Kalman Filter

A) Motivation

All of the models that we have talked about so far are static (i.e. they do not incorporate the concept of time).

However, many types of data (including neural data) vary with time.

In such settings, we will want to apply models that can capture temporal relationships between data points (i.e. a dynamical model)

B) Linear Dynamical Systems (LDS)

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

$\underline{z}_t \in \mathbb{R}^M$ be the state variable
(e.g. arm position and velocity)

$\underline{x}_t \in \mathbb{R}^D$ be the observation
(e.g. spike count vector)

State model:

$$\underline{z}_t \mid \underline{z}_{t-1} \sim N(A\underline{z}_{t-1}, Q) \qquad (1)$$

$$\underline{z}_1 \sim N(\underline{\pi}, V)$$

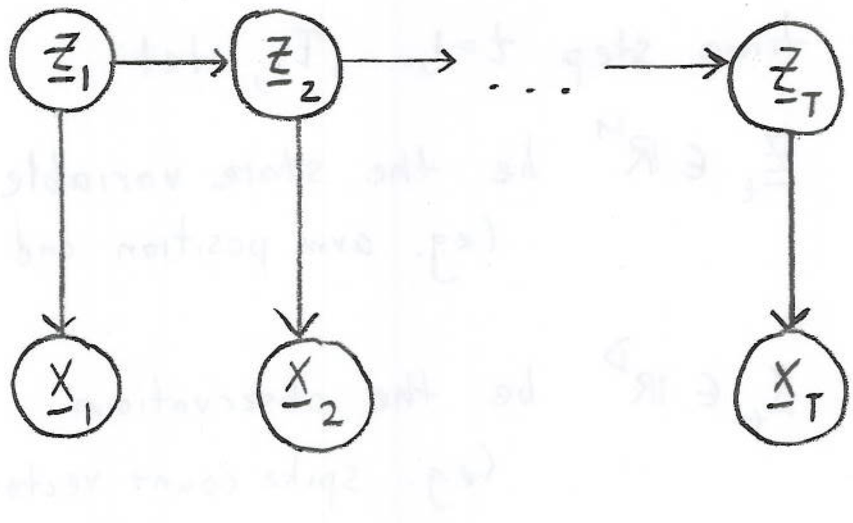
Observation model:

$$\underline{x}_t \mid \underline{z}_t \sim N(C\underline{z}_t, R) \qquad (2)$$

Model parameters are $\theta = \{A, Q, \underline{\pi}, V, C, R\}$

What are the dimensions of each of these parameters?

Graphical model:



- State model describes how state evolves over time.
- Observation model describes how observation relates to the state.
- State model uses a <u>Markov</u> assumption:

$$P(\underline{z}_1, \ldots, \underline{z}_T) = P(\underline{z}_1) P(\underline{z}_2 | \underline{z}_1) \ldots P(\underline{z}_T | \underline{z}_1, \ldots, \underline{z}_{T-1})$$

$$= P(\underline{z}_1) \prod_{t=2}^{T} P(\underline{z}_t | \underline{z}_{t-1})$$ Markov assumption

C) <u>Training phase</u>

<u>Goal</u>: Estimate the model parameters $\theta = \{A, Q, \underline{\pi}, V, C, R\}$ from the training data.

- If the values of the state variables $\underline{z}_t$ are <u>unknown</u> during training, use EM algorithm (unsupervised learning).

Maximize $P(\{\underline{x}\} | \theta)$ w.r.t. θ.

- Here, we will consider the simpler case, where the $\underline{z}_t$ are <u>known</u> during training (supervised learning).

Maximize $P(\{\underline{x}\}, \{\underline{z}\} | \theta)$ w.r.t. θ.

For decoding arm trajectories from neural activity, the $\underline{z}_t$ (arm states) are typically known during training.

$$P(\{\underline{x}\}, \{\underline{z}\} | \theta) = P(\underline{z}_1) \prod_{t=2}^{T} P(\underline{z}_t | \underline{z}_{t-1}) \left(\prod_{t=1}^{T} P(\underline{x}_t | \underline{z}_t) \right)$$

Let

$$\begin{aligned}
\mathcal{L}(\theta) &= \log P(\{\underline{x}\}, \{\underline{z}\} | \theta) \\
&= \log P(\underline{z}_1) + \sum_{t=2}^{T} \log P(\underline{z}_t | \underline{z}_{t-1}) + \sum_{t=1}^{T} \log P(\underline{x}_t | \underline{z}_t) \\
&= -\frac{M}{2} \log(2\pi) - \frac{1}{2} \log |V| - \frac{1}{2} (\underline{z}_1 - \underline{\pi})^T V^{-1} (\underline{z}_1 - \underline{\pi}) \\
&\quad + \sum_{t=2}^{T} \left(-\frac{M}{2} \log(2\pi) - \frac{1}{2} \log |Q| - \frac{1}{2} (\underline{z}_t - A \underline{z}_{t-1})^T Q^{-1} (\cdot) \right) \\
&\quad + \sum_{t=1}^{T} \left(-\frac{D}{2} \log(2\pi) - \frac{1}{2} \log |R| - \frac{1}{2} (\underline{x}_t - C \underline{z}_t)^T R^{-1} (\cdot) \right)
\end{aligned}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial A} = \frac{\partial}{\partial A}\left\{ \sum_{t=2}^{T}\left(-\underline{z}_{t-1}^{T} A^{T} Q^{-1} \underline{z}_{t} - \underline{z}_{t}^{T} Q^{-1} A \underline{z}_{t-1} + \underline{z}_{t-1}^{T} A^{T} Q^{-1} A \underline{z}_{t-1} \right) \right\}$$

$$= \frac{\partial}{\partial A}\left\{ -\operatorname{Tr}\left(A^{T} Q^{-1} \sum_{t=2}^{T} \underline{z}_{t} \underline{z}_{t-1}^{T} \right) - \operatorname{Tr}\left(A\left(\sum_{t=2}^{T} \underline{z}_{t-1} \underline{z}_{t}^{T} \right) Q^{-1} \right) + \operatorname{Tr}\left(Q^{-1} A \left(\sum_{t=2}^{T} \underline{z}_{t-1} \underline{z}_{t-1}^{T} \right) A^{T} \right) \right\}$$

$$= -Q^{-1}\left(\sum_{t=2}^{T} \underline{z}_{t} \underline{z}_{t-1}^{T} \right) - Q^{-1}\left(\sum_{t=2}^{T} \underline{z}_{t} \underline{z}_{t-1}^{T} \right) + Q^{-1} A \left(\sum_{t=2}^{T} \underline{z}_{t-1} \underline{z}_{t-1}^{T} \right) + Q^{-1} A \left(\sum_{t=2}^{T} \underline{z}_{t-1} \underline{z}_{t-1}^{T} \right)$$

$$= [0]$$

$$\boxed{A = \left(\sum_{t=2}^{T} \underline{z}_{t} \underline{z}_{t-1}^{T} \right)\left(\sum_{t=2}^{T} \underline{z}_{t-1} \underline{z}_{t-1}^{T} \right)^{-1}} \qquad (3)$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial Q} = \frac{\partial}{\partial Q}\left\{ -\frac{(T-1)}{2}\log|Q| - \frac{1}{2}\operatorname{Tr}\left(Q^{-1} \sum_{n=2}^{T} (\underline{z}_{t} - A\underline{z}_{t-1})(\underline{z}_{t} - A\underline{z}_{t-1})^{T} \right) \right\}$$

$$= -\frac{(T-1)}{2} Q^{-1} - \frac{1}{2}\left(-Q^{-1} \sum_{n=2}^{T} (\underline{z}_{t} - A\underline{z}_{t-1})(\underline{z}_{t} - A\underline{z}_{t-1})^{T} Q^{-1} \right)$$

$$= [0]$$

$$\boxed{Q = \frac{1}{T-1} \sum_{t=2}^{T} (\underline{z}_t - A\underline{z}_{t-1})(\underline{z}_t - A\underline{z}_{t-1})^T} \qquad (4)$$

use the A found in (3)

<u>Note</u>: The expressions for A and Q are entirely analogous to those in linear regression, where A is the "slope" and Q is the "minimum mean squared error".
(See p. 26 in Dimensionality Reduction (part 2) notes)

Similarly,

$$\frac{\partial \mathcal{L}(\theta)}{\partial C} = [0] \quad \text{yields}$$

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

$$\frac{\partial \mathcal{L}(\theta)}{\partial R} = [0] \quad \text{yields}$$

$$\boxed{R = \frac{1}{T} \sum_{t=1}^{T} (\underline{x}_t - C\underline{z}_t)(\underline{x}_t - C\underline{z}_t)^T} \qquad (6)$$

use the C found in (5).

For notational simplicity, we consider only one sequence $\begin{pmatrix} \underline{z}_1, \ldots, \underline{z}_T \\ \underline{x}_1, \ldots, \underline{x}_T \end{pmatrix}$ here. In general, there may be multiple sequences, each with a different number of time steps.

Let $\{\underline{x}\}_n$, $\{\underline{z}\}_n$ represent the nth sequence $(n = 1, \ldots, N)$

Now, the goal is to maximize $\prod_{n=1}^{N} P(\{\underline{x}\}_n, \{\underline{z}\}_n \mid \theta)$ w.r.t. θ.

The resulting expressions for (3) through (6) have the same form, but each summation sums over more elements.

$\underline{\Pi}$ and V are the sample mean and covariance, respectively, of the N instances of $\underline{z}_1$.

D) <u>Test phase</u>: Decoding arm trajectories from neural activity

<u>Goal</u>: To compute $P(\underline{z}_t \mid \underbrace{\underline{x}_1, \ldots, \underline{x}_t}_{\text{we will use the shorthand } \{\underline{x}\}_1^t})$ for $t = 1, \ldots, T$.

The variables $\underline{z}_1, \ldots, \underline{z}_T, \underline{x}_1, \ldots, \underline{x}_T$ are jointly Gaussian, so $P(\underline{z}_t \mid \{\underline{x}\}_1^t)$ is Gaussian.

Thus, we need only find its mean and covariance.

We can compute $P(\underline{z}_t \mid \{\underline{x}\}_1^t)$ recursively starting at $t=1$:

- One-step prediction

$$P(\underline{z}_t \mid \{\underline{x}\}_1^{t-1}) = \int \underbrace{P(\underline{z}_t \mid \underline{z}_{t-1})}_{\text{State model}} \underbrace{P(\underline{z}_{t-1} \mid \{\underline{x}\}_1^{t-1})} \, d\underline{z}_{t-1} \tag{7}$$

- Measurement update

$$\underbrace{P(\underline{z}_t \mid \{\underline{x}\}_1^t)}_{\text{our goal}} = \frac{\overbrace{P(\underline{x}_t \mid \underline{z}_t)}^{\text{obs model}} \overbrace{P(\underline{z}_t \mid \{\underline{x}\}_1^{t-1})}}{P(\underline{x}_t \mid \{\underline{x}\}_1^{t-1})} \tag{8}$$

Let

$$\underline{\mu}_t^{\tau} = E\left[\underline{z}_t \mid \{\underline{x}\}_1^{\tau}\right]$$

$$\Sigma_t^{\tau} = \operatorname{cov}\left(\underline{z}_t \mid \{\underline{x}\}_1^{\tau}\right)$$

We want to express (7) and (8) in terms of the model parameters.

Plugging the state and observation models into (7) and (8), then simplifying, is a method that will always work.

Here, we will recognize that all the distributions in (7) and (8) are Gaussian, and just solve for means and covariances.

- One-step prediction

$$\underline{z}_t \mid \{\underline{x}\}_1^{t-1} \sim N\left(\underline{\mu}_t^{t-1}, \Sigma_t^{t-1}\right)$$

Find $\underline{\mu}_t^{t-1}$ and Σ_t^{t-1}.

An equivalent way of writing (1) is

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

$$\begin{aligned}\underline{\mu}_t^{t-1} &= E\left[\underline{z}_t \mid \{\underline{x}\}_1^{t-1}\right] \\ &= A\, E\left[\underline{z}_{t-1} \mid \{\underline{x}\}_1^{t-1}\right] + E\left[\cancel{\underline{v}_t} \mid \{\underline{x}\}_1^{t-1}\right]\end{aligned}$$

$$\boxed{\underline{\mu}_t^{t-1} = A\,\underline{\mu}_{t-1}^{t-1}} \qquad (9)$$

$$\Sigma_t^{t-1} = \operatorname{cov}(\underline{z}_t \mid \{\underline{x}\}_1^{t-1})$$

$$= A \operatorname{cov}(\underline{z}_{t-1} \mid \{\underline{x}\}_1^{t-1}) A^T + \operatorname{cov}(\underline{v}_t \mid \{\underline{x}\}_1^{t-1})$$

$$\boxed{\Sigma_t^{t-1} = A \Sigma_{t-1}^{t-1} A^T + Q} \qquad (10)$$

- Measurement update

$$\underline{z}_t \mid \{\underline{x}\}_1^t \sim N(\underline{\mu}_t^t, \Sigma_t^t)$$

Find $\underline{\mu}_t^t$ and Σ_t^t.

Recognize that (8) is Bayes rule for $\underline{z}_t$ and $\underline{x}_t$, with all terms conditioned on $\{\underline{x}\}_1^{t-1}$.

Thus, we will first find the joint probability of $\underline{z}_t$ and $\underline{x}_t$ <u>given</u> $\{\underline{x}\}_1^{t-1}$, then apply the results of conditioning for jointly Gaussian random vars.

a)

$$\begin{bmatrix} \underline{x}_t \\ \underline{z}_t \end{bmatrix} \Bigg| \{\underline{x}\}_1^{t-1} \sim N\left(\begin{bmatrix} C\underline{\mu}_t^{t-1} \\ \underline{\mu}_t^{t-1} \end{bmatrix}, \begin{bmatrix} C\Sigma_t^{t-1}C^T + R & C\Sigma_t^{t-1} \\ \Sigma_t^{t-1}C^T & \Sigma_t^{t-1} \end{bmatrix} \right) \qquad (11)$$

b)

An equivalent way of writing (2) is

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

a) $$E[\underline{x}_t \mid \{\underline{x}\}_1^{t-1}] = C\,E[\underline{z}_t \mid \{\underline{x}\}_1^{t-1}] + \cancel{E[\underline{w}_t \mid \{\underline{x}\}_1^{t-1}]}$$
$$= C\underline{\mu}_t^{t-1}$$

$$\mathrm{cov}(\underline{x}_t \mid \{\underline{x}\}_1^{t-1}) = C\,\mathrm{cov}(\underline{z}_t \mid \{\underline{x}\}_1^{t-1})\,C^T + \mathrm{cov}(\underline{w}_t \mid \{\underline{x}\}_1^{t-1})$$
$$= C\Sigma_t^{t-1}C^T + R$$

b) $$E[\underline{x}_t\underline{z}_t^T \mid \{\underline{x}\}_1^{t-1}] - E[\underline{x}_t \mid \{\underline{x}\}_1^{t-1}]\,E[\underline{z}_t \mid \{\underline{x}\}_1^{t-1}]^T$$
$$= E\left[C\underline{z}_t\underline{z}_t^T + \underline{w}_t\underline{z}_t^T \mid \{\underline{x}\}_1^{t-1}\right] - C\underline{\mu}_t^{t-1}\underline{\mu}_t^{t-1\,T}$$
$$= C\left(\Sigma_t^{t-1} + \underline{\mu}_t^{t-1}\underline{\mu}_t^{t-1\,T}\right) + \cancel{E[\underline{w}_t \mid \{\underline{x}\}_1^{t-1}]}\,E[\underline{z}_t \mid \{\underline{x}\}_1^{t-1}]^T$$
$$- C\underline{\mu}_t^{t-1}\underline{\mu}_t^{t-1\,T}$$
$$= C\Sigma_t^{t-1}$$

Applying the results of conditioning for jointly Gaussian random variables to (11),

(see p. 22 in Dimensionality Reduction (part 2) notes)

$$\mu_t^t = E\left[\underline{z}_t \mid \underline{x}_t, \{\underline{x}\}_1^{t-1} \right]$$

$$= \mu_t^{t-1} + \underbrace{\Sigma_t^{t-1} C^T \left(C \Sigma_t^{t-1} C^T + R \right)^{-1}}_{\text{call this } K_t\text{, the "Kalman gain"}} \left(\underline{x}_t - C \mu_t^{t-1} \right)$$

Rewriting,

$$\boxed{\mu_t^t = \mu_t^{t-1} + K_t \left(\underline{x}_t - C \mu_t^{t-1} \right)} \qquad (12)$$

$$\Sigma_t^t = \text{cov}\left(\underline{z}_t \mid \underline{x}_t, \{\underline{x}\}_1^{t-1} \right)$$

$$= \Sigma_t^{t-1} - \Sigma_t^{t-1} C^T \left(C \Sigma_t^{t-1} C^T + R \right)^{-1} C \Sigma_t^{t-1}$$

$$\boxed{\Sigma_t^t = \Sigma_t^{t-1} - K_t C \Sigma_t^{t-1}} \qquad (13)$$

Taking the recursions defined by (9), (10), (12), (13), we obtain μ_t^t and Σ_t^t for $t=1,\ldots,T$.

- μ_t^t is the arm state estimate at time t
- Σ_t^t is our uncertainty around that estimate at time t.

Initialize recursions with
$$\mu_1^0 = \underline{\Pi}$$
$$\Sigma_1^0 = V.$$

Appendix: Useful matrix properties

$$\frac{d}{dX} \mathrm{Tr}\left(X A^T \right) = \frac{d}{dX} \mathrm{Tr}\left(X^T A \right) = A$$

$$\frac{d}{dX} \mathrm{Tr}\left(A X B X^T C \right) = A^T C^T X B^T + C A X B$$

$$\frac{d}{dX} \log |X| = X^{-T}$$

$$\frac{d}{dX} \mathrm{Tr}\left(X^{-1} A \right) = - X^{-T} A^T X^{-T}$$