DS-GA 3001.001 Special Topics in Data Science: Modeling Time Series
Handout: Kalman filtering and smoothing derivations

1 LSD model specification

The latents have first order Markov structure with linear dynamics and i.i.d. gaussian noise:

$$\mathbf{z}_i = \mathbf{A}\mathbf{z}_{i-1} + \mathbf{w}_i \tag{1}$$

where $\mathbf{z}_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $\mathbf{w}_i \sim \mathcal{N}(0, \mathbf{Q})$ (to avoid degeneracy, we usually take $\mathbf{Q} = \mathbf{I}$).

Observations are obtained by another linear projection with i.i.d. gaussian noise:

$$\mathbf{x}_i = \mathbf{C}\mathbf{z}_i + \mathbf{v}_i \tag{2}$$

with $\mathbf{v}_i \sim \mathcal{N}(0, \mathbf{R})$.

2 Kalman filter

The goal is to compute the posterior distribution of the latent state $\mathbf{z}_i$ given past observations $\{\mathbf{x}_1, \dots, \mathbf{x}_i\}$. The proof proceeds by induction: we start by inferring the posterior distribution of latent variable $\mathbf{z}_1$ condition on observation $\mathbf{x}_1$ (marginalizing out the unobserved initial state $\mathbf{z}_0$). The resulting posterior is also gaussian so it can be treated as prior for inferring the next latent state and so on.

To compute the posterior $\mathrm{P}(\mathbf{z}_1|\mathbf{x}_1)$ we start from the joint distribution $\mathrm{P}(\mathbf{z}_1, \mathbf{x}_1)$ and then use the general formula for conditioning in multivariate Gaussians:

$$\begin{aligned} \mu_{z|x} &= \mu_z + \Sigma_{zx}\Sigma_{xx}^{-1}(\mathbf{x} - \mu_x) && (3)\\ \Sigma_{z|x} &= \Sigma_{zz} - \Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz} && (4) \end{aligned}$$

To compute the means, we start with $\mathbf{z}_1 = \mathbf{A}\mathbf{z}_{i-1} + \mathbf{w}_i$ which is a sum of 2 independent multivariate gaussian distributions with means $\mathbf{A}\mu_0$ and 0, and covariances $\mathbf{A}\Sigma_0\mathbf{A}^{\mathrm{t}}$ and $\mathbf{Q}$, respectively. Putting the two together, yields the prior $\mathbf{z}_1 \sim \mathcal{N}\left(\mu_{1|0}, \Sigma_{1|0}\right)$, with parameters: [1]

$$\begin{aligned} \mu_z &= \mu_{1|0} = \mathbf{A}\mu_0 && (5)\\ \Sigma_{zz} &= \Sigma_{1|0} = \mathbf{A}\Sigma_0\mathbf{A}^{\mathrm{t}} + \mathbf{Q} && (6) \end{aligned}$$

Since $\mathbf{x}_1$ is generated using the same kind of process, we can get the corresponding moments exactly in the same way:

$$\begin{aligned} \mu_x &= \mathbf{C}\mu_{1|0} && (7)\\ \Sigma_{xx} &= \mathbf{C}\Sigma_{1|0}\mathbf{C}^{\mathrm{t}} + \mathbf{R} && (8) \end{aligned}$$

Lastly, we need to compute the covariance:[2]

$$\Sigma_{zx} = \mathrm{cov}[\mathbf{z}_1, \mathbf{x}_1] = \mathrm{cov}[\mathbf{z}_1, \mathbf{C}\mathbf{z}_1 + \mathbf{v}_t] = \Sigma_{1|0}\mathbf{C}^{\mathrm{t}} \tag{9}$$

where we have used the fact that the observation noise is independent (so the second term cancels out).

Plugging the different elements in Eq. 3 we obtain the posterior mean for $\mathbf{z}_1$:

$$\mu_{1|1} = \mu_{1|0} + \mathbf{K}\left(x_1 - \mathbf{C}\mu_{1|0}\right) \tag{10}$$

where we use matrix $\mathbf{K} = \Sigma_{zx}\Sigma_{xx}^{-1}$ as shorthand for the Kalman gain.

---

[1]Notational convention for the double indexing of $\mu_{i|j}$ and $\Sigma_{i|j}$ means that we are computing the moments of latent $\mathbf{z}_i$, conditioned on observations up to j, $\{\mathbf{x}_1, \dots, \mathbf{x}_j\}$.

[2]Remember that linear combinations distribute when computing covariances.

The corresponding posterior covariance matrix is (using Eq. 4):

$$\Sigma_{1|1} = \Sigma 1|0 - \mathbf{KC}\Sigma_{1|0}. \tag{11}$$

In general, for $i > 1$ we first compute the prior for $\mathbf{z}_{i+1}$ as:

$$\begin{aligned} \mu_{i|i-1} &= \mathbf{A}\mu_{i-1|i-1} && (12) \\ \Sigma_{i|i-1} &= \mathbf{A}\Sigma_{i-1|i-1}\mathbf{A}^{\mathrm{t}} + \mathbf{Q} && (13) \end{aligned}$$

Then we incorporate the evidence, to obtain the posterior with parameters:

$$\begin{aligned} \mu_{i|i} &= \mu_{i|i-1} + \mathbf{K}_i \left(x_i - \mathbf{C}\mu_{i|i-1}\right) && (14) \\ \Sigma_{i|i} &= \Sigma_{i|i-1} - \mathbf{K}_i\mathbf{C}\Sigma_{i|i-1} && (15) \end{aligned}$$

where the Kalman gain is $\mathbf{K}_i = \Sigma_{i|i-1}\mathbf{C}^{\mathrm{t}} \left(\mathbf{C}\Sigma_{i|i-1}\mathbf{C}^{\mathrm{t}} + R\right)^{-1}$.

3 Kalman smoothing

The goal of smoothing is to compute the posterior distribution of the latent state $\mathbf{z}_i$ but given the full sequence of observations observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$. There are very different ways of deriving the smoother; here we use the approach of Ainsley and Kohn (1982), as presented in Chapter 6 of Shumway and Stoffer.

The starting point of this version of the derivation is focusing on the dependency between $\mathbf{z}_i$ and $\mathbf{z}_{i+1}$, and explicitly separating out different components of the generative model such that the parts are (conditionally) independent. Specifically, we start by defining some shorthand variables $\mathbf{x}_{1:i} = \{\mathbf{x}_1, \ldots, \mathbf{x}_i\}$ and $\eta_{i+1} = \{\mathbf{w}_{i+2}, \ldots, \mathbf{w}_t, \mathbf{v}_{i+1}, \ldots, \mathbf{v}_t\}$.

We also introduce an additional auxiliary random variable m_i, defined as follows:

$$m_i = \mathbb{E}[\mathbf{z}_i|\mathbf{x}_{1:i}, \mathbf{z}_{i+1} - \mu_{i+1|i}, \eta_{i+1}] \tag{16}$$

This estimates the uncertainty about $\mathbf{z}_i$ after taking into account observations up to $\mathbf{x}_i$, the prediction error at step $i + 1$(which depends implicitly on $\mathbf{w}_i$) and all the subsequent noise terms, which are independent of $\mathbf{z}_i$. Note also that the variables $\mathbf{x}_{1:i}$, $z_{i+1} - \mu_{i+1|i}$, η_i are mutually independent.

To compute m_i we use again the trick of computing joint statistics for $\mathbf{z}_i$ and $\left(\mathbf{z}_{i+1} - \mu_{i+1|i}\right)$, conditioned on the rest and then recovering m_i via the multivariate normal conditioning formula, as done for the Kalman filter above.

First, the moments of $\mathbf{z}_i$ conditioned on $\mathbf{x}_{1:i}$ are $\mu_{i|i}$ and $\Sigma_{i|i}$, according to our original definition.

Second, the corresponding moments for random variable $\left(\mathbf{z}_{i+1} - \mu_{i+1|i}\right)$ are: [3]

$$\begin{aligned} \mathbb{E}[\mathbf{z}_{i+1} - \mu_{i+1|i}|\mathbf{x}_{1:i}] &= \mathbb{E}[\mathbf{z}_{i+1}|\mathbf{x}_{1:i}] - \mu_{i+1|i} = 0 && (17) \\ \mathrm{Var}[\mathbf{z}_{i+1} - \mu_{i+1|i}|\mathbf{x}_{1:i}] &= \Sigma_{i+1|i} && (18) \end{aligned}$$

Lastly the covariance term is:

$$\mathrm{cov}[\mathbf{z}_i, \mathbf{z}_{i+1} - \mu_{i+1|i}|\mathbf{x}_{1:i}] = \mathrm{cov}[\mathbf{z}_i, \mathbf{A}\mathbf{z}_i + w_{i+1} - \mu_{i+1|i}|\mathbf{x}_{1:i}] = \Sigma_{i|i}\mathbf{A}^{\mathrm{t}} \tag{19}$$

since the rest of the terms cancel out (because of conditional independence).

Putting everything together (Eq. 3) we get the final expression for m_i:

$$m_i = \mu_{i|i} + \mathbf{F}_i \left(\mathbf{z}_{i+1} - \mu_{i+1|i}\right) \tag{20}$$

where $\mathbf{F}_i = \Sigma_{i|i}\mathbf{A}^{\mathrm{t}}\Sigma_{i+1|i}^{-1}$.

[3] After conditioning, $\mu_{i+1|i}$ is just a constant.

Lastly, we use m_i to compute the mean of $\mathbf{z}_i$ conditioned on the full sequence of observations $\mathbf{x}_{1:t}$:

$$
\begin{aligned}
\mu_{i|t} &= \mathbb{E}[m_i|\mathbf{x}_{1:t}] && (21)\\
&= \mathbb{E}[\mu_{i|i} + \mathbf{F}_i \left(\mathbf{z}_{i+1} - \mu_{i+1|i}\right) |\mathbf{x}_{1:t}] && (22)\\
&= \mu_{i|i} + \mathbf{F}_i \left(\mathbb{E}[\mathbf{z}_{i+1}|\mathbf{x}_{1:t}] - \mu_{i+1|i}\right) && (23)\\
&= \mu_{i|i} + \mathbf{F}_i \left(\mu_{i+1|t} - \mu_{i+1|i}\right). && (24)
\end{aligned}
$$

For the covariance update rules, we start from the expression derived from the mean, $\mu_{i|t} = \mu_{i|i} + \mathbf{F}_i \left(\mu_{i+1|t} - \mu_{i+1|i}\right)$, subtracting $\mathbf{z}_i$ on both sides and rewriting slightly we get:

$$
\mathbf{z}_i - \mu_{i|t} + \mathbf{F}_i\mu_{i+1|t} = \mathbf{z}_i - \mu_{i|i} + \mathbf{F}_i\mathbf{A}\mu_{i|i} \qquad (25)
$$

where we have used $\mu_{i+1|i} = \mathbf{A}\mu_{i|i}$. We then multiply each side with its transpose and take expectations (for $\mathbf{z}_i$). With some appropriate ordering of terms this yields:

$$
\begin{aligned}
\mathbb{E}[(\mathbf{z}_i - \mu_{i|t})(\mathbf{z}_i - \mu_{i|t})^{\mathrm{t}}] + \mathbf{F}_i\mathbb{E}[\mu_{i+1|t}\mu_{i+1|t}^{\mathrm{t}}]\mathbf{F}_i^{\mathrm{t}} &= \mathbb{E}[(\mathbf{z}_i - \mu_{i|i})(\mathbf{z}_i - \mu_{i|i})^{\mathrm{t}}] + \mathbf{F}_i\mathbf{A}\mathbb{E}[(\mu_{i|i})(\mu_{i|i})^{\mathrm{t}}]\mathbf{A}^{\mathrm{t}}\mathbf{F}_i^{\mathrm{t}} && (26)\\
\Sigma_{i|t} + \mathbf{F}_i\mathbb{E}[\mu_{i+1|t}\mu_{i+1|t}^{\mathrm{t}}]\mathbf{F}_i^{\mathrm{t}} &= \Sigma_{i|i} + \mathbf{F}_i\mathbf{A}\mathbb{E}[\mu_{i|i}\mu_{i|i}^{\mathrm{t}}]\mathbf{A}^{\mathrm{t}}\mathbf{F}_i^{\mathrm{t}} && (27)
\end{aligned}
$$

where we used the fact that the cross-products cancel out. Finally, the remaining expectations can be rewritten as:[4]:

$$
\begin{aligned}
\mathbb{E}[\mu_{i+1|t}\mu_{i+1|t}^{\mathrm{t}}] &= \mathbb{E}[\mathbf{z}_{i+1}\mathbf{z}_{i+1}^{\mathrm{t}}] - \Sigma_{i+1|t} && (28)\\
&= \mathbf{A}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\mathrm{t}}]\mathbf{A}^{\mathrm{t}} + \mathbf{Q} - \Sigma_{i+1|t} && (29)\\
\mathbb{E}[\mu_{i|i}\mu_{i|i}^{\mathrm{t}}] &= \mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\mathrm{t}}] - \Sigma_{i|i}. && (30)
\end{aligned}
$$

where we have used the one step linear dependence between $\mathbf{z}_i$ and $\mathbf{z}_{i+1}$.

Including these expressions, and rewriting $\mathbf{A}\Sigma_{i|i}\mathbf{A}^{\mathrm{t}} + \mathbf{Q}$ back as $\Sigma_{i+1|i}$, the posterior covariance for smoothing simplifies to the final expression:

$$
\Sigma_{i|t} = \Sigma_{i|i} + \mathbf{F}_i \left(\Sigma_{i+1|t} - \Sigma_{i+1|i}\right) \mathbf{F}_i^{\mathrm{t}}. \qquad (31)
$$

Overall, filtering (forward pass) includes information flowing rightwards from index 1 to t with the current state $\mathbf{z}_i$ being updates as a function of past state $\mathbf{z}_{i-1}$ and current observation $\mathbf{x}_i$. In contrast, during smoothing (backward pass) the current state $\mathbf{z}_i$ is updated based on posterior for the next state $\mathbf{z}_{i+1}$ which already incorporates information from the full sequence.

[4]Here we use the fact that the covariance is 2nd moment - squared mean.