T^{t+1})') \tag{83}$$

$$= \mu_t^t\,(\mu_T^{t+1})' + g_t(\sigma_T^{t+1} + (\mu_T^{t+1} - \mu_t^{t+1})(\mu_T^{t+1})') \tag{84}$$

7 Steady State Filter and Smoother

7.1 Steady State Filter

At a steady state $\sigma_{t+1}^{t+2} = \sigma_t^{t+1}$. Thus we can calculate the steady state uncertainty σ by solving the following equation

$$\sigma = a(\sigma - \sigma b'(b\sigma b + \sigma_w)^{-1}b\sigma)a' + \sigma_z \tag{85}$$

This is known as the Algebraic Riccati Equation. The steady state Kalman gain can then computed as follows

$$k = \sigma b'(b\sigma b + \sigma_w)^{-1} \tag{86}$$

Example Let $a = b = 1$. In this case the steady state variance and gain satisfy the following equations

$$\sigma = (1 - k)\sigma + \sigma_z \tag{87}$$

$$k = \frac{\sigma}{\sigma + \sigma_w} \tag{88}$$

Thus

$$\sigma k = \sigma_z \tag{89}$$

$$k = \frac{\sigma_z/k}{\sigma_z/k + \sigma_w} \tag{90}$$

$$\sigma_z + k\sigma_w = \frac{\sigma_z}{k} \tag{91}$$

$$k^2\sigma_w + k\sigma_z - \sigma_z = 0 \tag{92}$$

$$k = \frac{-\sigma_z + \sqrt{\sigma_z^2 + 4\sigma_z\sigma_w}}{2\sigma_w} \tag{93}$$

$$k = -\frac{r}{2} + \sqrt{\frac{r^2}{4} + r} \tag{94}$$

$$r \stackrel{\text{def}}{=} \frac{\sigma_z}{\sigma_w} \tag{95}$$

Following similar steps it can be shown that

$$\sigma = \frac{\sigma_z + \sqrt{\sigma_z^2 + 4\sigma_z\sigma_w}}{2} \tag{96}$$

$$\sigma = \sigma_w\Big(\frac{r}{2} + \sqrt{\frac{r^2}{4} + r}\Big) \tag{97}$$

For example, if $\sigma_z = \sigma_w = 1$ then $k = 0.618$ and $\sigma = 1.618$. If $\sigma_z = 0.0001$, $\sigma_w = 1$ then $k = 0.01$ and $\sigma = 0.01$.

Kalman Filters and Exponential Averages Consider 1-D case with

$$X_{t+1} = X_t + Z_t \tag{98}$$

$$Y_{t+1} = X_{t+1} + W_{t+1} \tag{99}$$

In this case

$$\mu_{t+1}^{t+1} = \mu_t^t + k_t(Y_t - \hat{\mu}_t^t) \tag{100}$$

which is an exponential smoother. The steady state gain is

$$k = \frac{\sigma}{\sigma + \sigma_w} \leq 1 \tag{101}$$

where

$$\sigma = (1 - \frac{\sigma}{\sigma + \sigma_w})\sigma + \sigma_z \tag{102}$$

$$\tag{103}$$

from which it follows that

$$\sigma = \frac{1}{2}\Big(\sigma_z + \sqrt{\sigma_z^2 + 4\sigma_z\sigma_w}\Big) \tag{104}$$

As it turns out the value of K depends only on the ratio between the system noise variance σ_z and the sensor noise variance σ_w. Table 1 shows the Kalman gain for different value of the system to sensor variance ratio. It also shows the effective number of sensory observations N being averaged by the Kalman filter to produce the state estimate (see Primer on Exponential Smoothing). The table illustrates that relying on the past is only useful if the sensors are more noisy than the system dynamic. If the system dynamics are very unpredictable it is better to “live in the present”.

σ_z/σ_w	K	N
1000	0.999	2003
100	0.9902	202
10	0.9161	22.83
4	0.8284	10.65
2	0.7321	6.46
1	0.618	4.23
0.5	0.5	3
0.25	0.394	2.28
0.1	0.2702	1.74
0.01	0.0951	1.21
0.001	0.0311	1.06

Table 1: First column is the ratio between system uncertainty and sensor uncertainty. Second column is the optimal stationary Kalman gain. In this case the Kalman filter operates as an exponential smoother. The third column shows the effective "memory" of the smoother, i.e., how far it averages into the past. A value of 1 means that it only pays attention to the current sensory value.

7.2 Steady State Smoother

Once the filtering distribution σ_t^t has reached a steady state σ then the smoothing gain will also be constant

$$g = \sigma a'(a\sigma a' + \sigma_z)^{-1} \tag{105}$$

At that point the Kalman smoother is just an exponential smoother running backwards in time on the output of the forward filter

$$\mu_T^t = (1-g)\mu_t^t + g\mu_T^{t+1} \tag{106}$$

with initial condition given by the terminal output μ_T^T of the forward filter. We saw before that in steady state the Kalman filter is fundamentally an exponential smoother. The Kalman smoother is then an exponential smoother (running backwards in time) operating on the output of an exponential smoother that runs forward in time.

Example Let $a = b = 1$. We saw that in this case the steady state variance and gain satisfy the following equations

$$k = -\frac{r}{2} + \sqrt{\frac{r^2}{4} + r} \tag{107}$$

$$\sigma = \sigma_w\left(\frac{r}{2} + \sqrt{\frac{r^2}{4} + r}\right) \tag{108}$$

where

$$r \stackrel{\text{def}}{=} \frac{\sigma_z}{\sigma_w} \tag{109}$$

Thus the steady state smoother gain g is as follows

$$g = \frac{\sigma}{\sigma + \sigma_z} \tag{110}$$

and since

$$\sigma = (1-k)\sigma + \sigma_z \tag{111}$$
$$\sigma = \frac{\sigma_z}{k} \tag{112}$$

then

$$g = \frac{\sigma_z/k}{\sigma_z/k + \sigma_z} = \frac{1}{1+1/k} \tag{113}$$

8 Biographical Notes

Rudolph Emil Kalman was born in Budapest, Hungary on May 19, 1930. He received BS degree from MIT in 1953, MS degree from MIT in 1954, and a ScD in engineering from Columbia University (1957). During those early years he showed a highly individual approach to research which continued during his entire career. He was a staff engineer at IBM from 1957 to 1958. From 1958 to 1964 he was at the Research Institute for Advanced Studies in Baltimore. In 1960 he published his classic paper on what we now know as the discrete time Kalman filter E. (1960). In 1961 he and R. S. Bucy generalized the approach to the continuous time case E. and S. (1961). He was was at Stanford University from 1964 to 1971. In 1971, he became a graduate research professor and director of the Center for Mathematical System Theory at the University of Florida. He retired in the late 90's with with emeritus status. The backward recursion used in smoothing was first published in E. (1963).

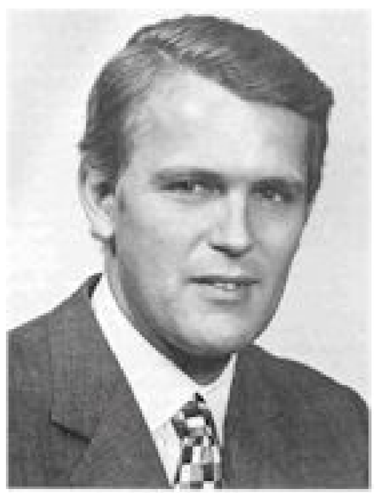

Figure 1: Rudolph Emil Kalman

9 Appendix: Unit Test for Kalman Filter and Smoother

Let

$$X_{t+1} = aX_t + Z_t \tag{114}$$
$$Y_{t+1} = bX_{t+1} + W_t \tag{115}$$

with

$$a = \begin{pmatrix} 1 & -0.5 \\ 0.5 & 1 \end{pmatrix} \tag{116}$$

$$b = (1, 2) \tag{117}$$

$$\sigma_z = \mathbb{V}(Z_t) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{118}$$

$$\sigma_w = \mathbb{V}(W_t) = 1 \tag{119}$$

In addition we let the prior distribution be Gaussian with the following parameters

$$\mu_1 = \mathbb{E}[X_1] = [1, -1]' \tag{120}$$

$$\sigma_1 = \mathbb{V}[X_1] = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{121}$$

We make the observations for the first 4 time steps be as follows

$$y_{1:4} = [-2, 4.5, 1.75, 7.625] \tag{122}$$

The Kalman filter should produce the following expected values

$$\mathbb{E}[X_1 \mid y_1] = [0.833, -1.333]' \tag{123}$$

$$\mathbb{E}[X_2 \mid y_{1:2}] = [2.8454, 0.5284]' \tag{124}$$

$$\mathbb{E}[X_3 \mid y_{1:3}] = [0.8237, 0.7109]' \tag{125}$$

$$\mathbb{E}[X_4 \mid y_{1:4}] = [2.5048, 2.3258]' \tag{126}$$

The Kalman smoother should produce the following expected values

$$\mathbb{E}[X_1 \mid y_{1:4}] = [1.3602, -1.3682]' \tag{127}$$

$$\mathbb{E}[X_2 \mid y_{1:4}] = [2.4797, 0.4091]' \tag{128}$$

$$\mathbb{E}[X_3 \mid y_{1:4}] = [2.1848, 0.2965]' \tag{129}$$

$$\mathbb{E}[X_4 \mid y_{1:4}] = [2.5048, 2.3258]' \tag{130}$$

10 Appendix: Gaussian densities

An n-dimensional random vector X with mean μ and covariance matrix σ is Gaussian if its density function has the following form

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |\sigma|}} \exp(\frac{1}{2}(x-\mu)^T \sigma^{-1}(x-\mu)) \tag{131}$$

10.1 Conditional Gaussian Densities

Let $X \in \mathbb{R}^n, Y \in \mathbb{R}^m$ be jointly Gaussian random vectors. It can be shown that the distribution of X given that Y takes the value y is Gaussian with the following mean and covariance matrix

$$\mathbb{E}[X \mid y] = \mathbb{E}[X] + k(y - \mathbb{E}[Y]) \tag{132}$$

$$\mathbb{V}[X \mid y] = \mathbb{V}[X] - k\mathbb{C}[X, Y]' == \mathbb{V}[X] - k\mathbb{V}[Y]k' \tag{133}$$

where the gain matrix k is defined as follows

$$k = \mathbb{C}[X, Y] \Big(\mathbb{V}[Y]\Big)^{-1} \tag{134}$$

10.2 Properties of Expected Values and Variances

For random vectors $X, Y \in \mathbb{R}^n$ let

$$\mathbb{C}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))'] \tag{135}$$

$$\mathbb{V}(X) = \mathbb{C}(X,X) \tag{136}$$

Then

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c \tag{137}$$

$$\mathbb{C}[aX + bY + c, Z] = a\mathbb{C}[X,Z] + b\mathbb{C}[Y,Z] \tag{138}$$

$$\mathbb{V}(aX) = a\mathbb{V}(X)a' \tag{139}$$

$$\mathbb{V}[aX + bY + c] = a\mathbb{V}[X]a' + b\mathbb{V}[Y]b' + 2a\mathbb{V}(X,Y)b' \tag{140}$$

10.3 Gaussian Densities and Quadratic Functions

Suppose we are told that a random variable X has a distribution of the following form

$$p(x) \propto e^{x'ax + bx + c} \tag{141}$$

where a is a symmetric positive definite matrix. Note this is the exponential of a quadratic function thus X must be a multivariate Gaussian distribution. To compute its mean and variance we want to put the exponent in the standard form, i.e., we want to find values of μ and σ such that

$$x'ax + b'x + c = -\frac{1}{2}(x - \mu)'\sigma^{-1}(x - \mu) + k \tag{142}$$

where k is a constant. We do so by matching first the terms quadratic on x

$$x'ax = -\frac{1}{2}x'\sigma^{-1}x \tag{143}$$

Thus

$$\sigma^{-1} = (\mathrm{Var}[X])^{-1} = -2a \tag{144}$$

We then match the terms linear on x

$$b'x = \mu'\sigma^{-1}x = -2\mu'ax \tag{145}$$

Thus, since this needs to hold for all x,

$$b' = -2\mu'a \tag{146}$$

and

$$\mu = \mathbb{E}[X] = -\frac{1}{2}a^{-1}b \tag{147}$$

10.4 Total Variance

Consider two arbitrary random vectors X, Y. Note

$$\mathbb{C}(X) = \int p(y) \int p(x \mid y)(x - \mathbb{E}[X|y] + \mathbb{E}[X|y] - \mathbb{E}[X]) \tag{148}$$

$$(x - \mathbb{E}[X|y] + \mathbb{E}[X|y] - \mathbb{E}[X])' dxdy \tag{149}$$

$$= \int p(y) \int p(x \mid y)(x - \mathbb{E}[X|y])(x - \mathbb{E}[X|y])' dxdy \tag{150}$$

$$+ \int p(y) \int p(x \mid y)(\mathbb{E}[X|y) - \mathbb{E}[X])(\mathbb{E}[X|y] - \mathbb{E}[X])' dxdy \tag{151}$$

$$= \int p(y)\mathbb{C}(X \mid y) dy$$

$$+ \int p(y)(\mathbb{E}[X|y) - \mathbb{E}[X])(\mathbb{E}[X|y] - \mathbb{E}[X])' dy \tag{152}$$

$$= \mathbb{E}[\mathbb{C}[X \mid Y]] + \mathbb{C}[\mathbb{E}[X \mid Y]] \tag{153}$$

10.5 Law of Iterated Expectations

Consider two arbitrary random vectors X, Y. Note

$$\mathbb{E}[X] = \int p(y) \int p(x \mid y) x dx dy = \int p(y) \mathbb{E}[X \mid y] dy = \mathbb{E}[\mathbb{E}[X \mid Y]] \tag{154}$$

We also use the following version of the law

$$\mathbb{E}[XY] = \int p(x, y) xy dx dy = \int p(x \mid y) y \int p(x \mid y) x dx dy \tag{155}$$

$$= \int p(y) \mathbb{E}[X \mid y] dy = \mathbb{E}[\mathbb{E}[X \mid Y]] \tag{156}$$

References

Kalman R. E. A new approach to linear filtering and prediction problems. *Transactions ASME J. of Basic Eng.*, 82:35–45, 1960.

Kalman R. E. and Bucy R. S. New results in linear filtering and prediction theory. *Transactions ASME J. of Basic Eng.*, 83:95–108, 1961.

Rauch H. E. Solutions to the linear smoothing problem. *IEEE Transactions on Automatic Control*, 8:371–372, 1963.