

Kalman Filter and EM (11/5/07)

Lecturer: Daniel Ting

Scribes:

- State space model + example and assumptions for the Kalman filter
- Graphical model (exactly the same as an HMM)
- Filtering and smoothing: Rauch-Tung-Streibel (RTS) algo (analogous to the $\alpha - \gamma$ recursion)
- E-step: what else we need to do this
- M-step: Estimating the parameters

Note: There is a lot of algebra/calculations for this.

1 State space model and the Kalman filter

The state space model used in the Kalman filter is of the form

$$\begin{aligned}X_{t+1} &= AX_t + W_t \\ Y_t &= CX_t + V_t\end{aligned}$$

$$\begin{aligned}W_t &\sim \text{Normal}(0, Q) \\ V_t &\sim \text{Normal}(0, R)\end{aligned}$$

where the noise terms, the W_t 's and V_t 's, are independent. (Note that there is a slight difference from the book which has GW_t rather than just W_t . This doesn't matter since $GW_t \sim N(0, GQG^T)$ so we may simply redefine Q appropriately)

Here the X_t 's are our hidden states, and the Y_t 's are the observed variables. The graphical model for this is the same as that for an HMM. One big difference, however, is that the X_t 's are now continuous variables rather than discrete.

We can see how this model is applicable in a radar system example. The states X_t encode say

- position
- velocity
- acceleration

The Y_t encode some noisy measurement of the position at the t^{th} radar sweep.

From the value of the current state, one can predict the state at the next timestep fairly well using a linear transformation from previous state. eg. new position = old position + velocity * δ + acceleration * $\delta^2/2$ where δ is the time between sweeps.

Compared to a more general state space model, the key assumptions to make estimating the parameters of the model possible and feasible are the

- homogeneity assumptions: ie. that the parameters A, C, Q, and R do not vary with time)
- Gaussianity assumptions: which gives us that $(X_0, \dots, X_T, Y_0, \dots, Y_T)$ is jointly Gaussian, and hence every conditional is also Gaussian.

Goals: Our goals are the same as those for the HMM. We will figure out how to solve the filtering and smoothing problems. Our solution will be the Rauch-Tung-Streifel algorithm. It is analogous to the α, γ recursion for HMMs. We will then see how that translates into an EM algorithm to estimate the parameters.

2 Filtering

Recall the filtering problem is to calculate $p(x_t|y_0, \dots, y_t)$. This is analogous to the α recursion for HMMs. From our Gaussianity assumptions, we know this density is also Gaussian. This means we can use the following important fact about Gaussians.

- The distribution of a Gaussian is completely determined by its mean and covariance matrix.

We will denote

$$\begin{aligned}\hat{x}_{t|s} &= \mathbb{E}(x_t|y_0, \dots, y_s) \\ P_{t|s} &= \text{Var}(x_t|y_0, \dots, y_s)\end{aligned}$$

So the filtering problem reduces to calculating $\hat{x}_{t|t}$ and $P_{t|t}$.

To do this efficiently, we use dynamic programming and make two updates:

- time update: $\hat{x}_{t|t} \rightarrow \hat{x}_{t+1|t}$
- measurement update: $\hat{x}_{t+1|t} \rightarrow \hat{x}_{t+1|t+1}$

and similarly for P

2.1 Key identities

To do the calculations, we will use some easily verified identities for expectation and (co)variance. Let A be a matrix and c be a constant. Let X, Y , and Z be random vectors (not necessarily gaussian)

$$\begin{aligned}\mathbb{E}(AX + Y) &= A\mathbb{E}(X) + \mathbb{E}(Y) \\ \text{Cov}(AX + Y + c, Z) &= A\text{Cov}(X, Z) + \text{Cov}(Y, Z) \\ \text{Var}(AX + Y) &= \text{Cov}(AX + Y, AX + Y) = A\text{Var}(X)A^T + \text{Var}(Y) \quad \text{if } \text{Cov}(X, Y) = 0 \\ \mathbb{E}(XX^T) &= \text{Var}(X) + \mathbb{E}(X)\mathbb{E}(X)^T\end{aligned}$$

2.2 Time update

The time update is an easy direct calculation.

$$\begin{aligned}
 \hat{x}_{t+1|t} &= \mathbb{E}(Ax_t + w_t | y_0, \dots, y_t) \\
 &= A\mathbb{E}(x_t | y_0, \dots, y_t) = A\hat{x}_{t|t} \\
 P_{t+1|t} &= \text{Var}(Ax_t + w_t | y_0, \dots, y_t) \\
 &= A \text{Var}(x_t | y_0, \dots, y_t) A^T + \text{Var}(w_t) = AP_{t|t}A^T + Q
 \end{aligned}$$

2.3 Measurement update

To do the measurement update we first look at the joint distribution $(X_{t+1}, Y_{t+1}) | y_0, \dots, y_t$. This joint is still Gaussian. We then condition on y_{t+1} as well to get the distribution of $X_{t+1} | y_0, \dots, y_t, y_{t+1}$ via a simple update.

Recall from chapter 13 (eqns 13.6 and 13.7): If

$$(Z_1, Z_2) \sim \text{Normal} \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

then

$$Z_1 | Z_2 \sim \text{Normal} (\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Z_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

We already have $\mu_1 = \hat{x}_{t+1|t}$ and $\Sigma_{11} = P_{t+1|t}$ from the filtering step.

$$\begin{aligned}
 \mu_2 &= \mathbb{E}(Y_{t+1} | y_0, \dots, y_t) = \mathbb{E}(CX_{t+1} + V_{t+1} | y_0, \dots, y_t) \\
 &= C\hat{x}_{t+1|t} \\
 \Sigma_{22} &= \text{Var}(Y_{t+1} | y_0, \dots, y_t) = \text{Var}(CX_{t+1} + V_{t+1} | y_0, \dots, y_t) \\
 &= CP_{t+1|t}C^T + R \\
 \Sigma_{21} &= \text{Cov}(Y_{t+1}, X_{t+1} | y_0, \dots, y_t) \\
 &= \text{Cov}(CX_{t+1} + V_{t+1}, X_{t+1} | y_0, \dots, y_t) \\
 &= CP_{t+1|t} \\
 \Sigma_{12} &= \Sigma_{21}^T
 \end{aligned}$$

This gives our measurement updates

$$\begin{aligned}
 \hat{x}_{t+1|t+1} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_{t+1} - \mu_2) \\
 &= \hat{x}_{t+1|t} + P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}(y_{t+1} - C\hat{x}_{t+1|t}) \\
 P_{t+1|t+1} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\
 &= P_{t+1|t} - P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1}(y_{t+1} - C\hat{x}_{t+1|t})
 \end{aligned}$$

Often this is written in terms of the Kalman gain matrix K_{t+1}

$$\begin{aligned} K_{t+1} &\stackrel{\text{def}}{=} P_{t+1|t}C^T(CP_{t+1|t}C^T + R)^{-1} \\ \hat{x}_{t+1|t+1} &= \hat{x}_{t+1|t} + K_{t+1}(y_{t+1} - C\hat{x}_{t+1|t}) \\ P_{t+1|t+1} &= P_{t+1|t} - K_{t+1}P_{t+1|t} \end{aligned}$$

3 Smoothing

Recall that the smoothing problem is calculating $p(x_t|y)$ where y is all the data y_0, \dots, y_T . Our strategy will be to

- Start with the joint $p(x_{t+1}, x_t|y_0, \dots, y_t)$
- Condition on X_{t+1} to make X_t independent of future observations (Y_{t+1}, \dots, Y_T) and calculate the conditional $p(x_t|x_{t+1}, y) = p(x_t|x_{t+1}, y_0, \dots, y_t)$
- “Uncondition” on X_{t+1} part to obtain $p(x_t|y)$

This step is analagous to the γ recursion for HMMs.

For the joint we already have

$$\begin{aligned} \mathbb{E}\left(\begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix} \middle| y_0, \dots, y_t\right) &= \\ \text{Var}\left(\begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix} \middle| y_0, \dots, y_t\right) &= \begin{pmatrix} P_{t|t} & ?^T \\ ? & P_{t+1|t} \end{pmatrix} \end{aligned}$$

That leaves us to calculate

$$? = \text{Cov}(X_{t+1}, X_t|y_0, \dots, y_t) = \text{Cov}(AX_t + V_t, X_t|y_0, \dots, y_t) = AP_{t|t}$$

Conditioning we get

$$\begin{aligned} \mathbb{E}(X_t|X_{t+1}, y) &= \mathbb{E}(X_t|X_{t+1}, y_0, \dots, y_t) \\ &= \hat{x}_{t|t} + P_{t|t}A^T P_{t+1|t}^{-1}(X_{t+1} - \hat{x}_{t+1|t}) \\ &= \hat{x}_{t|t} + L_t(X_{t+1} - \hat{x}_{t+1|t}) \\ \text{Var}(X_t|X_{t+1}, y) &= \text{Var}(X_t|X_{t+1}, y_0, \dots, y_t) \\ &= P_{t|t} - P_{t|t}A^T P_{t+1|t}^{-1}AP_{t|t} = P_{t|t} - L_tP_{t+1|t}L_t^T \end{aligned}$$

To get our final result, we need to “uncondition” on X_{t+1} . To do this we use the *Law of Iterated Expectation* (Tower rule)

$$\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X|Y)] \quad \text{or} \quad \mathbb{E}(X|Z) = \mathbb{E}[\mathbb{E}(X|Y, Z)|Z]$$

and the *Law of Total Variance*

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}[\mathbb{E}(X|Y)] \quad \text{or} \quad \text{Var}(X|Z) = \mathbb{E}[\text{Var}(X|Y, Z)|Z] + \text{Var}[\mathbb{E}(X|Y, Z)|Z]$$

The law of iterated expectation is analogous to the law of total probability. The law of total variance may be easily derived from the law of iterated expectation.

Applying these gives us

$$\begin{aligned}
\hat{x}_{t|T} &= \mathbb{E}(x_t|y) = \mathbb{E}[\mathbb{E}(X_t|X_{t+1}, y)|y] \\
&= \mathbb{E}[\hat{x}_{t|t} + L_t(X_{t+1} - \hat{x}_{t+1|t})|y] \\
&= \hat{x}_{t|t} + L_t(\hat{x}_{t+1|T} - \hat{x}_{t+1|t}) \\
P_{t|T} &= \text{Var}(x_t|y) = \mathbb{E}[\text{Var}(X_t|X_{t+1}, y)|y] + \text{Var}[\mathbb{E}(X_t|X_{t+1}, y)|y] \\
&= \mathbb{E}[P_{t|t} - L_t P_{t+1|t} L_t^T |y] + \text{Var}[\hat{x}_{t|t} + L_t(X_{t+1} - \hat{x}_{t+1|t})|y] \\
&= P_{t|t} - L_t P_{t+1|t} L_t^T + L_t P_{t+1|T} L_t^T \\
&= P_{t|t} - L_t(P_{t+1|t} - P_{t+1|T})L_t^T
\end{aligned}$$

This completes the RTS algo.

4 EM

We now turn to the problem of estimating the parameters of the model: A, C, Q, R .

As with HMMs, we derive an EM algorithm to do this. We start off by writing down the complete data log-likelihood and seeing what an appropriate set of sufficient statistics are.

4.1 E-step and the Complete data log-likelihood

For simplicity assume that $x_0 = 0$, and write $p(x_0|x_{-1}) = p(x_0) = \delta(x_0)$ where δ is the Dirac delta.

$$\begin{aligned}
l(A, C, Q, R|x, y) &= \log \prod_{t=0}^T p(x_t|x_{t-1})p(y_t|x_t) \\
&= \log p(x_0) + \sum_{t=0}^{T-1} \log p(x_{t+1}|x_t) + \sum_{t=0}^T \log p(y_t|x_t) \\
&= \left(\sum_{t=0}^{T-1} \frac{1}{2} \log |Q^{-1}| - \frac{1}{2} (x_{t+1} - Ax_t)^T Q^{-1} (x_{t+1} - Ax_t) \right) \\
&\quad + \left(\sum_{t=0}^T \frac{1}{2} \log |R^{-1}| - \frac{1}{2} (y_t - Cx_t)^T R^{-1} (y_t - Cx_t) \right) + \text{const} + \log p(x_0) \\
&= \frac{T}{2} \log |Q^{-1}| - \frac{1}{2} \left(\sum_{t=0}^{T-1} \text{Tr} \left((x_{t+1} - Ax_t)^T Q^{-1} (x_{t+1} - Ax_t) \right) \right) \\
&\quad + \frac{T+1}{2} \log |R^{-1}| - \frac{1}{2} \left(\sum_{t=0}^T \text{Tr} \left((y_t - Cx_t)^T R^{-1} (y_t - Cx_t) \right) \right) + \text{const} + \log p(x_0) \\
&= \frac{T}{2} \log |Q^{-1}| - \frac{1}{2} \text{Tr} \left(Q^{-1} \left(\sum_{t=0}^{T-1} (x_{t+1} - Ax_t)(x_{t+1} - Ax_t)^T \right) \right) \\
&\quad + \frac{T+1}{2} \log |R^{-1}| - \frac{1}{2} \text{Tr} \left(R^{-1} \left(\sum_{t=0}^T (y_t - Cx_t)(y_t - Cx_t)^T \right) \right) + \text{const} + \log p(x_0) \\
&= \frac{T}{2} \log |Q^{-1}| - \frac{1}{2} \text{Tr} \left(Q^{-1} \left(\sum_{t=0}^{T-1} x_{t+1}x_{t+1}^T - x_{t+1}x_t^T A^T - Ax_t x_{t+1}^T + Ax_t x_t^T A^T \right) \right) \\
&\quad + \frac{T+1}{2} \log |R^{-1}| - \frac{1}{2} \text{Tr} \left(R^{-1} \left(\sum_{t=0}^T y_t y_t - y_t x_t^T C^T - Cx_t y_t^T + Cx_t x_t^T C^T \right) \right) + \text{const} + \log p(x_0)
\end{aligned}$$

From this we see that the complete data log-likelihood is linear in the entries of the $x_t x_t^T$, $x_t x_{t+1}^T$, and x_t .

RTS already gives us

$$\begin{aligned}
\mathbb{E}(X_t|y) &= \hat{x}_{t|T} \\
\mathbb{E}(X_t X_t^T|y) &= \text{Var}(X_t|y) + \mathbb{E}(X_t|y)\mathbb{E}(X_t|y)^T = P_{t|T} + \hat{x}_{t|T}\hat{x}_{t|T}^T
\end{aligned}$$

This the only piece left we calculate using the law of iterated expectations again.

$$\begin{aligned}
\mathbb{E}(X_t X_{t+1}^T|y) &= \mathbb{E} [\mathbb{E}(X_t|X_{t+1}, y) X_{t+1}^T|y] \\
&= \mathbb{E} [\hat{x}_{t|t} + L_t (X_{t+1} - \hat{x}_{t+1|t}) X_{t+1}^T|y] \\
&= \hat{x}_{t|t} \mathbb{E} [X_{t+1}^T|y] + L_t \mathbb{E} [X_{t+1} X_{t+1}^T - \hat{x}_{t+1|t} X_{t+1}^T|y] \\
&= \hat{x}_{t|t} \hat{x}_{t+1|T}^T + L_t \left(P_{t+1|T} + \hat{x}_{t+1|T} \hat{x}_{t+1|T}^T - \hat{x}_{t+1|t} \hat{x}_{t+1|T}^T \right) \\
&= \hat{x}_{t|t} \hat{x}_{t+1|T}^T + L_t \left(P_{t+1|T} + (\hat{x}_{t+1|T} - \hat{x}_{t+1|t}) \hat{x}_{t+1|T}^T \right)
\end{aligned}$$

This gives us everything we need to calculate the expected complete data log-likelihood.

4.2 M-step

To do the M-step, note that the log-likelihood is concave with respect to the matrices A, C, Q, R . Thus we can maximize it by taking derivatives with respect to the matrices and setting the derivatives equal to 0.

Recall the following matrix calculus identities. For matrices A, B, M and vector μ of appropriate dimensions, we have

$$\frac{\partial \mu^T M \mu}{\partial \mu} = \mu^T (M^T + M)$$

Similarly

$$\frac{\partial A^T M A}{\partial A} = A^T (M^T + M) = 2A^T M \quad \text{if } M \text{ is symmetric}$$

Also,

$$\frac{\partial \text{Tr}(AB)}{\partial A} = \frac{\partial \text{Tr}(BA)}{\partial A} = \frac{\partial \text{Tr}(B^T A^T)}{\partial A} = B^T$$

and

$$\frac{\partial \log |M|}{\partial M} = M^{-T} = M^{-1} \quad \text{if } M \text{ is symmetric}$$

Let $\theta^{(n)}$ be the current value of the parameters. Taking the derivative w.r.t. A , we obtain

$$\frac{\partial \mathbb{E}(l|y, \theta^{(n)})}{\partial A} = \frac{1}{2} \left(Q^{-1} \left(\sum_{t=0}^{T-1} 2\mathbb{E}(x_{t+1}x_t^T | y, \theta^{(n)}) - 2A^T \mathbb{E}(x_t x_t^T | y, \theta^{(n)}) \right) \right)$$

Thus the maximizer for our update is

$$A^{(n+1)} = \left(\sum_{t=0}^{T-1} \mathbb{E}(x_{t+1}x_t^T | y, \theta^{(n)}) \right) \left(\sum_{t=0}^{T-1} \mathbb{E}(x_t x_t^T | y, \theta^{(n)}) \right)^{-1}$$

Similarly for C , the maximizer is

$$C^{(n+1)} = \left(\sum_{t=0}^{T-1} \mathbb{E}(x_{t+1}x_t^T | y, \theta^{(n)}) \right) \left(\sum_{t=0}^{T-1} \mathbb{E}(x_t x_t^T | y, \theta^{(n)}) \right)^{-1}$$

Taking the derivative w.r.t Q^{-1} , we obtain

$$\frac{\partial \mathbb{E}(l|y, \theta^{(n)})}{\partial Q^{-1}} = \frac{T}{2} Q - \frac{1}{2} \mathbb{E} \left(\sum_{t=0}^{T-1} x_{t+1}x_{t+1}^T - x_{t+1}x_t^T A^T - A x_t x_{t+1}^T + A x_t x_t^T A^T \mid y, \theta^{(n)} \right)^T$$

Thus gives the maximizer

$$Q^{(n+1)} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left(x_{t+1}x_{t+1}^T - x_{t+1}x_t^T A^T - A x_t x_{t+1}^T + A x_t x_t^T A^T \mid y, \theta^{(n)} \right)$$

Similarly

$$R^{(n+1)} = \frac{1}{T+1} \sum_{t=0}^T \mathbb{E} \left(y_t y_t^T - y_t x_t^T C^T - C x_t y_t^T + C x_t x_t^T C^T \mid y, \theta^{(n)} \right)$$

5 Summary

- Introduction to the Kalman filter state space model.
- Same setup as HMMs but with continuous as opposed to discrete hidden states and also Gaussian assumptions.
- The Gaussian assumptions reduce our filtering/smoothing problems to calculating some means and covariance matrices.
- We derive the RTS algorithm using an algorithm analagous to the α, γ recursion.
- To do the calculations, we rely on
 - some identities for expectation and (co)variance
 - eqns to convert a joint Gaussian to a conditional
 - law of iterated expectation and law of total variance
- We write down the complete data log-likelihood and use the trace trick to obtain something linear.
- E-step: We identify appropriate sufficient statistics and calculate their expected values given the data and current values of the parameters.
- M-step: We use some matrix calculus to find the maximizers of the expected complete data log-likelihood.