Lecture 11 — May 5

Lecturer: Lester Mackey *Scribe: Sidd Jagadish, Ben Nachman*

11.1 Linear Gaussian State Space Model

Last time we introduced the following linear Gaussian state space model:

- $z_0 \sim N(0, \Sigma_0)$
- $z_t = Az_{t-1} + w_{t-1}$ for independent $w_{t-1} \sim \mathcal{N}(0, Q)$ for all $t \geq 1$
- $x_t = Cz_t + v_t$ for independent $v_t \sim \mathcal{N}(0, R)$ for all $t \geq 0$.

11.2 Kalman Filter

Under the LGSSM, we can use the Kalman filter to compute the inferential quantity $p(z_t|x_{0:t})$ recursively assuming $\theta = (\Sigma_0, A, Q, C, R)$ known. We define the following shorthand notation to help us derive the recursive updates:

- $\hat{z}_{s|t} = \mathbb{E}[z_s|x_{0:t}]$
- $P_{s|t} = \mathbb{E}[(z_s - \hat{z_{s|t}})(z_s - \hat{z_{s|t}})^T|x_{0:t}]$

Last time, we described a two-step approach to deriving the Kalman filter consisting of a **time update** and a

11.2.1 Time Update

The first of our two updates computes the prediction distribution $p(z_{t+1}|x_{0:t})$ given the last filtered distribution $p(z_t|x_{0:t})$. Last time, we leveraged the fact that we know $z_{t+1}|x_{0:t}$ will take a normal distribution, and thus it is sufficient to calculate $\hat{z}_{t+1}$ and $P_{t+1|t}$. Last lecture, we found the following two update formulas.

$$\hat{z}_{t+1} = A\hat{z}_t \tag{11.1}$$

$$P_{t+1|t} = AP_{t|t}A^T + Q. \tag{11.2}$$

11.2.2 Measurement Update

The second of our two recursive updates computes the new filtered distribution $p(z_{t+1}|x_{0:t+1})$ given the prediction distribution $p(z_{t+1}|x_{0:t})$. We will do so by first computing the joint conditional density $p(x_{t+1}, z_{t+1}|x_{0:t})$. To do so, we need the mean and covariance of $x_{t+1}|x_{0:t}$.

Expectation of x_{t+1}

First, let's calculate $\mathbb{E}[x_{t+1}|x_{0:t}]$

$$\begin{aligned}\hat{x}_{t+1|t} &= \mathbb{E}[x_{t+1}|x_{0:t}] \\ &= \mathbb{E}[Cz_{t+1} + v_{t+1}|x_{0:t}] \\ &= C\mathbb{E}[z_{t+1}|x_{0:t}] = C\hat{z}_{t+1|t}\end{aligned}$$

Covariance of x_{t+1}

Now, let's calculate $\text{Cov}(x_{t+1}, x_{t+1}|x_{0:t})$

$$\begin{aligned}\text{Cov}(x_{t+1}, x_{t+1}|x_{0:t}) &= \mathbb{E}[(x_{t+1} - \hat{x}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})^T|x_{(0:t)}] \\ &= \mathbb{E}[(Cz_{t+1} + v_{t+1} - c\hat{z}_{t+1})(Cz_{t+1} + v_{t+1} - C\hat{z}_{t+1})^T|x_{0:t}] \\ &= CP_{t+1|t}C^T + R\end{aligned}$$

The final line makes use of the fact that z_{t+1} and v_{t+1} are independent, and the fact that $P_{t+1|t} = \mathbb{E}[(z_{t+1} - \hat{z}_{t+1|t})(z_{t+1} - \hat{z}_{t+1|t})^T]$.

We also need the "cross" covariance $\text{Cov}(x_{t+1}, z_{t+1}|x_{0:t})$

$$\begin{aligned}\text{Cov}(x_{t+1}, z_{t+1}|x_{0:t}) &= \mathbb{E}[(x_{t+1} - \hat{x}_{t+1|t})(z_{t+1} - \hat{z}_{t+1|t})^T|x_{0:t}] \\ &= \mathbb{E}[(Cz_{t+1} + v_{t+1} - C\hat{z}_{t+1|t})(Z_{t+1} - \hat{z}_{t+1|t})^T|X_{0:t}] \\ &= CP_{t+1|t}\end{aligned}$$

Thus, we have all we need for the joint distribution of $z_{t+1}, x_{t+1}|x_{0:t}$

$$\begin{bmatrix} z_{t+1} \\ x_{t+1} \end{bmatrix} \sim N\left(\begin{bmatrix} \hat{z}_{t+1|t} \\ C\hat{z}_{t+1|t} \end{bmatrix}, \begin{bmatrix} P_{t+1|T} & P_{t+1|t}C^T \\ CP_{t+1|t} & CP_{t+1|t}C^T + R \end{bmatrix}\right)$$

Now we take advantage of our knowledge of Gaussian conditional distributions to get $p(z_{t+1}|x_{0:t}, x_{t+1})$ (again, it suffices to know $\hat{z}_{t+1|t+1}$ and $P_{t+1|t+1}$). Our update formulas are

$$\hat{z}_{t+1|t+1} = \hat{z}_{t+1|t} + P_{t+1|t}C^T(CP_{t+1}C^T + R)^{-1}(x_{t+1} - C\hat{z}_{t+1|t}) \tag{11.3}$$

$$P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t}C^T(CP_{t+1}C^T + R)^{-1}CP_{t+1|t} \tag{11.4}$$

We see now that equations 11.1 through 11.4 together constitute an algorithm, the Kalman filter. We summarize the complete algorithm here

1. Initialize with $P_{0|-1} = \Sigma_0, \hat{z}_{0|-1} = 0$
2. Time Update

   $\hat{z}_{t+1} = A\hat{z}_{t+1}$

   $P_{t+1|t} = AP_{t|t}A^T + Q$

3. Measurement Update

$$\hat{z}_{t+1|t+1} = \hat{z}_{t+1|t} + P_{t+1|t}C^T(CP_{t+1}C^T + R)^{-1}(x_{t+1} - C\hat{z}_{t+1|t})$$
$$P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t}C^T(CP_{t+1}C^T + R)^{-1}CP_{t+1|t}$$

Kalman Gain Matrix

The matrix

$$K_{t+1} = P_{t+1|t}C^T(CP_{t+1}C^T + R)^{-1} \tag{11.5}$$

appearing in the measurement update is known as the **Kalman gain matrix**. Note that the current expression involves a $p \times p$ matrix inversion. However, due to Sherman-Morrison-Woodbury, we can rewrite the K_{t+1} as follows:

$$K_{t+1} = (P^{-1}_{t+1|t} + C^T RC)^{-1}C^T R^{-1} \tag{11.6}$$

We may initially worry that since R is a $p \times p$ matrix, we have not reduced our computational work; however, we note that R is fixed, and thus we only need to compute its inverse once. As such, for all iterations other than the first, we have reduced our work from inverting p x p matrix to inverting two q x q matrices.

11.3 Smoothing

The Kalman filter taught us how to recursively calculate $p(z_t|x_{0:t})$. We may also be interested in inference regarding a previous hidden state – that is, in calculating $p(z_t|x_{0:T}), t < T$, assuming our parameters θ are known. This is called the **smoothing** task. There are two standard recursive approaches to smoothing:

- The **two-filter algorithm** described in the SSM chapter is analogous to the forward-backward / α - β algorithm for HMMs
- The Rausch - Tung - Streibel smoother is analogous to the α-γ algorithm for HMM inference in the HMM chapter. This is the more common algorithm (in the LGSSM setting, not in the HMM setting), and hence it is the approach that we will focus on.

11.4 RTS Smoothing

Let us outline our plan of attack for smoothing.

1. We will first run the Kalman filter from time $0, ..., T$ to obtain the filtered and one-step prediction quantities $\hat{z_{t|t}}$, $\hat{z}_{t+1|t}$, $P_{t|t}$, $P_{t+1|t}$.
2. We will next initialize the smoother with $\hat{z_{T|T}}$ and $P_{T|T}$.
3. Finally, we will recursively compute $p(z_t|x_{0:T})$, given $p(z_{t+1}|x_{0:T})$. We know this is a Gaussian with mean $\hat{z}_{t|T}$ and covariance $P_{t|T}$. As such, we just need to find this mean and covariance.

To tackle the final item, we will make use of the fact that $z_t \perp x_{t+1:T} \mid z_{t+1}$ (recall our graphical model chain structure). As such, conditioning on z_{t+1} will simplify the smoothing computation and set us up nicely for recursion. To compute the full conditional distribution $p(z_t|z_{t+1}, x_{0:T}) = p(z_t|z_{t+1}, x_{0:t})$, we first compute the joint probability $p(z_t, z_{t+1}|x_{0:t})$ and then use Gaussian conditioning.

Computing $p(z_t, z_{t+1}|x_{0:t})$

We know that $p(z_t, z_{t+1}|x_{0:t})$ is Gaussian, so it suffices to compute its mean vector and its covariance matrix. We know the mean vector is

$$\begin{bmatrix} \hat{z}_{t|t} \\ \hat{z}_{t+1|t} \end{bmatrix},$$

where $\hat{z}_{t+1|t} = A\hat{z}_{t|t}$. We will use this when calculating the cross covariance:

$$\begin{aligned} \mathrm{Cov}(z_t, z_{t+1}|x_{0:t}) &= \mathbb{E}[(z_t - \hat{z}_{t|t})(z_{t+1|t} - \hat{z}_{t+1|t})^T|x_{0:t}] \\ &= \mathbb{E}[(z_t - \hat{z}_{t|t})(Az_t + w_t - A\hat{z}_{t|t})^T|x_{0:t}] \\ &= P_{t|t}A^T \end{aligned}$$

Thus, we obtain the covariance matrix

$$\begin{bmatrix} P_{t|t} & P_{t|t}A^T \\ AP_{t|t} & P_{t+1|t} \end{bmatrix}.$$

Computing $p(z_t|z_{t+1}, x_{0:T}) = p(z_t|z_{t+1}, x_{0:t})$

Now we will compute $p(z_t|z_{t+1}, x_{0:T})$ via $p(z_t|z_{t+1}, x_{0:t})$ by Gaussian conditioning. We find that

$$\mathbb{E}[z_t|z_{t+1}, x_{0:T}] = \mathbb{E}[z_t|z_{t+1}, x_{0:t}] = \hat{z}_{t|t} + L_t(z_{t+1} - \hat{z}_{t+1|t})$$

where

$$L_t = P_{t|t}A^T P_{t+1|t}^{-1}.$$

We also use Gaussian conditioning to compute

$$\mathrm{Cov}(z_t|z_{t+1}, x_{0:T}) = \mathrm{Cov}(z_t|z_{t+1}, x_{0:t}) = P_{t|t} - L_t P_{t+1|t} L_t^T$$

Computing $p(z_t|x_{0:T})$

Recall that our final goal is to compute $p(z_t|x_{0:T})$; we can achieve this by taking an expectation over z_{t+1} in $p(z_t|z_{t+1}, x_{0:T})$. To do so, two key conditioning properties of general random vectors X, Y, and Z.

- Tower Property: $\mathbb{E}[Z|X] = \mathbb{E}[\mathbb{E}[Z|Y, X]|X]$
- Law of Total Conditional Variance: $\mathrm{Cov}[Z|X] = \mathrm{Cov}[\mathbb{E}[Z|Y, X]|X] + \mathbb{E}[\mathrm{Cov}[Z|Y, X]|X]$

We will apply the above two properties, using $Z = z_t$, $Y = z_{t+1}$, $X = x_{0:T}$. First, we compute the smoothed mean $\hat{z}_{t|T}$.

$$\begin{aligned}\hat{z}_{t|T} &= \mathbb{E}[z_t|x_{0:T}] \\ &= \mathbb{E}[\mathbb{E}[z_t|z_{t+1}, x_{0:T}]|x_{0:T}] \\ &= \mathbb{E}[\hat{z}_{t|t} + L_t(z_{t+1} - \hat{z}_{t+1|t})|x_{0:T}] \\ &= \hat{z}_{t|t} + L_t(\hat{z}_{t+1|T} - \hat{z}_{t+1|t})\end{aligned}$$

We see that our final expression consists of our filtering estimate $\hat{z}_{t|t}$ added to a correction term $L_t(\hat{z}_{t+1|T} - \hat{z}_{t+1|t})$. Now, all that remains is to apply thte law of total conditional variance to find $P_{t|T}$:

$$\begin{aligned}P_{t|T} &= \mathrm{Cov}[z_t|x_{0:T}] \\ &= \mathrm{Cov}(\mathbb{E}[z_t|z_{t+1}, x_{0:T}]|x_{0:T}) + \mathbb{E}[\mathrm{Cov}[z_t|z_{t+1}, x_{0:T}]|x_{0:T}] \\ &= \mathrm{Cov}(\hat{z}_{t|t} + L_t(z_{t+1} - \hat{z}_{t+1|t})|x_{0:T}) + \mathbb{E}[P_{t|t} - L_t P_{t+1|t} L_t^T|X_{0:T}] \\ &= L_t P_{t+1|T} L_t^T + P_{t|t} - L_t P_{t+1|t} L_t^T \\ &= P_{t|t} + L_t(P_{t+1|T} - P_{t+1|t}) L_t^T\end{aligned}$$

We note again that our final expression consists of our filtering estimate and a correction term.

11.5 EM for the Linear Gaussian State Space Model

Now that we have learned how to conduct inference in LGSSMs for known model parameters θ, we turn to the question of estimating those parameters. Unfortunately, there are no closed-form MLEs, so we turn as usual to the EM algorithm. Let us begin by formulating the complete log likelihood:

$$\begin{aligned}\log p(x_{0:T}, z_{0:T}; \theta) = -\frac{1}{2}\Bigg(&\log|\Sigma_0| + z_0^T \Sigma_0^{-1} z_0 + \sum_{t=1}^{T} \log|Q| + (z_t - Az_{t-1})^T Q^{-1}(z_t - Az_{t-1}) \\ &+ \sum_{t=0}^{T} \log|R| + (x_t - Cz_t)^T R^{-1}(x_t - Cz_t)\Bigg) + \text{constants}\end{aligned}$$

Introduce the shorthand $M_0 = z_0 z_0^T$,

$$M = \frac{1}{T}\sum_{t=1}^{T}(z_t - Az_{t-1})(z_t - Az_{t-1})^T, \quad \text{and} \quad N = \frac{1}{T+1}\sum_{t=0}^{T}(x_t - Cz_t)(x_t - Cz_t)^T,$$

where N is the same conditional sample covariance that we saw in the factor analysis setting. With this notation, the complete log likelihood becomes

$$\begin{aligned}\log p(x_{0:T}, z_{0:T}; \theta) = -\frac{1}{2}\big(&\log|\Sigma_0| + \mathrm{tr}(M_0 \Sigma_0^{-1}) \\ &+ T[\log|Q| + \mathrm{tr}(MQ^{-1})] + (T+1)[\log|R| + \mathrm{tr}(NR^{-1})]\big) + \text{constants}\end{aligned}$$

where we have used the same trace trick as in factor analysis. Now, we turn to the E-step.

11.5.1 E step

In the E step we form the expected complete log likelihood under the conditional distribution

$$q_s(z_{0:T}) = p(z_{0:T}|x_{0:T}; \theta^{(s)}),$$

where $\theta^{(s)}$ are the parameters from the previous step in the algorithm. It suffices to compute $\mathbb{E}_{q_s}[M_0] = \mathbb{E}[z_0 z_0^T | x_{0:T}]$, which we know from smoothing, $\mathbb{E}_{q_s}[M]$, and $\mathbb{E}_{q_s}[N]$. $\mathbb{E}_{q_s}[N]$ depends on $\mathbb{E}[z_t|x_{0:T}]$ and

$$\mathbb{E}[z_t z_t^T | x_{0:T}] = \mathrm{Cov}(z_t|x_{0:T}) + \mathbb{E}[z_t|x_{0:T}]\mathbb{E}[z_t^T|x_{0:T}],$$

for each t, both of which are known from smoothing. $\mathbb{E}_{q_s}[M]$ also depends on $\mathbb{E}_{q_s}[z_t z_{t-1}^T|x_{0:T}]$ for all $t \geq 1$, which we have not directly computed. However, we note that

$$\mathbb{E}[z_t z_{t-1}^T | x_{0:T}] = \mathrm{Cov}(z_t z_{t-1}|x_{0:T}) + \mathbb{E}[z_t|x_{0:T}]\mathbb{E}[z_{t-1}^T|x_{0:T}],$$

and both of these terms are computable from smoothing/filtering (work this out for yourself - consider $p(z_{t-1}|z_t, x_{0:T})$). The takeaway message is that we can carry out the E step by running the Kalman filter and RTS smoothing.

11.5.2 M step

Now, we optimize the ECLL. As in factor analysis, there exists a closed form for the updates:

$$\begin{aligned}
C^{(s+1)} &= \left(\sum_{t=0}^{T} x_t \mathbb{E}_{q_s}[z_t^T]\right)\left(\sum_{t=0}^{T} \mathbb{E}_{q_s}[z_t z_t^T]\right)^{-1} \\
A^{(s+1)} &= \left(\sum_{t=1}^{T} \mathbb{E}_{q_s}[z_t z_{t-1}^T]\right)\left(\sum_{t=1}^{T} \mathbb{E}_{q_s}[z_{t-1} z_{t-1}^T]\right)^{-1} \\
\Sigma_0^{(s+1)} &= \mathbb{E}_{q_s}[M_0] \\
R^{(s+1)} &= \mathbb{E}_{q_s}[N] \\
Q^{(s+1)} &= \mathbb{E}_{q_s}[M],
\end{aligned}$$

where after some rearrangement we have

$$\begin{aligned}
R^{(s+1)} = \mathbb{E}_{q_s}[N] &= \frac{1}{T+1}\left[\sum_{t=0}^{T} x_t x_t^T - C^{(s+1)} \sum_{t=0}^{T} \mathbb{E}_{q_s}[z_t] x_t^T\right] \\
Q^{(s+1)} = \mathbb{E}_{q_s}[M] &= \frac{1}{T}\left[\sum_{t=1}^{T} \mathbb{E}_{q_s}[z_t z_t^T] - A^{(s+1)} \sum_{t=1}^{T} \mathbb{E}_{q_s}[z_{t-1} z_t^T]\right].
\end{aligned}$$