SVAE LDS KL Divergence

Kavi Dey

October 8, 2024

1 Starting Loss Function

Start with the KL Divergence between two sequences of length T (predicted and updated):

$$
\begin{aligned}
KL(q||p) &= \mathbb{E}_{q(z|x)}\left[\log \frac{q(z|x)}{p(z)}\right] && (1)\\
&= \mathbb{E}_{q(z|x)}\left[\log q(z|x) - logp(z)\right] && (2)\\
&= \mathbb{E}_{q(z|x)}\left[\log q(z_1|x) + \sum_{i=2}^{T} \log q(z_i|z_{i-1}, x) - \log p(z_1) - \sum_{i=2}^{T} \log p(z_i|z_{i-1})\right] && (3)
\end{aligned}
$$

The terms outside of the summation can be separated and form their own KL divergence:

$$
= \underbrace{KL(q(z_1|x)||p(z_1))}_{\text{Term A}} + \sum_{i=2}^{T} \mathbb{E}_{q(z_i, z_{i-1}|x)}\left[\underbrace{\log q(z_i|z_{i-1}, x)}_{\text{Term B}} - \underbrace{\log p(z_i|z_{i-1})}_{\text{Term C}}\right] \quad (4)
$$

The KL divergence in term A can be calculated in closed form, terms B and C are calculated below. Term C is simpler than term B so we will calculate it first.

2 Kalman Filter Equations

In *Bayesian Filtering and Smoothing* (Särkkä and Svensson) describe a kalman filter as follows:

$$
\begin{aligned}
z_i &= Az_{i-1}b + q && (5)\\
x_i &= Hz_i + r && (6)
\end{aligned}
$$

Where z_i is the state, A is the transition matrix, and b is a constant input. x_i is the measurement and H is the measurement matrix. $q \sim \mathcal{N}(0, Q)$ is the process noise and $r \sim \mathcal{N}(0, R)$ is the measurement noise, and $z_0 \sim \mathcal{N}(\mu_0, P_0)$ is the prior.

Then we can write the prediction and update equations as follows:
Prediction Step Equations:

$$\begin{aligned} z_{i|i-1} &= \mathbf{A}z_{i-1|i-1} + \mathbf{b} && (7)\\ P_{i|i-1} &= \mathbf{A}P_{i-1|i-1} + \mathbf{A}^T + Q && (8)\\ & && (9) \end{aligned}$$

Update Step Equations

$$\begin{aligned} K_i &= P_{i|i-1}H^T(HP_{i|i-1}H^T + R)^{-1} \qquad \text{(Kalman Gain)} && (10)\\ z_{i|i} &= z_{i|i-1} + K_i(x_i - Hz_{i|i-1}) && (11)\\ P_{i|i} &= P_{i|i-1} - K_iHP_{i|i-1} = (I - K_iH)P_{i|i-1} && (12) \end{aligned}$$

It will also be useful to write out the covariance of the joint distribution between $z_{i|i-1}$ and $z_{i-1|i-1}$ as well as $z_{i|i}$ and $zi|i-1$. Following *Probabilistic Machine Learning: Advanced Topics* by Kevin Murphy (8.40, 8.45), we can write the covariance as:

$$\Sigma_{z_{i|i-1},z_{i-1|i-1}} = \begin{pmatrix} \Sigma_{z_{i|i-1}z_{i|i}} & \Sigma_{z_{i|i-1}z_{i-1|i-1}} \\ \Sigma_{z_{i-1|i-1}z_{i|i-1}} & \Sigma_{z_{i-1|i-1}z_{i-1|i-1}} \end{pmatrix} = \begin{pmatrix} P_{i-1|i-1} & P_{i-1|i-1}A^T \\ AP_{i-1|i-1} & AP_{i-1|i-1}A^T + Q \end{pmatrix} \quad (13)$$

$$\Sigma_{z_{i|i},z_{i|i-1}} = \begin{pmatrix} \Sigma_{z_{i|i}z_{i|i}} & \Sigma_{z_{i|i}z_{i|i-1}} \\ \Sigma_{z_{i|i-1}z_{i|i}} & \Sigma_{z_{i|i-1}z_{i|i-1}} \end{pmatrix} = \begin{pmatrix} P_{i|i-1} & P_{i|i-1}H^T \\ H^TP_{i|i-1} & HP_{i|i-1}^{-1}H^T + R \end{pmatrix} \quad (14)$$

3 Calculating Term C

Lets start with term C from section 1. From the kalman filter update step $\mu_{i|i-1} = \mathbf{A}z_{i-1|i-1} + \mathbf{b}$ and $P_{i|i-1} = \mathbf{A}P_{i-1}\mathbf{A}^T + Q$. Additionally we use the shorthand $\mathbb{E}_q = \mathbb{E}_{q(z_i,z_{i-1}|x)}$.

$$\begin{aligned} \mathbb{E}_q\left[\log p(z_i|z_{i-1})\right] &= -\frac{1}{2}\mathbb{E}_q\left[k\log(2\pi) + \log\det P_i\right] - \frac{1}{2}\mathbb{E}_q\left[(z_{i|i-1} - \mu_{i|i-1})^T P_i (z_{i|i-1} - \mu_{i|i-1})\right] && (15)\\ &= -\frac{1}{2}\left[k\log(2\pi) + \log\det P_i\right] - \frac{1}{2}\mathbb{E}_q\left[(z_{i|i-1} - \mu_{i|i-1})^T P_i (z_{i|i-1} - \mu_{i|i-1})\right] && (16)\\ &= -\frac{1}{2}\left[k\log(2\pi) + \log\det P_i\right] \\ &\quad - \frac{1}{2}\mathbb{E}_q\Big[\underbrace{z_{i|i-1}^T P_{i|i-1}^{-1} z_{i|i-1}}_{①} - \underbrace{z_{i|i-1}^T P_{i|i-1}^{-1}\mathbf{A}z_{i-1|i-1}}_{②} - \underbrace{z_{i|i-1}^T P_{i|i-1}^{-1}\mathbf{b}}_{③} - \underbrace{\mathbf{A}^T z_{i-1|i-1}^T P_{i|i-1}^{-1} z_{i|i-1}}_{④} \\ &\quad + \underbrace{\mathbf{A}^T z_{i-1|i-1}^T P_{i|i-1}^{-1}\mathbf{A}z_{i-1|i-1}}_{⑤} + \underbrace{\mathbf{A}^T z_{i-1|i-1}^T P_i^{-1}\mathbf{b}}_{⑥} - \underbrace{\mathbf{b}^T P_i^{-1} z_{i|i-1}}_{⑦} + \underbrace{\mathbf{b}^T P_i^{-1}\mathbf{A}z_{i-1|i-1}}_{⑧} + \underbrace{\mathbf{b}^T P_i^{-1}\mathbf{b}}_{⑨}\Big] \\ & && (17) \end{aligned}$$

Term ⑨ is constant and can be pulled out of the expectation, all other terms can be calculated with expectation identities from appendix A.

$$\begin{aligned}
① :&\quad \mathbb{E}_q[z_{i|i-1}^T P_{i|i-1}^{-1} z_{i|i-1}] = \mathrm{tr}[P_{i|i-1}^{-1}(\mu_{i|i-1}\mu_{i|i-1}^T + P_{i|i-1}^{-1})] \\
② :&\quad \mathbb{E}_q[z_{i|i-1}^T P_{i|i-1}^{-1} \mathbf{A} z_{i-1|i-1}] = 1 \\
③ :&\quad \mathbb{E}_q[z_{i|i-1}^T P_{i|i-1}^{-1} \mathbf{b}] = 1 \\
④ :&\quad \mathbb{E}_q[\mathbf{A}^T z_{i-1|i-1}^T P_{i|i-1}^{-1} z_{i|i-1}] = 1 \\
⑤ :&\quad \mathbb{E}_q[\mathbf{A}^T z_{i-1|i-1}^T P_{i|i-1}^{-1} \mathbf{A} z_{i-1|i-1}] = 1 \\
⑥ :&\quad \mathbb{E}_q[\mathbf{A}^T z_{i-1|i-1}^T P_i^{-1} \mathbf{b}] = 1 \\
⑦ :&\quad \mathbb{E}_q[\mathbf{b}^T P_i^{-1} z_{i|i-1}] = 1 \\
⑧ :&\quad \mathbb{E}_q[\mathbf{b}^T P_i^{-1} \mathbf{A} z_{i-1|i-1}] = 1 \\
⑨ :&\quad \mathbb{E}_q[\mathbf{b}^T P_i^{-1} \mathbf{b}] = \mathbf{b}^T P_i^{-1} \mathbf{b}
\end{aligned}$$

4 Calculating Term B

The process for calculating term B starts the same way as term A. $q(z_i|z_{i-1})$ represents the update step of the kalman filter, so the equations for μ_i and Σ_i are more complicated.

$$\mathbb{E}_q\left[\log p(z_i|z_{i-1})\right] = -\frac{1}{2}\mathbb{E}_q\left[k\log(2\pi) + \log\det P_i\right] - \frac{1}{2}\mathbb{E}_q\left[(z_i-\mu_i)^T P_i (z_i-\mu_i)\right] \tag{18}$$

Using traditional kalman filter from section 2 we get:

$$\mu_i = \mathbf{A}z_{i-1|i-1} + \mathbf{b} + K_i(x_i - H(\mathbf{A}z_{i-1|i-1} + \mathbf{b})) \tag{19}$$

$$P_{i|i} = P_{i|i-1} - K_i H P_{i|i-1} = (I - K_i H)P_{i|i-1} \tag{20}$$

$P_{i|i}$ is constant with respect to z, so we don't need to expand the full equation. It will be helpful to simplify the equation for μ_i as follows:

$$\begin{aligned}
\mu_i &= \mathbf{A}z_{i-1|i-1} + \mathbf{b} + K_i(x_i - H(\mathbf{A}z_{i-1|i-1} + \mathbf{b})) && (21)\\
&= \mathbf{A}z_{i-1|i-1} + \mathbf{b} + K_i x_i - K_i H\mathbf{A}z_{i-1|i-1} - K_i H\mathbf{b} && (22)\\
&= \mathbf{A}z_{i-1|i-1} - K_i H\mathbf{A}z_{i-1|i-1} + \underbrace{\mathbf{b} - K_i H\mathbf{b} + K_i x_i}_{\text{constant } C} && (23)
\end{aligned}$$

Now we can expand equation 18 into:

$$\begin{aligned}
\mathbb{E}_q\left[\log p(z_i|z_{i-1})\right] &= -\frac{1}{2}\left[k\log(2\pi) + \log\det P_i\right] \\
&\quad - \frac{1}{2}\mathbb{E}_q\left[(z_{i|i}-\mu_i)^T P_i (z_{i|i}-\mu_i)\right] && (24)\\
&= -\frac{1}{2}\left[k\log(2\pi) + \log\det P_i\right] \\
&\quad - \frac{1}{2}\mathbb{E}_q\left[(z_{i|i} - \mathbf{A}z_{i-1|i-1} + K_i H\mathbf{A}z_{i-1|i-1} - C)^T P_i (z_{i|i} - \mathbf{A}z_{i-1|i-1} + K_i H\mathbf{A}z_{i-1|i-1} - C)\right] \\
& && (25)\\
& && (26)
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}\left[k\log(2\pi) + \log\det P_i\right] \\
&\quad - \frac{1}{2}\mathbb{E}_q\Big[\underbrace{z_{i|i}^T P_{i|i}^{-1} z_{i|i}}_{①} - \underbrace{z_{i|i}^T P_{i|i}^{-1} A z_{i|i-1}}_{②} + \underbrace{z_{i|i}^T P_{i|i}^{-1} K_i H A z_{i|i-1}}_{③} - \underbrace{z_{i|i}^T P_{i|i}^{-1} C}_{④} \\
&\quad - \underbrace{A^T z_{i|i-1}^T P_{i|i}^{-1} z_{i|i}}_{⑤} + \underbrace{A^T z_{i|i-1}^T P_{i|i}^{-1} A z_{i|i-1}}_{⑥} - \underbrace{A^T z_{i|i}^T P_{i|i}^{-1} K_i H A z_{i|i-1}}_{⑦} + \underbrace{A^T z_{i|i}^T P_{i|i}^{-1} C}_{⑧} \\
&\quad + \underbrace{K_i^T H^T A^T z_{i|i-1}^T P_{i|i}^{-1} z_{i|i}}_{⑨} - \underbrace{K_i^T H^T A^T z_{i|i-1}^T P_{i|i}^{-1} A z_{i|i-1}}_{⑩} \\
&\qquad + \underbrace{K_i^T H^T A^T z_{i|i-1}^T P_{i|i}^{-1} K H A z_{i|i-1}}_{⑪} - \underbrace{K_i^T H^T A^T z_{i|i-1}^T P_{i|i}^{-1} C}_{⑫} \\
&\quad - \underbrace{C^T P_{i|i}^{-1} z_{i|i}}_{⑬} + \underbrace{C^T P_{i|i}^{-1} A z_{i|i-1}}_{⑭} - \underbrace{C^T P_{i|i}^{-1} K_i H A z_{i|i-1}}_{⑮} + \underbrace{C^T P_{i|i}^{-1} C}_{⑯}\Big]
\end{aligned} \tag{27}
$$

Term ⑯ is constant and can be pulled out of the expectation, all other terms can be calculated with expectation identities from appendix A.

$$① : \quad \mathbb{E}_q[z_{i|i}^T P_{i|i}^{-1} z_{i|i}] = \mathrm{tr}[P_{i|i}^{-1}(\mu_{i|i}\mu_{i|i}^T + P_{i|i}^{-1})] \tag{28}$$

$$② : \quad \mathbb{E}_q[z_{i|i}^T P_{i|i}^{-1} A z_{i|i-1}] = 1 \tag{29}$$

$$③ : \quad \mathbb{E}_q[z_{i|i}^T P_{i|i}^{-1} K_i H A z_{i|i-1}] = 1 \tag{30}$$

$$④ : \quad \mathbb{E}_q[z_{i|i}^T P_{i|i}^{-1} C] = \mu_{i|i}^T P_{i|i}^{-1} C \tag{31}$$

$$⑤ : \quad \mathbb{E}_q[A^T z_{i|i-1}^T P_{i|i}^{-1} z_{i|i}] = 1 \tag{32}$$

$$⑥ : \quad \mathbb{E}_q[A^T z_{i|i-1}^T P_{i|i}^{-1} A z_{i|i-1}] = 1 \tag{33}$$

$$⑦ : \quad \mathbb{E}_q[A^T z_{i|i}^T P_{i|i}^{-1} K_i H A z_{i|i-1}] = 1 \tag{34}$$

$$⑧ : \quad \mathbb{E}_q[A^T z_{i|i}^T P_{i|i}^{-1} C] = 1 \tag{35}$$

$$⑨ : \quad \mathbb{E}_q[K_i^T H^T A^T z_{i|i-1}^T P_{i|i}^{-1} z_{i|i}] = 1 \tag{36}$$

$$⑩ : \quad \mathbb{E}_q[K_i^T H^T A^T z_{i|i-1}^T P_{i|i}^{-1} A z_{i|i-1}] = 1 \tag{37}$$

$$⑪ : \quad \mathbb{E}_q[K_i^T H^T A^T z_{i|i-1}^T P_{i|i}^{-1} K H A z_{i|i-1}] = 1 \tag{38}$$

$$⑫ : \quad \mathbb{E}_q[K_i^T H^T A^T z_{i|i-1}^T P_{i|i}^{-1} C] = 1 \tag{39}$$

$$⑬ : \quad \mathbb{E}_q[C^T P_{i|i}^{-1} z_{i|i}] = 1 \tag{40}$$

$$⑭ : \quad \mathbb{E}_q[C^T P_{i|i}^{-1} A z_{i|i-1}] = 1 \tag{41}$$

$$⑮ : \quad \mathbb{E}_q[C^T P_{i|i}^{-1} K_i H A z_{i|i-1}] = 1 \tag{42}$$

$$⑯ : \quad \mathbb{E}_q[C^T P_{i|i}^{-1} C] = C^T P_{i|i}^{-1} C \tag{43}$$

A Expectation Identities

$$\mathbb{E}_{z\sim\mathcal{N}(\mu,\Sigma)}[z] = \mu \tag{44}$$

$$\mathbb{E}_{z\sim\mathcal{N}(\mu,\Sigma)}[zz^T] = \mu\mu^T + \Sigma \tag{45}$$

$$\mathbb{E}_{y\sim\mathcal{N}(\mu_y,\Sigma_y),z\sim\mathcal{N}(\mu_z,\Sigma_z)}[zy^T] = \mu_z\mu_y^T + \Sigma_{zy} \tag{46}$$

B Matrix Identities

The following matrix identities are copied from Max Welling's wonderful notes on Kalman Filters (stat.columbia.edu/ liam/teaching/neurostat-spr12/papers/hmm/KF-welling-notes.pdf)

$$\mathbf{a}^T\mathbf{A}\mathbf{b} = \mathrm{tr}[\mathbf{A}\mathbf{b}\mathbf{a}^T] \tag{47}$$

$$\mathrm{tr}[\mathbf{A}\mathbf{B}] = \mathrm{tr}[\mathbf{B}\mathbf{A}] \tag{48}$$

$$\log\det[\mathbf{A}] = -\log\det[\mathbf{A}^{-1}] \tag{49}$$