ELEC 548 Kalman Filter

A. Motivation - Continuous Latent Variable Models

We have implemented latent variable models for Classification, Clustering, and Dimensionality Reduction. In all of these cases, the data we were modeling were *static* – they came from a single snap shot or measurement and our models did not incorporate the concept of time or continuity.

However, for many machine learning problems, the observed data *vary with time*. For these problems, we will want our underlying latent variable model to capture the continuity present in the temporal relationship between data points. We want models that are *dynamical*.

B. Linear Dynamical Systems (LDS)

At time step $t = 1, \dots, T$, let:

$$\begin{aligned} \mathbf{z}_t \in \mathbb{R}^M \quad & \text{be the (latent) "state" variable at time } t \\ \mathbf{x}_t \in \mathbb{R}^D \quad & \text{be the observation at time } t \end{aligned}$$

State Model:

The state model describes how the state evolves over time. Similarly to how we used a linear Gaussian model for Dimensionality Reduction, pointing out that both the *linear* and *Gaussian* aspects make this the simplest interesting model, a Linear Dynamical System with Gaussian "innovations" (changes from one time step to the next) is the simplest interesting model for how the underlying state evolves over time.

$$\begin{aligned} \mathbf{z}_t \mid \mathbf{z}_{t-1} &\sim \mathcal{N}(\mathbf{A}\mathbf{z}_{t-1}, \mathbf{Q}) \\ \mathbf{z}_1 &\sim \mathcal{N}(\boldsymbol{\pi}, \mathbf{V}) \end{aligned} \tag{1}$$

Observation Model:

The observation model describes how the observed data relates to the state. Similarly to the state mode, the simplest way of relating the state to our observed data is a linear, Gaussian observation model.

$$\mathbf{x}_t \mid \mathbf{z}_t \sim \mathcal{N}(\mathbf{C}\mathbf{z}_t, \mathbf{R}) \tag{2}$$

Thus, the model parameters are $\theta = \{\mathbf{A}, \mathbf{Q}, \boldsymbol{\pi}, \mathbf{V}, \mathbf{C}, \mathbf{R}$. *Exercise to the reader:* What are the dimensions of each of these parameters?

Markov assumption

Notice that the state model uses a first-order Markov assumption. Specifically, we assume that $\mathbf{z}_{t-1}$, the state at time $t-1$, contains all the relevant information necessary to predict $\mathbf{z}_t$, the state at time t. This means that we can write down a simple formula for the joint probability of the state over all times:

$$\Pr(\mathbf{z}_1, \dots, \mathbf{z}_T) = \Pr(\mathbf{z}_1) \Pr(\mathbf{z}_2 \mid \mathbf{z}_1) \Pr(\mathbf{z}_3 \mid \mathbf{z}_2, \cancel{\mathbf{z}_1}) \dots \Pr(\mathbf{z}_T \mid \mathbf{z}_1, \cancel{\mathbf{z}_2 \dots \mathbf{z}_{T-1}}) \qquad \text{(Markov assumption)}$$

$$= \Pr(\mathbf{z}_1) \prod_{t=2}^{T} \Pr(\mathbf{z}_t \mid \mathbf{z}_{t-1})$$

B.1 Training phase

Goal: Estimate the model parameters $\theta = \{\mathbf{A}, \mathbf{Q}, \boldsymbol{\pi}, \mathbf{V}, \mathbf{C}, \mathbf{R}$ from the training data.

unsupervised learning If the values of the state variables $\mathbf{z}_t$ are **unknown** during training, use the **EM algorithm** Maximize $\Pr(\{\mathbf{x}\} \mid \theta)$ with respect to θ.

supervised learning If the values of the state variables $\mathbf{z}_t$ are **known** (the simpler case) during training Maximize $\Pr(\{\mathbf{x}\}, \{\mathbf{z}\} \mid \theta)$ with respect to θ.

We will consider the simpler, *supervised learning* case, where the $\mathbf{z}_t$ are known during training. This makes sense if we are designing, for example, a real time neural signal processing system in which the thing we wish to decode (e.g., arm trajectories) can be measured during trianing. In this context, the latent variable is known during training and unknown during testing/operation.

(i). Maxmimum Likelihood Parameters in Supervised Training

$$\Pr(\{\mathbf{x}\}, \{\mathbf{z}\} \mid \theta) = \Pr(\mathbf{z}_1) \prod_{t=2}^{T} \Pr(\mathbf{z}_t \mid \mathbf{z}_{t-1}) \Big(\prod_{t=1}^{T} \Pr(\mathbf{x}_t \mid \mathbf{x}_t) \Big)$$

Writing down the training data log-likelihood, we have

$$\begin{aligned}
\mathcal{L}(\theta) &= \log \Pr(\{\mathbf{x}\}, \{\mathbf{z}\} \mid \theta) \\
&= \log \Pr(\mathbf{z}_1) + \sum_{t=2}^{T} \log \Pr(\mathbf{z}_t \mid \mathbf{z}_{t-1}) + \sum_{t=1}^{T} \log \Pr(\mathbf{x}_t \mid \mathbf{x}_t) \\
&= -\frac{M}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{V}| - \frac{1}{2} (\mathbf{z}_1 - \boldsymbol{\pi})^\mathsf{T} \mathbf{V}^{-1} (\mathbf{z}_1 - \boldsymbol{\pi}) \\
&\quad + \sum_{t=2}^{T} \Big(-\frac{M}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{Q}| - \frac{1}{2} (\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1})^\mathsf{T} \mathbf{Q}^{-1} (\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1}) \Big) \\
&\quad + \sum_{t=1}^{T} \Big(-\frac{D}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{R}| - \frac{1}{2} (\mathbf{x}_t - \mathbf{C}\mathbf{z}_t)^\mathsf{T} \mathbf{R}^{-1} (\mathbf{x}_t - \mathbf{C}\mathbf{z}_t) \Big)
\end{aligned}$$

Solving for **A** (which describes the *dynamics* of the state variable):

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta)}{\partial \mathbf{A}} &= \frac{\partial}{\partial \mathbf{A}} \Big\{ \sum_{t=2}^{T} \Big(-\frac{1}{2} (\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1})^\mathsf{T} \mathbf{Q}^{-1} (\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1}) \Big) \Big\} \\
&= -\frac{1}{2} \frac{\partial}{\partial \mathbf{A}} \Big\{ \sum_{t=2}^{T} \Big(-\mathbf{z}_{t-1}^\mathsf{T} \mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{z}_t - \mathbf{z}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{A}\mathbf{z}_{t-1} + \mathbf{z}_{t-1}^\mathsf{T} \mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{A}\mathbf{z}_{t-1} \Big) \Big\} \\
&= -\frac{1}{2} \frac{\partial}{\partial \mathbf{A}} \Big\{ \mathrm{Tr} \left(\mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \sum_{t=2}^{T} \mathbf{z}_t \mathbf{z}_{t-1}^\mathsf{T} \right) + \mathrm{Tr} \left(\mathbf{A} \sum_{t=2}^{T} \mathbf{z}_t \mathbf{z}_{t-1}^\mathsf{T} \mathbf{Q}^{-1} \right) + \mathrm{Tr} \left(\mathbf{Q}^{-1} \mathbf{A} \sum_{t=2}^{T} \mathbf{z}_t \mathbf{z}_{t-1}^\mathsf{T} \mathbf{A}^\mathsf{T} \right) \Big\}
\end{aligned}$$

$$= -\frac{1}{2}\left(-\mathbf{Q}^{-1}\sum_{t=2}^{T}\mathbf{z}_t\mathbf{z}_{t-1}^{\mathsf{T}} - \mathbf{Q}^{-1}\sum_{t=2}^{T}\mathbf{z}_t\mathbf{z}_{t-1}^{\mathsf{T}} + \mathbf{Q}^{-1}\mathbf{A}\sum_{t=2}^{T}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\mathsf{T}} + \mathbf{Q}^{-1}\mathbf{A}\sum_{t=2}^{T}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^{\mathsf{T}}\right)$$
$$= 0$$

Here, we have made use of the fact that because its a symmetric covariance matrix, $\mathbf{Q}^{-\mathsf{T}} = \mathbf{Q}^{-1}$, and $\frac{d}{d\mathbf{X}}\mathrm{Tr}(\mathbf{X}\mathbf{A}^{\mathsf{T}}) = \frac{d}{d\mathbf{X}}\mathrm{Tr}(\mathbf{X}^{\mathsf{T}}\mathbf{A}) = \mathbf{A}$ and $\frac{d}{d\mathbf{X}}\mathrm{Tr}(\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{X}^{\mathsf{T}}\mathbf{C}) = \mathbf{A}^{\mathsf{T}}\mathbf{C}^{\mathsf{T}}\mathbf{X}\mathbf{B}^{\mathsf{T}} + \mathbf{C}\mathbf{A}\mathbf{X}\mathbf{B}$. Solving, we have

$$\mathbf{A} = \left(\sum_{t=2}^{T}\mathbf{z}_t\mathbf{z}_{t-1}\right)\left(\sum_{t=2}^{T}\mathbf{z}_{t-1}\mathbf{z}_{t-1}\right)^{-1} \tag{3}$$

Solving for $\mathbf{Q}$ (which describes the variability of the *innovation*, or change from one time step to the next of the state variable):

$$\frac{\partial\mathcal{L}(\theta)}{\partial\mathbf{Q}} = \frac{\partial}{\partial\mathbf{Q}}\left\{-\frac{T-1}{2}\log|\mathbf{Q}| - \frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}^{-1}\sum_{n=2}^{T}(\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1})(\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1})^{\mathsf{T}}\right)\right\}$$
$$= -\frac{T-1}{2}\mathbf{Q}^{-1} - \frac{1}{2}\left(\mathbf{Q}^{-1}\sum_{n=2}^{T}(\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1})(\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1})^{\mathsf{T}}\mathbf{Q}^{-1}\right) \qquad = 0$$

Here, we have additionally used the fact that $\frac{d}{d\mathbf{X}}\mathrm{Tr}(\mathbf{X}^{-1}\mathbf{A}) = -\mathbf{X}^{-1}\mathbf{A}\mathbf{X}^{-1}$ and $fracdd\mathbf{X}\log|\mathbf{X}| = \mathbf{X}^{-1}$. Solving, we have

$$\mathbf{Q} = \frac{1}{T-1}\sum_{t=2}^{T}(\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1})(\mathbf{z}_t - \mathbf{A}\mathbf{z}_{t-1})^{\mathsf{T}} \tag{4}$$

(using the $\mathbf{A}$ found above.) Note that these solutions are identical to the solutions to linear regreissions!

The solutions for $\frac{\partial\mathcal{L}(\theta)}{\partial\mathbf{C}} = 0$ and $\frac{\partial\mathcal{L}(\theta)}{\partial\mathbf{R}} = 0$ follow very similar math and give us the solutions

$$\mathbf{C} = \left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{z}_t\right)\left(\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t\right)^{-1} \tag{5}$$

and

$$\mathbf{R} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_t - \mathbf{R}\mathbf{z}_t)(\mathbf{x}_t - \mathbf{C}\mathbf{z}_t)^{\mathsf{T}} \tag{6}$$

(using the solution for $\mathbf{C}$ found above).

In these solutions, for simplicity we have considered only one sequence of state and observation variables. In most scenarios, we would have **multiple sequences of training data**, potentially each with a different length (T). If we define $\{\mathbf{x}\}_n$ and $\{\mathbf{z}\}_n$ as the n-th training sequence ($n = 1, \ldots, N$), then the goal for training would be to find the parameters, $theta$, which maximize $\prod_{n=1}^{N}\Pr(\{\mathbf{x}\}_n, \{\mathbf{z}\}_n \mid \theta)$.

The maximimum likelihood solutions in the **multiple sequence** training case for equations (3) – (6) have the same form, but each summation is over more elements. The solution is *almost* the same as concatenating all the sequences together, with the exception that the terms in the dynamics equations, (3) and (4) that would involve $\mathbf{z}_{T,n}$ and $\mathbf{z}_{1,n+1}$ are not included.

In the case of **multiple sequences**, we can also solve for the initial state distribution. $\boldsymbol{\pi}$ and $\mathbf{V}$ are the sample mean and covariance, respectively, of the N instances of $\mathbf{z}_1$.

C. Test Phase / Decoding

Goal: Compute $\mathsf{Pr}(\mathbf{z}_t \mid \mathbf{x}_1, \ldots, \mathbf{x}_t)$ for $t = 1, \ldots, T$.

The variables $\mathbf{z}_1, \ldots, \mathbf{z}_T, \mathbf{x}_1, \ldots, \mathbf{x}_T$ are jointly Gaussian, so $\mathsf{Pr}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^T)$ is Gaussian (using the notation $\{\mathbf{x}\}_1^T = \mathbf{x}_1, \ldots, \mathbf{x}_t$). Thus, we only need to find its mean and covariance.

We can compute $\mathsf{Pr}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^T)$ recursively starting at $t = 1$.

One-step (forward) prediction

$$\underbrace{\mathsf{Pr}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t-1})}_{\text{}} = \int \underbrace{\mathsf{Pr}(\mathbf{z}_t \mid \mathbf{z}_{t-1})}_{\text{state model}} \underbrace{\mathsf{Pr}(\mathbf{z}_{t-1} \mid \{\mathbf{x}\}_1^{t-1})}_{\text{}} \tag{7}$$

Measurement update

$$\overbrace{\mathsf{Pr}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t})}^{\text{}} = \frac{\overbrace{\mathsf{Pr}(\mathbf{z}_t \mid \mathbf{z}_t)}^{\text{obs. model}} \overbrace{\mathsf{Pr}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t-1})}^{\text{}}}{\mathsf{Pr}(\mathbf{x}_t \mid \{\mathbf{x}\}_1^{t-1})} \tag{8}$$

Note that equations (7) and (8) are always a valid way of describing a dynamical system with Markovian properties. When we specify "linear" and "Gaussian", then each component is Gaussian, and all that we need to calculate are the mean and covariance. Define

$$\begin{aligned} \boldsymbol{\mu}_t^\tau &= \mathsf{E}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^\tau) \\ \boldsymbol{\Sigma}_t^\tau &= \mathsf{Cov}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^\tau) \end{aligned}$$

One-step prediction

$$\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_t^{t-1}, \boldsymbol{\Sigma}_t^{t-1})$$

So we need to find $\boldsymbol{\mu}_t^{t-1}$ and $\boldsymbol{\Sigma}_t^{t-1}$. We can equivalently write (1) as

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$$

Thus,

$$\begin{aligned} \boldsymbol{\mu}_t^{t-1} &= \mathsf{E}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t-1}) \\ &= \mathbf{A}\,\mathsf{E}(\mathbf{z}_{t-1} \mid \{\mathbf{x}\}_1^{t}) + \cancel{\mathsf{E}(\mathbf{v}_t \mid \{\mathbf{x}\}_1^{t-1})} \end{aligned}$$

$$\boldsymbol{\mu}_t^{t-1} = \mathbf{A}\boldsymbol{\mu}_{t-1}^{t-1} \tag{9}$$

and

$$\begin{aligned}
\boldsymbol{\Sigma}_t^{t-1} &= \mathrm{Cov}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t-1}) \\
&= \mathrm{E}((\mathbf{A}\mathbf{z}_{t-1} + \mathbf{v}_t)(\mathbf{A}\mathbf{z}_{t-1} + \mathbf{v}_t)^\mathsf{T}) - \mathrm{E}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t-1})\,\mathrm{E}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t-1})^\mathsf{T} \\
&= \mathrm{E}(\mathbf{A}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^\mathsf{T}\mathbf{A}^\mathsf{T} + \cancel{\mathbf{v}_t\mathbf{z}_{t-1}^\mathsf{T}\mathbf{A}^\mathsf{T}} + \cancel{\mathbf{A}\mathbf{z}_{t-1}\mathbf{v}_t^\mathsf{T}} + \mathbf{v}_t\mathbf{v}_t)^\mathsf{T}) - \mathbf{A}\boldsymbol{\mu}_{t-1}^{t-1}(\boldsymbol{\mu}_{t-1}^{t-1})^\mathsf{T}\mathbf{A}^\mathsf{T} \\
&= \mathbf{A}\,\mathrm{Cov}(\mathbf{z}_{t-1} \mid \{\mathbf{x}\}_1^t)\mathbf{A}^\mathsf{T} + \mathrm{Cov}(\mathbf{v}_t \mid \{\mathbf{x}\}_1^{t-1})
\end{aligned}$$

$$\boldsymbol{\Sigma}_t^{t-1} = \mathbf{A}\boldsymbol{\Sigma}_{t-1}^{t-1}\mathbf{A}^\mathsf{T} + \mathbf{Q} \tag{10}$$

Measurement update

$$\mathbf{z}_t \mid \{\mathbf{x}\}_1^t \sim \mathcal{N}(\boldsymbol{\mu}_t^t, \boldsymbol{\Sigma}_t^t)$$

So we need to find $\boldsymbol{\mu}_t^t$ and $\boldsymbol{\Sigma}_t^t$. Notice that (8) is just Bayes rule for $\mathbf{z}_t$ given $\mathbf{x}_t$ with everything conditioned on $\{\mathbf{x}\}_1^{t-1}$. So inspired by what we learned in our analysis of Dimensionality Reduction (P-PCA) about conditional Gaussian distributions from the joint distribution, lets start by finding the joint distribution $\Pr(\mathbf{z}_t, \mathbf{x}_t \mid \{\mathbf{x}\}_1^{t-1})$.

We will start by noting the equivalent form of (2) is

$$\mathbf{x}_t = \mathbf{C}\mathbf{z}_t + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$$

We need $\mathrm{E}(\mathbf{x}_t \mid \{\mathbf{x}\}_1^{t-1})$, $\mathrm{Cov}(\mathbf{x}_t \mid \{\mathbf{x}\}_1^{t-1})$, and the cross covariance $\mathrm{E}(\mathbf{x}_t\mathbf{z}_t^\mathsf{T} \mid \{\mathbf{x}\}_1^{t-1})$.

$$\begin{aligned}
\mathrm{E}(\mathbf{x}_t \mid \{\mathbf{x}\}_1^{t-1}) &= \mathbf{C}\,\mathrm{E}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t-1}) + \cancel{\mathrm{E}(\mathbf{w}_t \mid \{\mathbf{x}\}_1^{t-1})} \\
&= \mathbf{C}\boldsymbol{\mu}_t^{t-1}
\end{aligned} \tag{11}$$

$$\begin{aligned}
\mathrm{Cov}(\mathbf{x}_t \mid \{\mathbf{x}\}_1^{t-1}) &= \mathbf{C}\,\mathrm{Cov}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t-1})\mathbf{C}^\mathsf{T} + \mathrm{Cov}(\mathbf{w}_t \mid \{\mathbf{x}\}_1^{t-1}) + \cancel{\mathrm{E}(\text{cross terms})} \\
&= \mathbf{C}\boldsymbol{\Sigma}_t^{t-1}\mathbf{C}^\mathsf{T} + \mathbf{R}
\end{aligned} \tag{12}$$

$$\mathrm{E}(\mathbf{x}_t\mathbf{z}_t^\mathsf{T} \mid \{\mathbf{x}\}_1^{t-1}) - \mathrm{E}(\mathbf{x}_t \mid \{\mathbf{x}\}_1^{t-1})\,\mathrm{E}(\mathbf{z}_t \mid \{\mathbf{x}\}_1^{t-1})^\mathsf{T} \tag{13}$$

$$= \mathbf{C}\,\mathrm{E}(\mathbf{z}_t\mathbf{z}_t^\mathsf{T} \mid \{\mathbf{x}\}_1^{t-1}) + \cancel{\mathrm{E}(\mathbf{v}_t\mathbf{z}_t^\mathsf{T} \mid \{\mathbf{x}\}_1^{t-1})} - \mathbf{C}\boldsymbol{\mu}_t^{t-1}\boldsymbol{\mu}_t^{t-1^\mathsf{T}} \tag{14}$$

$$= \mathbf{C}\boldsymbol{\Sigma}_t^{t-1} \tag{15}$$

Filling in, we have

$$\begin{bmatrix}\mathbf{x}_t \\ \mathbf{z}_t\end{bmatrix} \mid \{\mathbf{x}\}_1^{t-1} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{C}\boldsymbol{\mu}_t^{t-1} \\ \boldsymbol{\mu}_t^{t-1}\end{bmatrix}, \begin{bmatrix}\mathbf{C}\boldsymbol{\Sigma}_t^{t-1}\mathbf{C}^\mathsf{T} + \mathbf{R} & \mathbf{C}\boldsymbol{\Sigma}_t^{t-1} \\ \boldsymbol{\Sigma}_t^{t-1}\mathbf{C}^\mathsf{T} & \boldsymbol{\Sigma}_t^{t-1}\end{bmatrix}\right)$$

Now, applying the rule for conditioning in jointly Gaussian random variables (see Dimensionality Reduction notes), we have

$$\begin{aligned}\boldsymbol{\mu}_t^t &= \mathsf{E}(\mathbf{z}_t \mid \mathbf{x}_t, \{\mathbf{x}\}_1^{t-1}) \\ &= \boldsymbol{\mu}_t^{t-1} + \underbrace{\boldsymbol{\Sigma}_t^{t-1}\mathbf{C}^\mathsf{T}(\mathbf{C}\boldsymbol{\Sigma}_t^{t-1}\mathbf{C}^\mathsf{T} + \mathbf{R})^{-1}}_{\mathbf{K}_t \equiv \text{“Kalman gain”}}(\mathbf{x}_t - \mathbf{C}\boldsymbol{\mu}_t^{t-1})\end{aligned}$$