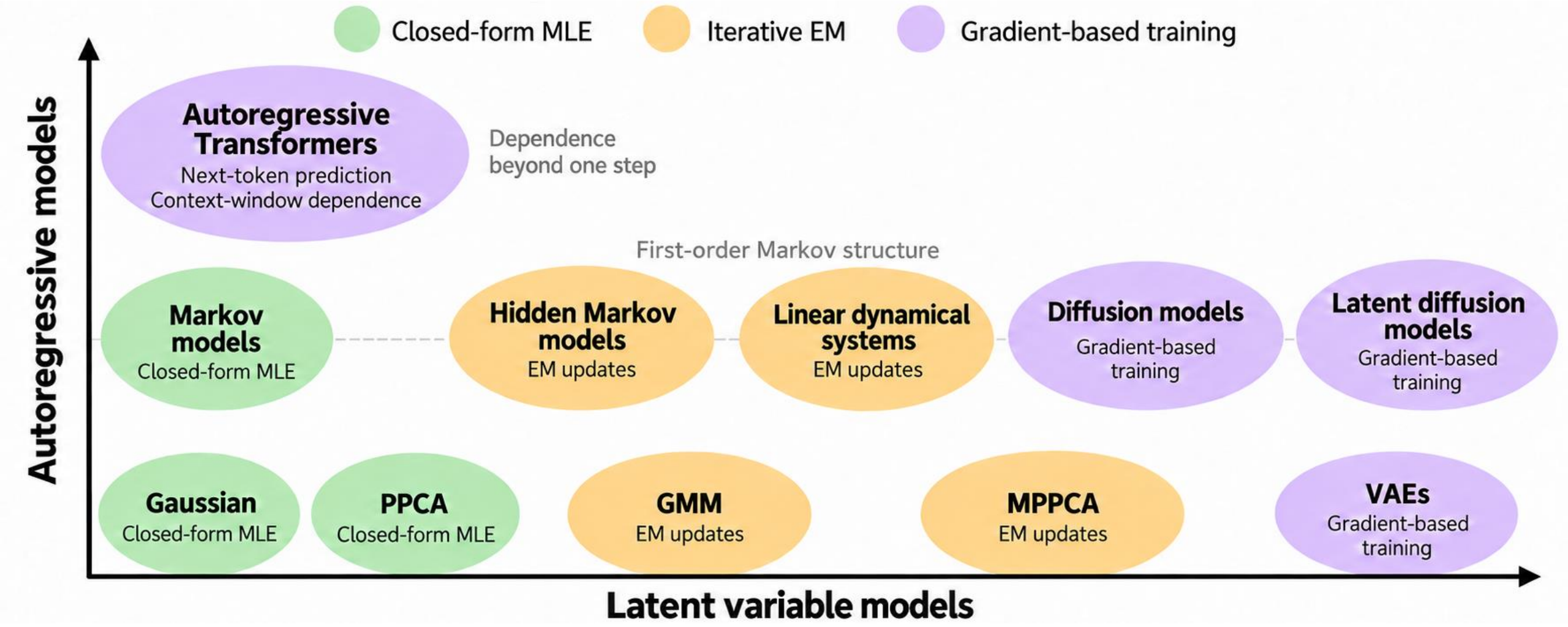
Deep Generative Models: Linear Dynamical Systems

Fall Semester 2026

René Vidal
Director of the Center for Innovation in Data Engineering and Science (IDEAS),
Rachleff University Professor, University of Pennsylvania
Amazon Scholar & Chief Scientist at NORCE

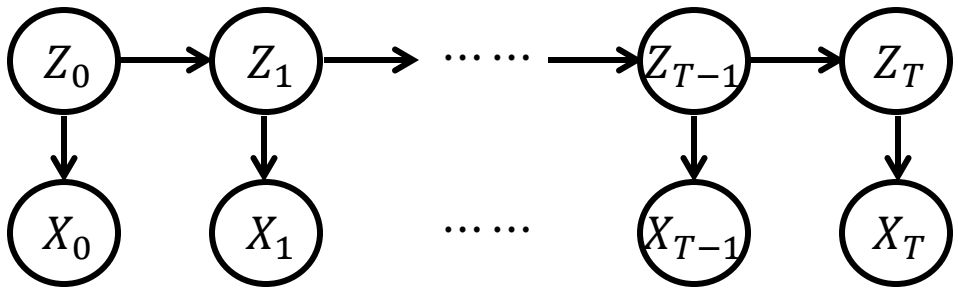


Taxonomy of Generative Models



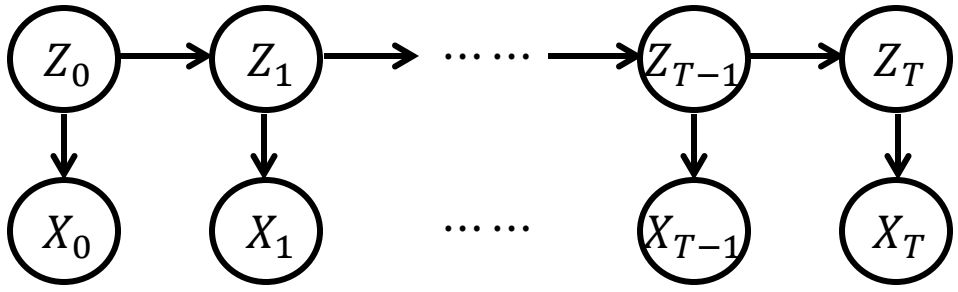
HMMs and Linear Dynamical Systems

Hidden Markov Models (HMMs)



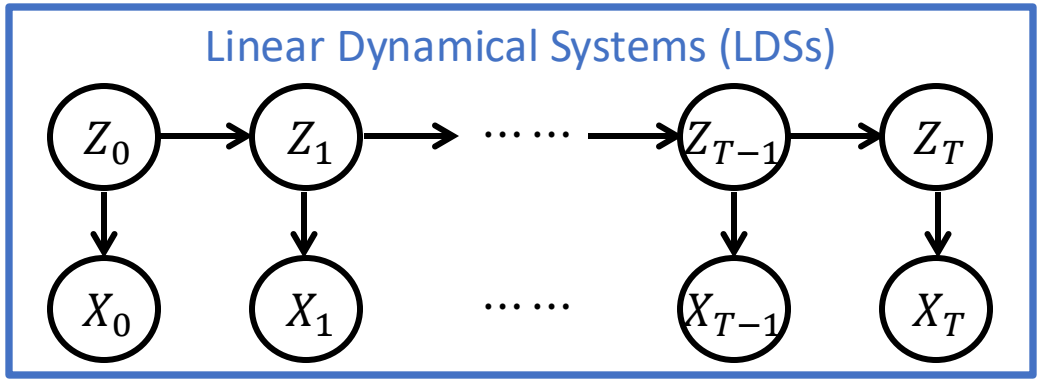
- Hidden state Z_t and observation X_t are discrete random variables
- State transition and emission probabilities are categorical

Linear Dynamical Systems (LDSs)



- Hidden state Z_t and observation X_t are continuous random vectors
- State transition and emission probabilities are Gaussian

Linear Dynamical Systems



- Hidden state Z_t and observation X_t are continuous random vectors
- State transition and emission are Gaussian

Model Parameters:
- $\theta := (\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{C}, \boldsymbol{R})$

d: state dimension
D: output dimension

$\boldsymbol{S}_0 \in \mathbb{R}^{d\times d}$
$\boldsymbol{S}_0 \succ 0$

$\boldsymbol{A} \in \mathbb{R}^{d\times d}$
$\boldsymbol{C} \in \mathbb{R}^{D\times d}$

$\boldsymbol{Q} \in \mathbb{R}^{d\times d}$
$\boldsymbol{Q} \succ 0$

$\boldsymbol{R} \in \mathbb{R}^{D\times D}$
$\boldsymbol{R} \succ 0$

- Initial Distribution:

$$p(Z_0) = \mathcal{N}(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0)$$

- State Transition:

$$p(Z_t \mid \boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q})$$

We assume $\boldsymbol{w}_t, \boldsymbol{v}_t$ are independent from each other and independent from $\boldsymbol{z}_0$

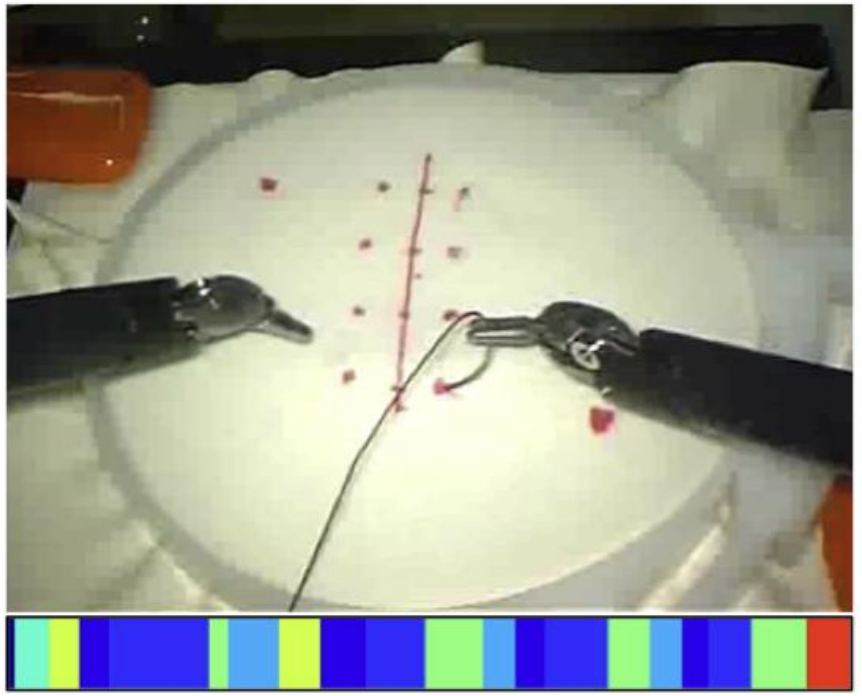
- State Emission:

$$p(X_t \mid \boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{C}\boldsymbol{z}_t, \boldsymbol{R})$$

	Probabilistic Description	Algebraic Description
State Transition	$p(Z_t \mid \boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q})$	$Z_t = \boldsymbol{A}Z_{t-1} + \boldsymbol{w}_t$ with $\boldsymbol{w}_t \sim \mathcal{N}(0, \boldsymbol{Q})$
State Emission	$p(X_t \mid \boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{C}\boldsymbol{z}_t, \boldsymbol{R})$	$X_t = \boldsymbol{C}Z_t + \boldsymbol{v}_t$ with $\boldsymbol{v}_t \sim \mathcal{N}(0, \boldsymbol{R})$

Example: Force Estimation in Surgical Procedures

- A surgical task (e.g., suturing) is decomposed as a sequence of gestures (e.g., grab the needle, insert the needle, pull the suture)



- X_t : trajectory (position and velocity) of each one of the joints of the Davinci robot
- Z_t : forces and accelerations in the end effector
- **Inference**: estimate the forces given the trajectories
- **Learning**: given kinematic trajectories, learn the parameters (A, C, Q, R)

Filtering and Smoothing

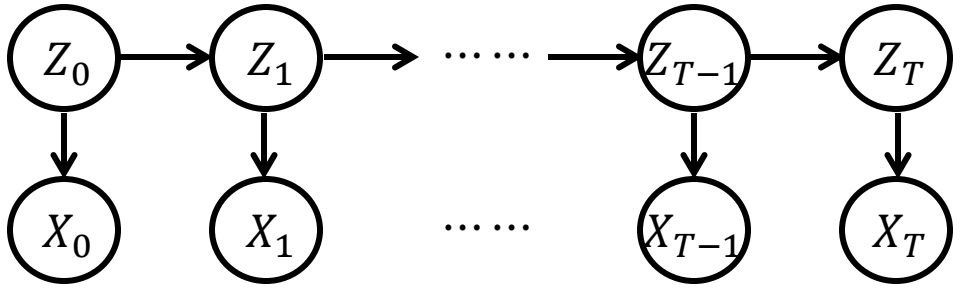
- P1: Filtering. Given θ and $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$, infer the current state $\boldsymbol{z}_t$, that is to compute
$$p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$$
  - e.g., what is the current force in the end effector given the joint positions thus far?

- P2: Smoothing. Given θ and $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$, infer the past state $\boldsymbol{z}_t$, that is to compute
$$p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$$
  - e.g., what was the force in the end effector at time t given we full trajectory of all the joints?

- Remark. You may find P1 and P2 familiar
  - In HMMs, we solved them via recursively updating $\alpha_j(t)$, $\gamma_j(t)$

Basic Properties

Linear Dynamical Systems (LDSs)



- Hidden state Z_t and observation X_t are continuous random vectors
- State transition and emission are Gaussian

Model Parameters:
- $\theta := (\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{C}, \boldsymbol{R})$

d: state dimension
D: output dimension

$\boldsymbol{S}_0 \in \mathbb{R}^{d\times d}$
$\boldsymbol{S}_0 \succ 0$

$\boldsymbol{A} \in \mathbb{R}^{d\times d}$
$\boldsymbol{C} \in \mathbb{R}^{D\times d}$

$\boldsymbol{Q} \in \mathbb{R}^{d\times d}$
$\boldsymbol{Q} \succ 0$

$\boldsymbol{R} \in \mathbb{R}^{D\times D}$
$\boldsymbol{R} \succ 0$

- Joint distribution is Gaussian:

Gaussian

$$p_\theta(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T, \boldsymbol{z}_0, \ldots, \boldsymbol{z}_T) = p_\theta(\boldsymbol{z}_0) \prod_{t=0}^{T} p_\theta(\boldsymbol{x}_t \mid \boldsymbol{z}_t) \prod_{t=1}^{T} p_\theta(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1})$$

- Therefore, "any conditional distribution" is Gaussian
  - Vague, but look at your question 1 of homework 1 (next page)

Gaussian Conditioning (HW 1)

- If $\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}\right)$, then the conditional distribution $p(\boldsymbol{a} \mid \boldsymbol{b})$ is Gaussian with mean $\boldsymbol{\mu}_{a|b}$ and covariance $\boldsymbol{\Sigma}_{a|b}$ given by

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{b} - \boldsymbol{\mu}_b)$$

original mean & variance of $\boldsymbol{a}$

correction upon observing $\boldsymbol{b}$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

Gaussian Combining (Extension of Problem 1a, HW 1)

Gaussian Combining: If $\boldsymbol{x} = \boldsymbol{C}\boldsymbol{z} + \boldsymbol{v}$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ and $\boldsymbol{v} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_v)$ then

$$\begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_z \\ \boldsymbol{C}\boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_z \boldsymbol{C}^\top \\ \boldsymbol{C}\boldsymbol{\Sigma}_z & \boldsymbol{C}\boldsymbol{\Sigma}_z \boldsymbol{C}^\top + \boldsymbol{\Sigma}_v \end{bmatrix} \right)$$

- Proof: The proof is finished by computing the following quantities:
  - $\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\boldsymbol{C}\boldsymbol{z} + \boldsymbol{v}] = \boldsymbol{C}\mathbb{E}[\boldsymbol{z}] + 0 = \boldsymbol{C}\boldsymbol{\mu}_z$
  - $\mathrm{Cov}(\boldsymbol{z}, \boldsymbol{x}) = \mathbb{E}[(\boldsymbol{z} - \boldsymbol{\mu}_z)(\boldsymbol{x} - \boldsymbol{C}\boldsymbol{\mu}_z)^\top] = \mathbb{E}[(\boldsymbol{z} - \boldsymbol{\mu}_z)(\boldsymbol{C}\boldsymbol{z} + \boldsymbol{v} - \boldsymbol{C}\boldsymbol{\mu}_z)^\top] = \boldsymbol{\Sigma}_z \boldsymbol{C}^\top$
  - $\mathrm{Cov}(\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{x} - \boldsymbol{C}\boldsymbol{\mu}_z)(\boldsymbol{x} - \boldsymbol{C}\boldsymbol{\mu}_z)^\top] = \mathbb{E}[(\boldsymbol{C}\boldsymbol{z} + \boldsymbol{v} - \boldsymbol{C}\boldsymbol{\mu}_z)(\boldsymbol{C}\boldsymbol{z} + \boldsymbol{v} - \boldsymbol{C}\boldsymbol{\mu}_z)^\top]$
    $= \mathbb{E}[\boldsymbol{C}(\boldsymbol{z} - \boldsymbol{\mu}_z)(\boldsymbol{z} - \boldsymbol{\mu}_z)^\top \boldsymbol{C}^\top + \boldsymbol{C}(\boldsymbol{z} - \boldsymbol{\mu}_z)\boldsymbol{v}^\top + \boldsymbol{v}(\boldsymbol{z} - \boldsymbol{\mu}_z)^\top \boldsymbol{C}^\top + \boldsymbol{v}\boldsymbol{v}^\top]$
    $= \boldsymbol{C} \cdot \mathbb{E}[(\boldsymbol{z} - \boldsymbol{\mu}_z)(\boldsymbol{z} - \boldsymbol{\mu}_z)^\top] \cdot \boldsymbol{C}^\top + 0 + 0 + \boldsymbol{R} = \boldsymbol{C}\boldsymbol{\Sigma}_z \boldsymbol{C}^\top + \boldsymbol{\Sigma}_v$

Filtering and Smoothing

- P1: Filtering. Given θ and $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$, compute

$$p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$$

- P2: Smoothing. Given θ and $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$, compute

$$p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$$

- Since $p_\theta(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$ is Gaussian $(\forall s, t)$, it suffices to compute

$$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$$
$$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathbb{E}\left[(\boldsymbol{z}_s - \hat{\boldsymbol{z}}_{s|t})(\boldsymbol{z}_s - \hat{\boldsymbol{z}}_{s|t})^\top \middle| \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t\right] = \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$$

- We will do so recursively (first for filtering and then for smoothing)

Filtering: Compute $\hat{\boldsymbol{z}}_{0|0}, \hat{\boldsymbol{\Sigma}}_{0|0}$

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

- P1: Filtering. Given θ and $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$, compute
$$p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$$

- Let's begin with the simplest case:
  - What are the mean $\hat{\boldsymbol{z}}_{0|0} = \mathbb{E}[\boldsymbol{z}_0 \mid \boldsymbol{x}_0]$ and covariance $\hat{\boldsymbol{\Sigma}}_{0|0} = \mathrm{Cov}(\boldsymbol{z}_0 \mid \boldsymbol{x}_0)$ of $\boldsymbol{z}_0$ given $\boldsymbol{x}_0$?

- High-level Idea.
  1. find the mean and covariance of $\begin{bmatrix}\boldsymbol{z}_0 \\ \boldsymbol{x}_0\end{bmatrix}$ via Gaussian combining
  2. find the mean $\hat{\boldsymbol{z}}_{0|0}$ and covariance $\hat{\boldsymbol{\Sigma}}_{0|0}$ of $\boldsymbol{z}_0 \mid \boldsymbol{x}_0$ via Gaussian conditioning

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{b} - \boldsymbol{\mu}_b),\ \boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

Filtering: Compute $\hat{\boldsymbol{z}}_{0|0}$, $\hat{\boldsymbol{\Sigma}}_{0|0}$

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

$p(Z_0) = \mathcal{N}(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0)$
$p(X_t \mid \boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{C}\boldsymbol{z}_t, \boldsymbol{R})$

- Step 1: find the mean and covariance of $\begin{bmatrix} \boldsymbol{z}_0 \\ \boldsymbol{x}_0 \end{bmatrix}$

Gaussian Combining: If $\boldsymbol{x} = \boldsymbol{C}\boldsymbol{z} + \boldsymbol{v}$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ and $\boldsymbol{v} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_v)$ then

$$\begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_z \\ \boldsymbol{C}\boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_z \boldsymbol{C}^\top \\ \boldsymbol{C}\boldsymbol{\Sigma}_z & \boldsymbol{C}\boldsymbol{\Sigma}_z \boldsymbol{C}^\top + \boldsymbol{\Sigma}_v \end{bmatrix} \right)$$

- Applying Gaussian combining yields

$$\begin{bmatrix} \boldsymbol{z}_0 \\ \boldsymbol{x}_0 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\pi}_0 \\ \boldsymbol{C}\boldsymbol{\pi}_0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_0 \boldsymbol{C}^\top \\ \boldsymbol{C}\boldsymbol{\Sigma}_0 & \boldsymbol{C}\boldsymbol{\Sigma}_0 \boldsymbol{C}^\top + \boldsymbol{R} \end{bmatrix} \right)$$

Filtering: Compute $\hat{\boldsymbol{z}}_{0|0}$, $\hat{\boldsymbol{\Sigma}}_{0|0}$

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

- Step 2: apply Gaussian conditioning to $\begin{bmatrix} \boldsymbol{z}_0 \\ \boldsymbol{x}_0 \end{bmatrix}$

$$\begin{bmatrix} \boldsymbol{z}_0 \\ \boldsymbol{x}_0 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\pi}_0 \\ \boldsymbol{C}\boldsymbol{\pi}_0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_0\boldsymbol{C}^\top \\ \boldsymbol{C}\boldsymbol{\Sigma}_0 & \boldsymbol{C}\boldsymbol{\Sigma}_0\boldsymbol{C}^\top + \boldsymbol{R} \end{bmatrix} \right)$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{b} - \boldsymbol{\mu}_b),\ \boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$$

- We have

$$\hat{\boldsymbol{z}}_{0|0} = \mathbb{E}[\boldsymbol{z}_0 \mid \boldsymbol{x}_0] = \boldsymbol{\pi}_0 + \boldsymbol{\Sigma}_0\boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{\Sigma}_0\boldsymbol{C}^\top + \boldsymbol{R})^{-1}(\boldsymbol{x}_0 - \boldsymbol{C}\boldsymbol{\pi}_0)$$

$$\hat{\boldsymbol{\Sigma}}_{0|0} = \mathrm{Cov}(\boldsymbol{z}_0 \mid \boldsymbol{x}_0) = \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{\Sigma}_0\boldsymbol{C}^\top + \boldsymbol{R})^{-1}\boldsymbol{C}\boldsymbol{\Sigma}_0$$

Filtering: From 0 to t

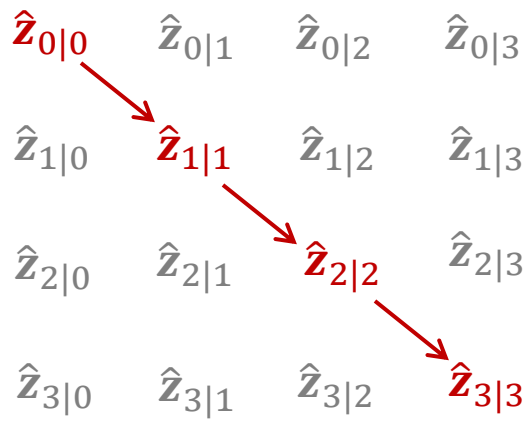
$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

- P1: Filtering. Given θ and $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$, compute
$$p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$$

- We've now computed $\hat{\boldsymbol{z}}_{0|0}$ and $\hat{\boldsymbol{\Sigma}}_{0|0}$. This solves P1 for the case $t = 0$

- To proceed, we will update $\hat{\boldsymbol{z}}_{t-1|t-1}$, $\hat{\boldsymbol{\Sigma}}_{t-1|t-1}$ into $\hat{\boldsymbol{z}}_{t|t}$, $\hat{\boldsymbol{\Sigma}}_{t|t}$ for every t

The planned computational flow



Filtering: From 0 to t

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

- To compute $\hat{\boldsymbol{z}}_{0|0}$ and $\hat{\boldsymbol{\Sigma}}_{0|0}$, we
  - (Step 0) found the mean and covariance of $\boldsymbol{z}_0$ (already known)
  - (Step 1) found the mean and covariance of $\begin{bmatrix}\boldsymbol{z}_0\\ \boldsymbol{x}_0\end{bmatrix}$ via Gaussian combining
  - (Step 2) found the mean and covariance of $\boldsymbol{z}_0 \mid \boldsymbol{x}_0$ via Gaussian conditioning

Question: How can we generalize these steps for general t?

- To update $\hat{\boldsymbol{z}}_{t-1|t-1}, \hat{\boldsymbol{\Sigma}}_{t-1|t-1}$ into $\hat{\boldsymbol{z}}_{t|t}, \hat{\boldsymbol{\Sigma}}_{t|t}$, we will condition on $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$ and
  - (Step 0) find the mean and covariance of $\boldsymbol{z}_t \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$ (using $\hat{\boldsymbol{z}}_{t-1|t-1}, \hat{\boldsymbol{\Sigma}}_{t-1|t-1}$)
  - (Step 1) find the mean and covariance of $\begin{bmatrix}\boldsymbol{z}_t\\ \boldsymbol{x}_t\end{bmatrix} \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$ via Gaussian combining
  - (Step 2) find the mean and covariance of $\boldsymbol{z}_t \mid \boldsymbol{x}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$ via Gaussian conditioning

Step 0 and conditioning on $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$ are the only differences
- You should be able to figure out all the details without looking at the rest slides

Filtering: Compute $\hat{\boldsymbol{z}}_{t|t}, \hat{\boldsymbol{\Sigma}}_{t|t}$

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_t)$

$p(Z_t \mid \boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q})$

- Step 0: find the mean and covariance of $\boldsymbol{z}_t | \boldsymbol{x}_0, \dots, \boldsymbol{x}_{t-1}$
  - By definition, this is to compute $\hat{\boldsymbol{z}}_{t|t-1}, \hat{\boldsymbol{\Sigma}}_{t|t-1}$

- We have

$$\hat{\boldsymbol{z}}_{t|t-1} = \mathbb{E}[\boldsymbol{z}_t \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_{t-1}] = \mathbb{E}[\boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{w}_t \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_{t-1}] = \boldsymbol{A}\hat{\boldsymbol{z}}_{t-1|t-1}$$

$$\hat{\boldsymbol{\Sigma}}_{t|t-1} = \mathbb{E}\left[\left(\boldsymbol{z}_t - \hat{\boldsymbol{z}}_{t|t-1}\right)\left(\boldsymbol{z}_t - \hat{\boldsymbol{z}}_{t|t-1}\right)^{\top} \middle| \boldsymbol{x}_0, \dots, \boldsymbol{x}_{t-1}\right] = \cdots = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{t-1|t-1}\boldsymbol{A}^{\top} + \boldsymbol{Q}$$

similar to how we computed $\mathrm{Cov}(\boldsymbol{x})$ in the proof of Gaussian combining

Filtering: Compute $\hat{\boldsymbol{z}}_{t|t}, \hat{\boldsymbol{\Sigma}}_{t|t}$

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

$\hat{\boldsymbol{z}}_{t|t-1} = \boldsymbol{A}\hat{\boldsymbol{z}}_{t-1|t-1}$
$\hat{\boldsymbol{\Sigma}}_{t|t-1} = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{t-1|t-1}\boldsymbol{A}^\top + \boldsymbol{Q}$

$\hat{\boldsymbol{z}}_{0|-1} \coloneqq \boldsymbol{\pi}_0, \ \hat{\boldsymbol{\Sigma}}_{0|-1} \coloneqq \boldsymbol{\Sigma}_0$

- Step 1: find the mean and covariance of $\begin{bmatrix}\boldsymbol{z}_t \\ \boldsymbol{x}_t\end{bmatrix} | \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$
- We applied Gaussian combining to $\begin{bmatrix}\boldsymbol{z}_0 \\ \boldsymbol{x}_0\end{bmatrix}$ and obtained
  - $\begin{bmatrix}\boldsymbol{z}_0 \\ \boldsymbol{x}_0\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\hat{\boldsymbol{z}}_{0|-1} \\ \boldsymbol{C}\hat{\boldsymbol{z}}_{0|-1}\end{bmatrix}, \begin{bmatrix}\hat{\boldsymbol{\Sigma}}_{0|-1} & \hat{\boldsymbol{\Sigma}}_{0|-1}\boldsymbol{C}^\top \\ \boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{0|-1} & \boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{0|-1}\boldsymbol{C}^\top + \boldsymbol{R}\end{bmatrix}\right)$
- Similarly, now, applying Gaussian combining to $\begin{bmatrix}\boldsymbol{z}_t \\ \boldsymbol{x}_t\end{bmatrix} | \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$ gives:
  - $\begin{bmatrix}\boldsymbol{z}_t \\ \boldsymbol{x}_t\end{bmatrix} | \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1} \sim \mathcal{N}\left(\begin{bmatrix}\hat{\boldsymbol{z}}_{t|t-1} \\ \boldsymbol{C}\hat{\boldsymbol{z}}_{t|t-1}\end{bmatrix}, \begin{bmatrix}\hat{\boldsymbol{\Sigma}}_{t|t-1} & \hat{\boldsymbol{\Sigma}}_{t|t-1}\boldsymbol{C}^\top \\ \boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{t|t-1} & \boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{t|t-1}\boldsymbol{C}^\top + \boldsymbol{R}\end{bmatrix}\right)$

Filtering: Compute $\hat{\boldsymbol{z}}_{t|t}, \hat{\boldsymbol{\Sigma}}_{t|t}$

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

$\hat{\boldsymbol{z}}_{t|t-1} = \boldsymbol{A}\hat{\boldsymbol{z}}_{t-1|t-1}$
$\hat{\boldsymbol{\Sigma}}_{t|t-1} = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{t-1|t-1}\boldsymbol{A}^\top + \boldsymbol{Q}$

$\hat{\boldsymbol{z}}_{0|-1} \coloneqq \boldsymbol{\pi}_0,\ \hat{\boldsymbol{\Sigma}}_{0|-1} \coloneqq \boldsymbol{\Sigma}_0$

- Step 2: apply Gaussian conditioning to $\begin{bmatrix}\boldsymbol{z}_t\\ \boldsymbol{x}_t\end{bmatrix} \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$

- We applied Gaussian conditioning to $\begin{bmatrix}\boldsymbol{z}_0\\ \boldsymbol{x}_0\end{bmatrix}$ and obtained:

$$\hat{\boldsymbol{z}}_{0|0} = \hat{\boldsymbol{z}}_{0|-1} + \hat{\boldsymbol{\Sigma}}_{0|-1}\boldsymbol{C}^\top\left(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{0|-1}\boldsymbol{C}^\top + \boldsymbol{R}\right)^{-1}(\boldsymbol{x}_0 - \boldsymbol{C}\hat{\boldsymbol{z}}_{0|-1})$$

$$\hat{\boldsymbol{\Sigma}}_{0|0} = \hat{\boldsymbol{\Sigma}}_{0|-1} - \hat{\boldsymbol{\Sigma}}_{0|-1}\boldsymbol{C}^\top\left(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{0|-1}\boldsymbol{C}^\top + \boldsymbol{R}\right)^{-1}\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{0|-1}$$

- Similarly, now, applying Gaussian conditioning to $\begin{bmatrix}\boldsymbol{z}_t\\ \boldsymbol{x}_t\end{bmatrix} \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}$ gives:

$$\hat{\boldsymbol{z}}_{t|t} = \hat{\boldsymbol{z}}_{t|t-1} + \hat{\boldsymbol{\Sigma}}_{t|t-1}\boldsymbol{C}^\top\left(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{t|t-1}\boldsymbol{C}^\top + \boldsymbol{R}\right)^{-1}(\boldsymbol{x}_t - \boldsymbol{C}\hat{\boldsymbol{z}}_{t|t-1})$$

$$\hat{\boldsymbol{\Sigma}}_{t|t} = \hat{\boldsymbol{\Sigma}}_{t|t-1} - \hat{\boldsymbol{\Sigma}}_{t|t-1}\boldsymbol{C}^\top\left(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{t|t-1}\boldsymbol{C}^\top + \boldsymbol{R}\right)^{-1}\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{t|t-1}$$

"Kalman gain matrix". Let us denote it by $\boldsymbol{K}_t$

Summary: Filtering for LDS

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_t)$

$p(Z_0) = \mathcal{N}(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0)$
$p(Z_t \mid \boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q})$
$p(X_t \mid \boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{C}\boldsymbol{z}_t, \boldsymbol{R})$

- Putting everything together gives Kalman Filter:

- Initialization: $\hat{\boldsymbol{z}}_{0|-1} \coloneqq \boldsymbol{\pi}_0,\ \hat{\boldsymbol{\Sigma}}_{0|-1} \coloneqq \boldsymbol{\Sigma}_0$
- Recursion ($\forall t = 0, \dots, T$):

"Correction":

$$\boldsymbol{K}_t = \hat{\boldsymbol{\Sigma}}_{t|t-1}\boldsymbol{C}^\top\left(\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{t|t-1}\boldsymbol{C}^\top + \boldsymbol{R}\right)^{-1}$$

$$\hat{\boldsymbol{z}}_{t|t} = \hat{\boldsymbol{z}}_{t|t-1} + \boldsymbol{K}_t(\boldsymbol{x}_t - \boldsymbol{C}\hat{\boldsymbol{z}}_{t|t-1})$$

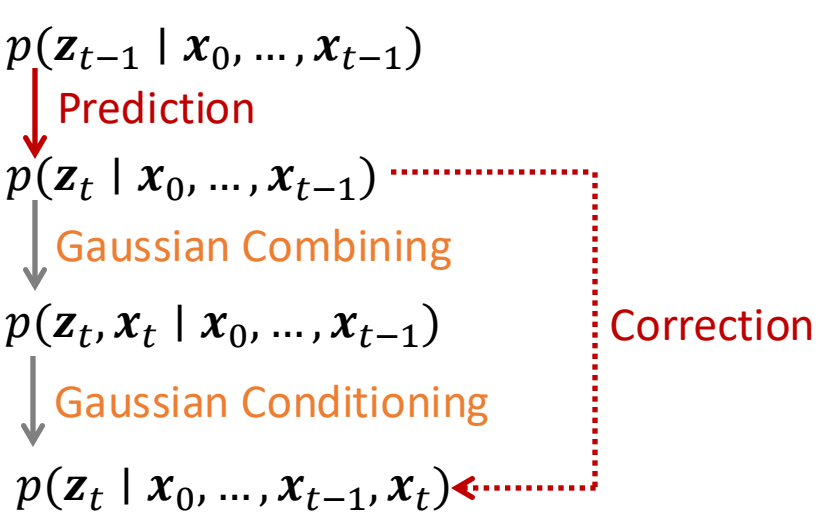
$$\hat{\boldsymbol{\Sigma}}_{t|t} = \hat{\boldsymbol{\Sigma}}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{C}\hat{\boldsymbol{\Sigma}}_{t|t-1}$$

"Prediction":

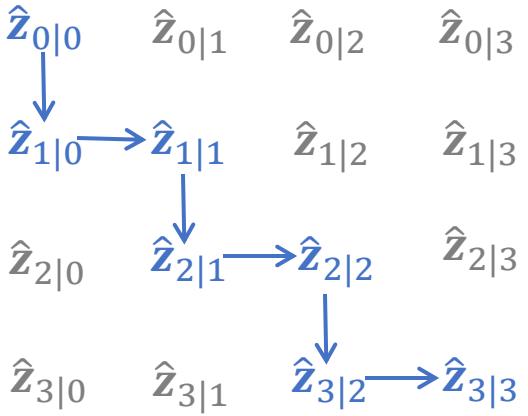
$$\hat{\boldsymbol{z}}_{t+1|t} = \boldsymbol{A}\hat{\boldsymbol{z}}_{t|t}$$

$$\hat{\boldsymbol{\Sigma}}_{t+1|t} = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{t|t}\boldsymbol{A}^\top + \boldsymbol{Q}$$

How To Derive Kalman Filter



Computational Trajectory of Kalman Filter



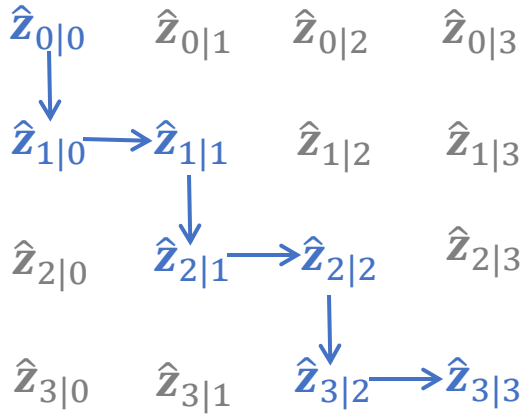
From Filtering to Smoothing

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_t)$

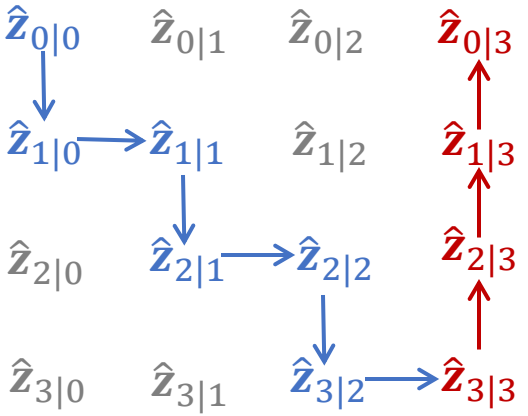
- P1: Filtering. Given θ and $(\boldsymbol{x}_0, \dots, \boldsymbol{x}_t)$, compute
$$p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_t)$$
- P2: Smoothing. Given θ and $(\boldsymbol{x}_0, \dots, \boldsymbol{x}_T)$, compute
$$p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_T)$$

- Since everything is Gaussian, to solve P2 it suffices to compute $\hat{\boldsymbol{z}}_{t|T}, \hat{\boldsymbol{\Sigma}}_{t|T}$ for all t

What Kalman Filter gives us:



What we will do to solve P2:



Goal: Given $\hat{\boldsymbol{z}}_{t|t}, \hat{\boldsymbol{\Sigma}}_{t|t}$ and $\hat{\boldsymbol{z}}_{t|t-1}, \hat{\boldsymbol{\Sigma}}_{t|t-1}$ for every t, update $\hat{\boldsymbol{z}}_{t|T}, \hat{\boldsymbol{\Sigma}}_{t|T}$ into $\hat{\boldsymbol{z}}_{t-1|T}, \hat{\boldsymbol{\Sigma}}_{t-1|T}$

Smoothing

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

Goal: Given $\hat{\boldsymbol{z}}_{t|t}, \hat{\boldsymbol{\Sigma}}_{t|t}$ and $\hat{\boldsymbol{z}}_{t|t-1}, \hat{\boldsymbol{\Sigma}}_{t|t-1}$ for every t, update $\hat{\boldsymbol{z}}_{t|T}, \hat{\boldsymbol{\Sigma}}_{t|T}$ into $\hat{\boldsymbol{z}}_{t-1|T}, \hat{\boldsymbol{\Sigma}}_{t-1|T}$

- Observation:
  - Since $\boldsymbol{z}_{t-1}$ is independent of $\boldsymbol{x}_t, \ldots \boldsymbol{x}_T$ given $\boldsymbol{z}_t$, we have
$$p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}) = p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$$
- High-level Idea:
  1. Compute $p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1})$ via Gaussian combining and Gaussian conditioning
     - This gives us $p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$
  2. Given $p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$, update $\hat{\boldsymbol{z}}_{t|T}, \hat{\boldsymbol{\Sigma}}_{t|T}$ into $\hat{\boldsymbol{z}}_{t-1|T}, \hat{\boldsymbol{\Sigma}}_{t-1|T}$
     - via *law of total probability* and *law of total expectation*
- Derivation Details: see the appendix at the end of slides

Summary: Smoothing for LDSs

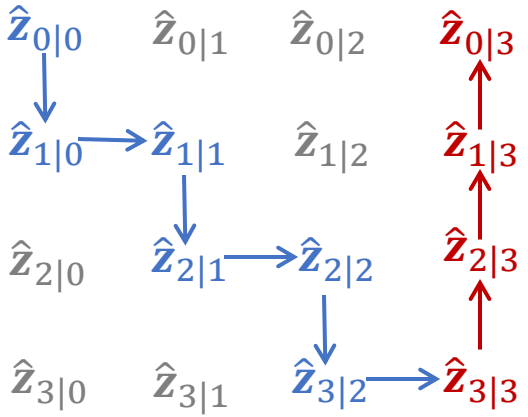
$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

- Goal. Given θ and $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$, compute
$$p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$$

- Algorithm (known as "Rauch-Tung-Striebel smoother").
  1. (Forward Pass) Run Kalman filtering to compute $\hat{\boldsymbol{z}}_{t|t}$, $\hat{\boldsymbol{\Sigma}}_{t|t}$ and $\hat{\boldsymbol{z}}_{t+1|t}$, $\hat{\boldsymbol{\Sigma}}_{t+1|t}$ for all t
  2. (Backward Pass) For $t = T, \ldots, 1$, compute the following:
     - $\boldsymbol{L}_{t-1} = \hat{\boldsymbol{\Sigma}}_{t-1|t-1} \boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_{t|t-1}^{-1}$
     - $\hat{\boldsymbol{z}}_{t-1|T} = \hat{\boldsymbol{z}}_{t-1|t-1} + \boldsymbol{L}_{t-1}\left(\hat{\boldsymbol{z}}_{t|T} - \hat{\boldsymbol{z}}_{t|t-1}\right)$
     - $\hat{\boldsymbol{\Sigma}}_{t-1|T} = \hat{\boldsymbol{\Sigma}}_{t-1|t-1} + \boldsymbol{L}_{t-1}(\hat{\boldsymbol{\Sigma}}_{t|T} - \hat{\boldsymbol{\Sigma}}_{t|t-1})\boldsymbol{L}_{t-1}^\top$

Remark: $\boldsymbol{L}_{t-1}$ is sometimes called the "backward Kalman gain matrix"

Computational Trajectory of Smoothing:



State Estimation and Learning

- State Estimation ("Decoding"). Given θ and $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$, solve:

$$\underset{\boldsymbol{z}_0,\ldots,\boldsymbol{z}_T}{\operatorname{argmax}}\, p_\theta(\boldsymbol{z}_0, \ldots, \boldsymbol{z}_T \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$$

- Learning. Given N observations $\left\{\boldsymbol{x}_{0:T}^{(n)}\right\}_{n=1}^{N}$, find best θ:

$$\max_\theta \prod_{n=1}^{N} p_\theta(\boldsymbol{x}_{0:T}^{(n)})$$

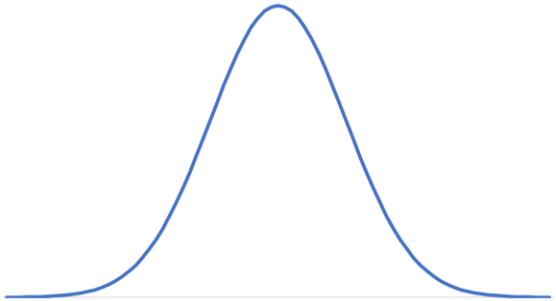
State Estimation

$$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s | \boldsymbol{x}_0, \dots, \boldsymbol{x}_t]$$
$$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s | \boldsymbol{x}_0, \dots, \boldsymbol{x}_t)$$

- State Estimation. Given θ and $(\boldsymbol{x}_0, \dots, \boldsymbol{x}_T)$, solve:

$$\underset{\boldsymbol{z}_0, \dots, \boldsymbol{z}_T}{\operatorname{argmax}} \, p_\theta(\boldsymbol{z}_0, \dots, \boldsymbol{z}_T \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_T)$$

- Solution:
  - Since $p_\theta(\boldsymbol{z}_0, \dots, \boldsymbol{z}_T \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_T)$ is Gaussian, the optimal solution to state estimation is

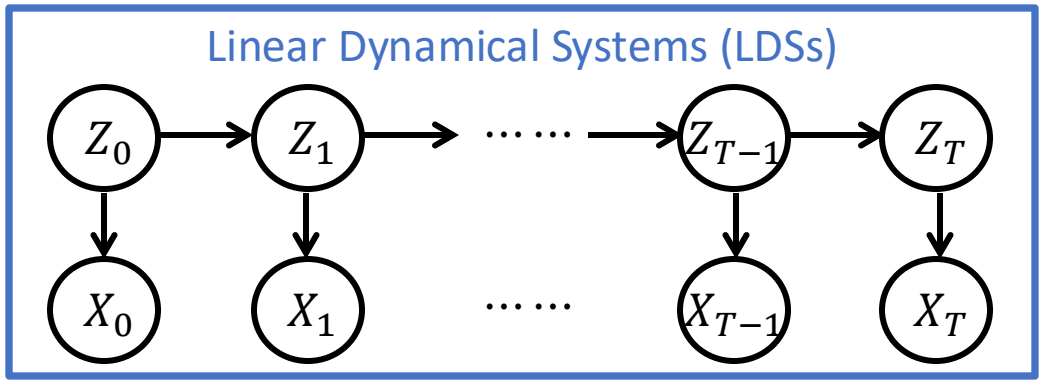


$$\mathbb{E}\left[\begin{bmatrix}\boldsymbol{z}_0\\ \boldsymbol{z}_1\\ \vdots\\ \boldsymbol{z}_T\end{bmatrix} \middle| \boldsymbol{x}_0, \dots, \boldsymbol{x}_T\right] = \begin{bmatrix}\mathbb{E}[\boldsymbol{z}_0 \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_T]\\ \mathbb{E}[\boldsymbol{z}_1 \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_T]\\ \vdots\\ \mathbb{E}[\boldsymbol{z}_T \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_T]\end{bmatrix} = \begin{bmatrix}\hat{\boldsymbol{z}}_{0|T}\\ \hat{\boldsymbol{z}}_{1|T}\\ \vdots\\ \hat{\boldsymbol{z}}_{T|T}\end{bmatrix}$$

  - Therefore, the state estimation problem can be solved by smoothing
  - Similarly, we can prove the output $\hat{\boldsymbol{z}}_{t|t}$ of the Kalman filter solves $\max_{\boldsymbol{z}_t} p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_t)$. It is in this sense that we say the Kalman filter is optimal

Learning

Model Parameters:
- $\theta := (\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{C}, \boldsymbol{R})$



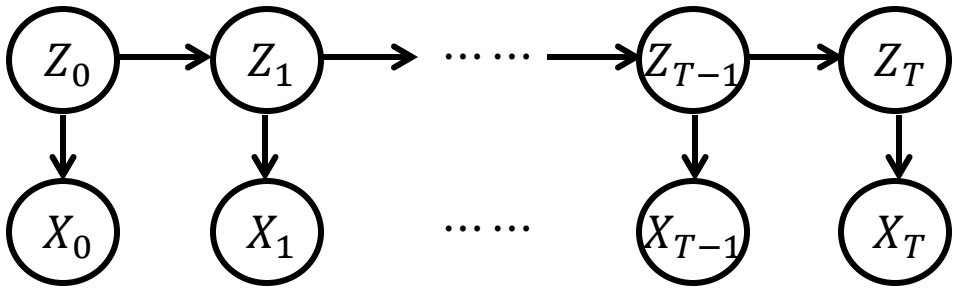
- Learning. Given N observations $\left\{\boldsymbol{x}_{0:T}^{(n)}\right\}_{n=1}^{N}$, find best θ:

$$\max_{\theta} \prod_{n=1}^{N} p_\theta(\boldsymbol{x}_{0:T}^{(n)})$$

- We can perform learning via the EM algorithm
  - we provide high level intuition in what follows
  - full derivations are given at the end of the slides

Learning

Linear Dynamical Systems (LDSs)



Model Parameters:
- $\theta := (\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{C}, \boldsymbol{R})$

- Learning. Given N observations $\left\{\boldsymbol{x}_{0:T}^{(n)}\right\}_{n=1}^{N}$, find best θ:

$$\max_{\theta} \prod_{n=1}^{N} p_\theta(\boldsymbol{x}_{0:T}^{(n)})$$

- We are going to apply the EM algorithm (iteration: k):

E-step:

$$q^k\left(\boldsymbol{z}_{0:T}|\boldsymbol{x}_{0:T}^{(n)}\right) = p_{\theta^k}\left(\boldsymbol{z}_{0:T}\,|\boldsymbol{x}_{0:T}^{(n)}\right)$$

M-step:

$$\theta^{k+1} = \text{argmax}_\theta \sum_{n=1}^{N} \mathbb{E}_{z\sim q^k(z|x^{(n)})}\left[\log p_\theta\left(\boldsymbol{x}_{0:T}^{(n)}, \boldsymbol{z}_{0:T}\right)\right]$$

- we provide high level intuition in what follows
- full derivations are given at the end of the slides

in subscripts: $z := \boldsymbol{z}_{0:T}, x^{(n)} := \boldsymbol{x}_{0:T}^{(n)}$

Guessing

Model Parameters:
- $\theta := (\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{C}, \boldsymbol{R})$

E-step:

$$q^k\left(\boldsymbol{z}_{0:T}|\boldsymbol{x}_{0:T}^{(n)}\right) = p_{\theta^k}\left(\boldsymbol{z}_{0:T}\,|\boldsymbol{x}_{0:T}^{(n)}\right)$$

M-step:

$$\theta^{k+1} = \text{argmax}_\theta \sum_{n=1}^{N} \mathbb{E}_{z\sim q^k(z|x^{(n)})}\left[\log p_\theta\left(\boldsymbol{x}_{0:T}^{(n)}, \boldsymbol{z}_{0:T}\right)\right]$$

$$p(Z_0) = \mathcal{N}(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0) \qquad p(Z_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q}) \qquad p(Y_t|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{C}\boldsymbol{z}_t, \boldsymbol{R})$$

- Let us guess a solution to the M-step… (one of the guesses is wrong)
  - Since $\boldsymbol{\pi}_0 = \mathbb{E}[\boldsymbol{z}_0]$, we guess:

$$\boldsymbol{\pi}_0^{k+1} = \sum_{n=1}^{N} \frac{\mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_0]}{N}$$

← "empirical mean"

  - And we guess the covariance might be

$$\boldsymbol{\Sigma}_0^{k+1} = \sum_{n=1}^{N} \frac{\left(\mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_0] - \boldsymbol{\pi}_0^{k+1}\right)\left(\mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_0] - \boldsymbol{\pi}_0^{k+1}\right)^\top}{N}$$

"empirical covariance" ↖

    - Try having a guess for $\boldsymbol{\Sigma}_0^{k+1}, \boldsymbol{A}^{k+1}, \boldsymbol{Q}^{k+1}, \boldsymbol{C}^{k+1}, \boldsymbol{R}^{k+1}$ yourself…

Summary: EM for LDSs (iteration: k)

- $\hat{z}_{s|t} := \mathbb{E}[z_s | x_0, \dots, x_t]$
- $\hat{\Sigma}_{s|t} := \mathrm{Cov}(z_s | x_0, \dots, x_t)$
- $L_{t-1} := \hat{\Sigma}_{t-1|t-1} A^\top \hat{\Sigma}^{-1}_{t|t-1}$

E-step

Given θ^k, for each $x^{(n)}_{0:T}$, use Kalman Filter & Smoothing to compute:

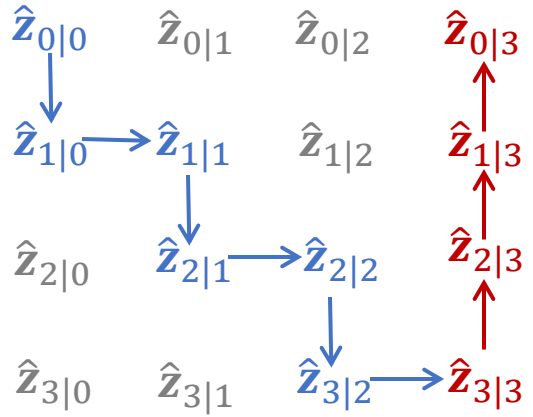
- $\mathbb{E}^{(n)}_k[z_t] := \mathbb{E}_{z \sim q^k(z|x^{(n)})}\left[z_t | x^{(n)}_{0:T}\right]$
- $\mathbb{E}^{(n)}_k[z_t z_t^\top] := \mathbb{E}_{z \sim q^k(z|x^{(n)})}\left[z_t z_t^\top | x^{(n)}_{0:T}\right]$
- $\mathbb{E}^{(n)}_k[z_t z_{t-1}^\top] := \mathbb{E}_{z \sim q^k(z|x^{(n)})}\left[z_t z_{t-1}^\top | x^{(n)}_{0:T}\right]$

add indices n, k

Kalman Filter & Smoothing can compute

- $\mathbb{E}[z_t | x_{0:T}] = \hat{z}_{t|T}$
- $\mathbb{E}[z_t z_t^\top | x_{0:T}] = \hat{\Sigma}_{t|T} + \hat{z}_{t|T}\hat{z}^\top_{t|T}$
- $\mathbb{E}[z_t z_{t-1}^\top | x_{0:T}] = L_{t-1}\hat{\Sigma}_{t|T} + \hat{z}_{t|T}\hat{z}^\top_{t-1|T}$

via forward & backward passes



M-step

Update parameters:

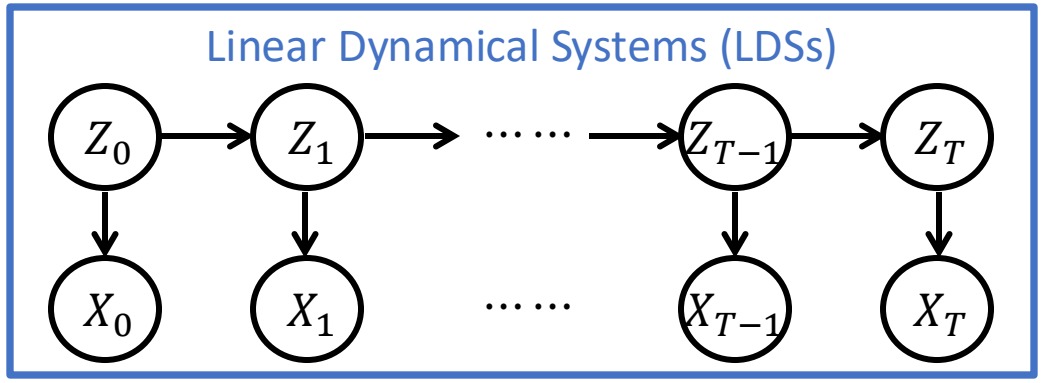
- $\pi_0^{k+1} = \frac{\sum_{n=1}^N \mathbb{E}^{(n)}_k[z_0]}{N}$
- $\Sigma_0^{k+1} = \frac{\sum_{n=1}^N \mathbb{E}^{(n)}_k[z_0 z_0^\top]}{N} - \pi_0^{k+1}(\pi_0^{k+1})^\top$
- $C^{k+1} = \left(\sum_{n=1}^N \sum_{t=0}^T x_t^{(n)} \mathbb{E}^{(n)}_k[z_t]^\top\right)\left(\sum_{n=1}^N \sum_{t=0}^T \mathbb{E}^{(n)}_k[z_t z_t^\top]\right)^{-1}$
- $R^{k+1} = \frac{\sum_{n=1}^N \sum_{t=0}^T x_t^{(n)}\left(x_t^{(n)}\right)^\top - C^{k+1}\sum_{n=1}^N \sum_{t=0}^T \mathbb{E}^{(n)}_k[z_t]\left(x_t^{(n)}\right)^\top}{N(T+1)}$
- $A^{k+1} = \left(\sum_{n=1}^N \sum_{t=1}^T \mathbb{E}^{(n)}_k[z_t z_{t-1}^\top]\right)\left(\sum_{n=1}^N \sum_{t=1}^T \mathbb{E}^{(n)}_k[z_{t-1} z_{t-1}^\top]\right)^{-1}$
- $Q^{k+1} = \frac{\sum_{n=1}^N \sum_{t=0}^T \mathbb{E}^{(n)}_k[z_t z_t^\top] - A^{k+1}\sum_{n=1}^N \sum_{t=0}^T \mathbb{E}^{(n)}_k[z_{t-1} z_t^\top]}{NT}$

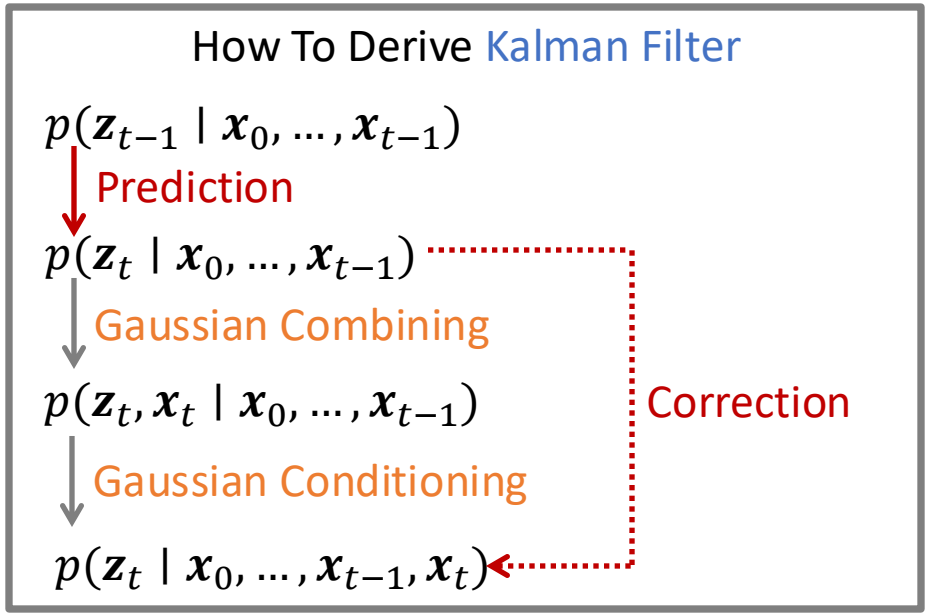
Model Parameters:

- $\theta := (\pi_0, \Sigma_0, A, Q, C, R)$

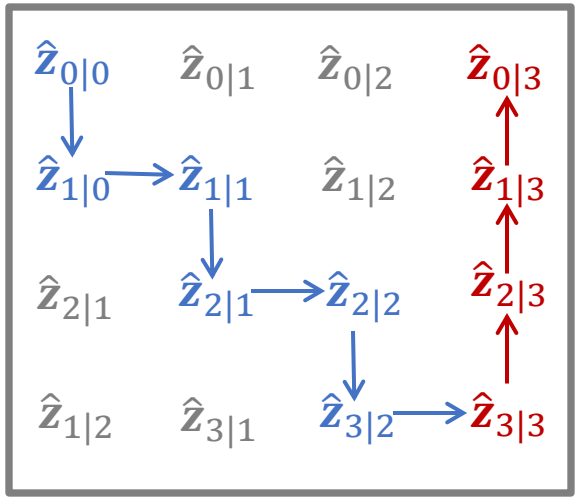
$$p(Z_0) = \mathcal{N}(\pi_0, \Sigma_0)$$
$$p(Z_t | z_{t-1}) = \mathcal{N}(Az_{t-1}, Q)$$
$$p(X_t | z_t) = \mathcal{N}(Cz_t, R)$$

Possible Extensions of Linear Dynamical Systems





- What if ……
  - we do not have Gaussians?
  - we have time-varying dynamics?
    - $Z_t = \boldsymbol{A}_t \boldsymbol{z}_{t-1} + \boldsymbol{w}_t,\ X_t = \boldsymbol{C}_t \boldsymbol{z}_t + \boldsymbol{v}_t$
  - we have control over states?
    - $Z_t = \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{U}q_t + \boldsymbol{w}_t$
  - we have nonlinear dynamics?
    - $Z_t = f(\boldsymbol{z}_{t-1}) + \boldsymbol{w}_t,\ X_t = g(\boldsymbol{z}_t) + \boldsymbol{v}_t$



Appendix (Optional)

- In the subsequent slides, we provide mathematical details for smoothing and learning

Smoothing: Prerequisite

- For smoothing, we need the following basic results:

  - Law of Total Expectation (LoTE)

$$\mathbb{E}[\boldsymbol{a}] = \mathbb{E}_b[\mathbb{E}_a[\boldsymbol{a} \mid \boldsymbol{b}]]$$

  - Law of Total Covariance (LoTC)

$$\mathrm{Cov}(\boldsymbol{a}) = \mathbb{E}[\mathrm{Cov}(\boldsymbol{a} \mid \boldsymbol{b})] + \mathrm{Cov}(\mathbb{E}[\boldsymbol{a} \mid \boldsymbol{b}])$$

$$\mathrm{Cov}(\boldsymbol{a}, \boldsymbol{b}) \coloneqq \mathbb{E}[(\boldsymbol{a} - \mathbb{E}[\boldsymbol{a}])(\boldsymbol{b} - \mathbb{E}[\boldsymbol{b}])^{\mathsf{T}}]$$
$$\mathrm{Cov}(\boldsymbol{a}) \coloneqq \mathrm{Cov}(\boldsymbol{a}, \boldsymbol{a})$$

Smoothing

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

Goal: Given $\hat{\boldsymbol{z}}_{t|t}, \hat{\boldsymbol{\Sigma}}_{t|t}$ and $\hat{\boldsymbol{z}}_{t|t-1}, \hat{\boldsymbol{\Sigma}}_{t|t-1}$ for every t, update $\hat{\boldsymbol{z}}_{t|T}, \hat{\boldsymbol{\Sigma}}_{t|T}$ into $\hat{\boldsymbol{z}}_{t-1|T}, \hat{\boldsymbol{\Sigma}}_{t-1|T}$

- Observation:
  - Since $\boldsymbol{z}_{t-1}$ is independent of $\boldsymbol{x}_t, \ldots \boldsymbol{x}_T$ given $\boldsymbol{z}_t$, we have

$$p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1}) = p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$$

- High-level Idea:
  1. Compute $p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{t-1})$ via Gaussian combining and Gaussian conditioning
     - This gives us $p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$
  2. Given $p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$, update $\hat{\boldsymbol{z}}_{t|T}, \hat{\boldsymbol{\Sigma}}_{t|T}$ into $\hat{\boldsymbol{z}}_{t-1|T}, \hat{\boldsymbol{\Sigma}}_{t-1|T}$ via LoTE and LoTC

Smoothing

$p(Z_t \mid \boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q})$

$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_t)$

- Step 1: Compute $p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \dots, \boldsymbol{x}_{t-1})$

Gaussian Combining: If $\boldsymbol{x} = \boldsymbol{C}\boldsymbol{z} + \boldsymbol{v}$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ and $\boldsymbol{v} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_v)$ then

$$\begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{x} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_z \\ \boldsymbol{C}\boldsymbol{\mu}_z \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_z \boldsymbol{C}^\top \\ \boldsymbol{C}\boldsymbol{\Sigma}_z & \boldsymbol{C}\boldsymbol{\Sigma}_z\boldsymbol{C}^\top + \boldsymbol{\Sigma}_v \end{bmatrix} \right)$$

Gaussian Conditioning: $\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{b} - \boldsymbol{\mu}_b),\ \boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$

- Step 1.1: Applying Gaussian combining to $\begin{bmatrix} \boldsymbol{z}_{t-1} \\ \boldsymbol{z}_t \end{bmatrix} \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_{t-1}$ gives:

$$\begin{bmatrix} \boldsymbol{z}_{t-1} \\ \boldsymbol{z}_t \end{bmatrix} \mid \boldsymbol{x}_0, \dots, \boldsymbol{x}_{t-1} \sim \mathcal{N}\left(\begin{bmatrix} \hat{\boldsymbol{z}}_{t-1|t-1} \\ \hat{\boldsymbol{z}}_{t|t-1} \end{bmatrix}, \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{t-1|t-1} & \hat{\boldsymbol{\Sigma}}_{t-1|t-1}\boldsymbol{A}^\top \\ \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{t-1|t-1} & \hat{\boldsymbol{\Sigma}}_{t|t-1} \end{bmatrix} \right)$$

$\hat{\boldsymbol{\Sigma}}_{t|t-1} = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{t-1|t-1}\boldsymbol{A}^\top + \boldsymbol{Q}$

- Step 1.2: From Gaussian conditioning we see $\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t, \boldsymbol{x}_0, \dots, \boldsymbol{x}_{t-1}$ has distribution:

$$\mathcal{N}\left(\hat{\boldsymbol{z}}_{t-1|t-1} + \boldsymbol{L}_{t-1}\left(\boldsymbol{z}_t - \hat{\boldsymbol{z}}_{t|t-1}\right), \hat{\boldsymbol{\Sigma}}_{t-1|t-1} - \boldsymbol{L}_{t-1}\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{t-1|t-1}\right)$$

$\boldsymbol{L}_{t-1} \coloneqq \hat{\boldsymbol{\Sigma}}_{t-1|t-1}\boldsymbol{A}^\top\hat{\boldsymbol{\Sigma}}_{t|t-1}^{-1}$

Remark. Note that $\boldsymbol{L}_{t-1}\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{t-1|t-1} = \boldsymbol{L}_{t-1}\hat{\boldsymbol{\Sigma}}_{t|t-1}\boldsymbol{L}_{t-1}^\top$, so the covariance $\hat{\boldsymbol{\Sigma}}_{t-1|t-1} - \boldsymbol{L}_{t-1}\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{t-1|t-1}$ is symmetric

Smoothing

$$L_{t-1} = \hat{\Sigma}_{t-1|t-1} A^\top \hat{\Sigma}_{t|t-1}^{-1}$$

$$\hat{z}_{s|t} := \mathbb{E}[z_s \mid x_0, \dots, x_t]$$
$$\hat{\Sigma}_{s|t} := \mathrm{Cov}(z_s \mid x_0, \dots, x_t)$$

- We have obtained

$$p_\theta(z_{t-1} \mid z_t, x_0, \dots, x_T) = \mathcal{N}\left(\hat{z}_{t-1|t-1} + L_{t-1}\left(z_t - \hat{z}_{t|t-1}\right), \hat{\Sigma}_{t-1|t-1} - L_{t-1}\hat{\Sigma}_{t|t-1}L_{t-1}^\top\right)$$

- Step 2: update $\hat{z}_{t|T}$, $\hat{\Sigma}_{t|T}$ into $\hat{z}_{t-1|T}$, $\hat{\Sigma}_{t-1|T}$ via LoTE and LoTC

$$\begin{aligned}
\hat{z}_{t-1|T} &= \mathbb{E}[z_{t-1} | x_0, \dots, x_T] = \mathbb{E}_{z_t}\left[\mathbb{E}_{z_{t-1}}[z_{t-1} \mid z_t, x_0, \dots, x_T]\right] \\
&= \mathbb{E}_{z_t}\left[\hat{z}_{t-1|t-1} + L_{t-1}\left(z_t - \hat{z}_{t|t-1}\right) \mid x_0, \dots, x_T\right] \\
&= \hat{z}_{t-1|t-1} + L_{t-1}\left(\hat{z}_{t|T} - \hat{z}_{t|t-1}\right)
\end{aligned}$$

$$\begin{aligned}
\hat{\Sigma}_{t-1|T} &= \mathrm{Cov}(z_{t-1} \mid x_0, \dots, x_T) = \mathbb{E}[\mathrm{Cov}(z_{t-1} \mid z_t, x_0, \dots, x_T)] + \mathrm{Cov}(\mathbb{E}[z_{t-1} \mid z_t, x_0, \dots, x_T]) \\
&= \mathbb{E}\left[\hat{\Sigma}_{t-1|t-1} - L_{t-1}\hat{\Sigma}_{t|t-1}L_{t-1}^\top\right] + \mathrm{Cov}\left(\hat{z}_{t-1|t-1} + L_{t-1}\left(z_t - \hat{z}_{t|t-1}\right)\right) \\
&= \hat{\Sigma}_{t-1|t-1} - L_{t-1}\hat{\Sigma}_{t|t-1}L_{t-1}^\top + \mathrm{Cov}\left(\hat{z}_{t-1|t-1} + L_{t-1}\left(z_t - \hat{z}_{t|t-1}\right) | x_0, \dots, x_T\right) \\
&= \hat{\Sigma}_{t-1|t-1} + L_{t-1}\left(\hat{\Sigma}_{t|T} - \hat{\Sigma}_{t|t-1}\right)L_{t-1}^\top
\end{aligned}$$

Summary: Smoothing for LDSs

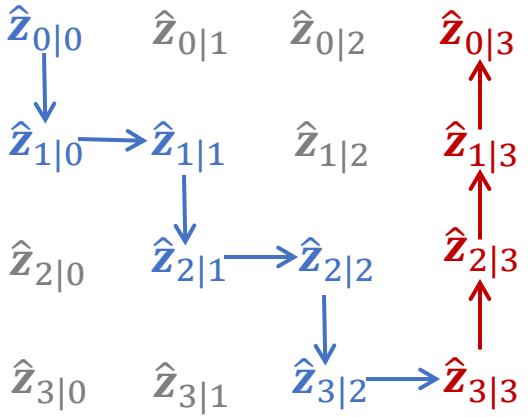
$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$
$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$

- Goal. Given θ and $(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$, compute
$$p_\theta(\boldsymbol{z}_t \mid \boldsymbol{x}_0, \ldots, \boldsymbol{x}_T)$$

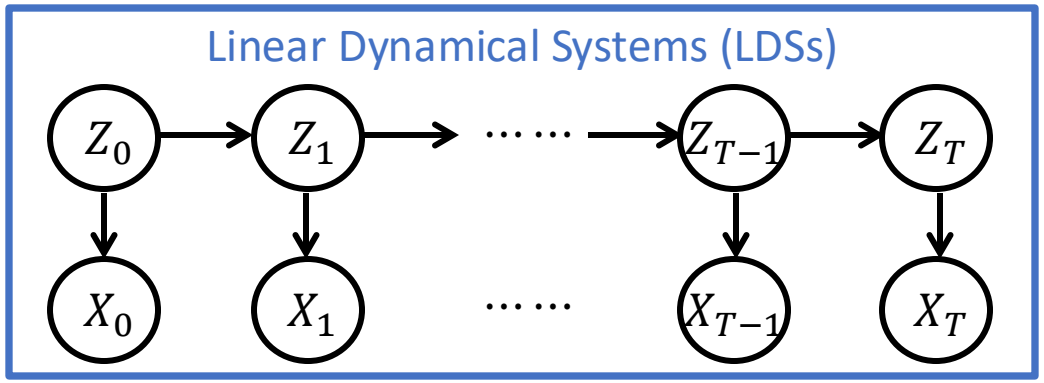
- Algorithm (known as "Rauch-Tung-Striebel smoother").
  1. (Forward Pass) Run Kalman filtering to compute $\hat{\boldsymbol{z}}_{t|t}$, $\hat{\boldsymbol{\Sigma}}_{t|t}$ and $\hat{\boldsymbol{z}}_{t+1|t}$, $\hat{\boldsymbol{\Sigma}}_{t+1|t}$ for all t
  2. (Backward Pass) For $t = T, \ldots, 1$, compute the following:
     - $\boldsymbol{L}_{t-1} = \hat{\boldsymbol{\Sigma}}_{t-1|t-1} \boldsymbol{A}^\top \hat{\boldsymbol{\Sigma}}_{t|t-1}^{-1}$
     - $\hat{\boldsymbol{z}}_{t-1|T} = \hat{\boldsymbol{z}}_{t-1|t-1} + \boldsymbol{L}_{t-1}\left(\hat{\boldsymbol{z}}_{t|T} - \hat{\boldsymbol{z}}_{t|t-1}\right)$
     - $\hat{\boldsymbol{\Sigma}}_{t-1|T} = \hat{\boldsymbol{\Sigma}}_{t-1|t-1} + \boldsymbol{L}_{t-1}(\hat{\boldsymbol{\Sigma}}_{t|T} - \hat{\boldsymbol{\Sigma}}_{t|t-1})\boldsymbol{L}_{t-1}^\top$

Remark: $\boldsymbol{L}_{t-1}$ might also be computed in the forward pass

Computational Trajectory of Smoothing:



Learning



Model Parameters:
- $\theta := (\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{C}, \boldsymbol{R})$

- Learning. Given N observations $\left\{\boldsymbol{x}_{0:T}^{(n)}\right\}_{n=1}^{N}$, find best θ:

$$\max_{\theta} \prod_{n=1}^{N} p_\theta(\boldsymbol{x}_{0:T}^{(n)})$$

- We can perform learning via the EM algorithm
  - derivations are algebraically involved and are given at the end of the slides.

Learning (Review Lecture 2)

- The likelihood $\prod_{i=1}^{N} p_\theta(\boldsymbol{x}_i) = \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right)}{(2\boldsymbol{\pi})^{\frac{ND}{2}} \det(\boldsymbol{\Sigma})^{\frac{N}{2}}}$ is maximized at

$$\boldsymbol{\mu}^* = \frac{\sum_{i=1}^{N}\boldsymbol{x}_i}{N}, \qquad \boldsymbol{\Sigma}^* = \frac{\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu}^*)(\boldsymbol{x}_i - \boldsymbol{\mu}^*)^\top}{N}$$

empirical mean

empirical covariance

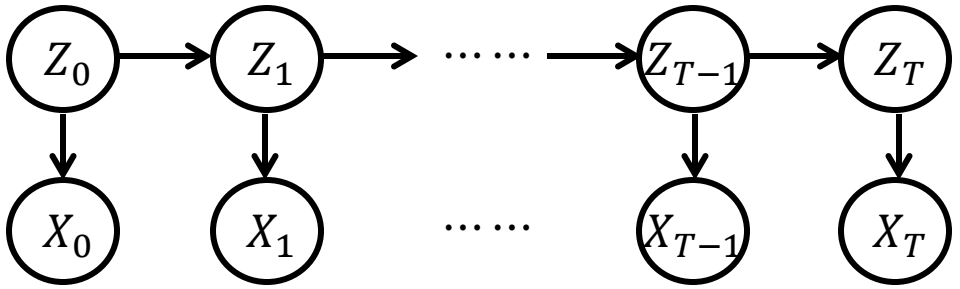
- How did we obtain $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$?
  - Step 0: Rewrite the objective
    - maximizing log-likelihood is minimizing $N \log\det\boldsymbol{\Sigma} + \sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$
  - Step 1: Set the derivative w.r.t. $\boldsymbol{\mu}$ to 0
    - solving it gives $\boldsymbol{\mu}^*$
  - Step 2: Substitute $\boldsymbol{\mu} = \boldsymbol{\mu}^*$ into the objective, and set the derivative w.r.t. $\boldsymbol{\Sigma}^{-1}$ to 0

$m \cdot \log\det\boldsymbol{\Sigma} + \mathrm{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})$ *is minimized at* $\boldsymbol{\Sigma} = \boldsymbol{S}/m$

Learning

Model Parameters:
- $\theta := (\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{C}, \boldsymbol{R})$

Linear Dynamical Systems (LDSs)



- Learning. Given N observations $\left\{\boldsymbol{x}_{0:T}^{(n)}\right\}_{n=1}^{N}$, find best θ:

$$\max_{\theta} \prod_{n=1}^{N} p_\theta(\boldsymbol{x}_{0:T}^{(n)})$$

- We are going to apply the EM algorithm (iteration: k):

E-step:

$$q^k\left(\boldsymbol{z}_{0:T}|\boldsymbol{x}_{0:T}^{(n)}\right) = p_{\theta^k}\left(\boldsymbol{z}_{0:T}|\boldsymbol{x}_{0:T}^{(n)}\right)$$

M-step:

$$\theta^{k+1} = \text{argmax}_\theta \sum_{n=1}^{N} \mathbb{E}_{z\sim q^k(z|x^{(n)})}\left[\log p_\theta\left(\boldsymbol{x}_{0:T}^{(n)}, \boldsymbol{z}_{0:T}\right)\right]$$

Guessing

Model Parameters:
- $\theta \coloneqq (\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{C}, \boldsymbol{R})$

E-step:

$$q^k\left(\boldsymbol{z}_{0:T}|\boldsymbol{x}_{0:T}^{(n)}\right) = p_{\theta^k}\left(\boldsymbol{z}_{0:T}\,|\boldsymbol{x}_{0:T}^{(n)}\right)$$

M-step:

$$\theta^{k+1} = \text{argmax}_\theta \sum_{n=1}^{N} \mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}\left[\log p_\theta\left(\boldsymbol{x}_{0:T}^{(n)}, \boldsymbol{z}_{0:T}\right)\right]$$

$$p(Z_0) = \mathcal{N}(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0) \qquad p(Z_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q}) \qquad p(Y_t|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{C}\boldsymbol{z}_t, \boldsymbol{R})$$

- Let guess a solution to the M-step:
  - Since $\boldsymbol{\pi}_0 = \mathbb{E}[\boldsymbol{z}_0]$, we guess… :

$$\boldsymbol{\pi}_0^{k+1} = \sum_{n=1}^{N} \frac{\mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}[\boldsymbol{z}_0]}{N}$$

← "empirical mean"

  - And we guess the covariance should be

$$\boldsymbol{\Sigma}_0^{k+1} = \sum_{n=1}^{N} \frac{\left(\mathbb{E}_{q^k\left(z|x^{(n)}\right)}[\boldsymbol{z}_0] - \boldsymbol{\pi}_0^{k+1}\right)\left(\mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}[\boldsymbol{z}_0] - \boldsymbol{\pi}_0^{k+1}\right)^\top}{N}$$

"empirical covariance" ↖

    - Try having a guess for $\boldsymbol{A}^{k+1}, \boldsymbol{Q}^{k+1}, \boldsymbol{C}^{k+1}, \boldsymbol{R}^{k+1}$ yourself…

E-step (iteration: k)

$$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s|\boldsymbol{x}_0, \ldots, \boldsymbol{x}_t]$$
$$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s|\boldsymbol{x}_0, \ldots, \boldsymbol{x}_t)$$

E-step:

$$q^k\left(\boldsymbol{z}_{0:T}|\boldsymbol{x}_{0:T}^{(n)}\right) = p_{\theta^k}\left(\boldsymbol{z}_{0:T}|\boldsymbol{x}_{0:T}^{(n)}\right)$$

M-step:

$$\theta^{k+1} = \mathrm{argmax}_\theta \sum_{n=1}^{N} \mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}\left[\log p_\theta\left(\boldsymbol{x}_{0:T}^{(n)}, \boldsymbol{z}_{0:T}\right)\right]$$

- In E-step, we will need to compute the expectation $\mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}[\cdot]$.

- *"It turns out that"* …… we only need to compute the following expectations:

$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t] \coloneqq \mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}[\boldsymbol{z}_t]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top] \coloneqq \mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top] \coloneqq \mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]$$

  - later you will see why…

E-step (iteration: k)

$$\hat{\boldsymbol{z}}_{s|t} \coloneqq \mathbb{E}[\boldsymbol{z}_s|\boldsymbol{x}_0, \dots, \boldsymbol{x}_t]$$
$$\hat{\boldsymbol{\Sigma}}_{s|t} \coloneqq \mathrm{Cov}(\boldsymbol{z}_s|\boldsymbol{x}_0, \dots, \boldsymbol{x}_t)$$

E-step:

$$q^k\left(\boldsymbol{z}_{0:T}|\boldsymbol{x}_{0:T}^{(n)}\right) = p_{\theta^k}\left(\boldsymbol{z}_{0:T}|\boldsymbol{x}_{0:T}^{(n)}\right)$$

M-step:

$$\theta^{k+1} = \mathrm{argmax}_\theta \sum_{n=1}^{N} \mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}\left[\log p_\theta\left(\boldsymbol{x}_{0:T}^{(n)}, \boldsymbol{z}_{0:T}\right)\right]$$

$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t] \coloneqq \mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}[\boldsymbol{z}_t]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top] \coloneqq \mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top] \coloneqq \mathbb{E}_{z\sim q^k\left(z|x^{(n)}\right)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]$$

- These expectations can all be computed via smoothing using θ^k and $\boldsymbol{x}_{0:T}^{(n)}$
  - To see this, dropping indices k, n for clarity, we have

$$\mathbb{E}[\boldsymbol{z}_t|\boldsymbol{x}_{0:T}] = \mathbb{E}[\boldsymbol{z}_t|\boldsymbol{x}_0, \dots, \boldsymbol{x}_T] = \hat{\boldsymbol{z}}_{t|T}$$
$$\mathbb{E}[\boldsymbol{z}_t\boldsymbol{z}_t^\top|\boldsymbol{x}_{0:T}] = \mathrm{Cov}(\boldsymbol{z}_t|\boldsymbol{x}_0, \dots, \boldsymbol{x}_T) + \mathbb{E}[\boldsymbol{z}_t|\boldsymbol{x}_0, \dots, \boldsymbol{x}_T]\mathbb{E}[\boldsymbol{z}_t^\top|\boldsymbol{x}_0, \dots, \boldsymbol{x}_T]$$
$$= \hat{\boldsymbol{\Sigma}}_{t|T} + \hat{\boldsymbol{z}}_{t|T}\hat{\boldsymbol{z}}_{t|T}^\top$$
$$\mathbb{E}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top|\boldsymbol{x}_{0:T}] = \hat{\boldsymbol{\Sigma}}_{t|T}\boldsymbol{L}_{t-1}^\top + \hat{\boldsymbol{z}}_{t|T}\hat{\boldsymbol{z}}_{t-1|T}^\top$$

$$\boldsymbol{L}_{t-1} = \hat{\boldsymbol{\Sigma}}_{t-1|t-1}\boldsymbol{A}^\top\hat{\boldsymbol{\Sigma}}_{t|t-1}^{-1}$$

M-step (iteration: k)

$$p(Z_0) = \mathcal{N}(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0)$$
$$p(Z_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q})$$
$$p(X_t|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{C}\boldsymbol{z}_t, \boldsymbol{R})$$

Model Parameters:
- $\theta \coloneqq (\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{C}, \boldsymbol{R})$

$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]$$

$$\theta^{k+1} = \text{argmax}_\theta \sum_{n=1}^N \mathbb{E}_{z\sim q^k(z|x^{(n)})}\left[\log p_\theta\left(\boldsymbol{x}_{0:T}^{(n)}, \boldsymbol{z}_{0:T}\right)\right]$$

- Observation. In the joint log-likelihood

$$p_\theta(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T, \boldsymbol{z}_0, \ldots, \boldsymbol{z}_T) = p_\theta(\boldsymbol{z}_0)\prod_{t=0}^T p_\theta(\boldsymbol{x}_t|\boldsymbol{z}_t)\prod_{t=1}^T p_\theta(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}),$$

  - $\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0$ only appear in $p_\theta(\boldsymbol{z}_0)$
  - $\boldsymbol{A}, \boldsymbol{Q}$ only appear in $\prod_{t=1}^T p_\theta(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$
  - $\boldsymbol{C}, \boldsymbol{R}$ only appear in $\prod_{t=0}^T p_\theta(\boldsymbol{x}_t|\boldsymbol{z}_t)$

  - So the objective of the M-step is separable (as in HMMs), this gives … (next page)

M-step (iteration: k)

$$p(Z_0) = \mathcal{N}(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0)$$
$$p(Z_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q})$$
$$p(X_t|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{C}\boldsymbol{z}_t, \boldsymbol{R})$$

Model Parameters:
- $\theta \coloneqq (\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{C}, \boldsymbol{R})$

$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]$$

$$\theta^{k+1} = \text{argmax}_\theta \sum_{n=1}^N \mathbb{E}_{z\sim q^k(z|x^{(n)})}\left[\log p_\theta\left(\boldsymbol{x}_{0:T}^{(n)}, \boldsymbol{z}_{0:T}\right)\right]$$

- We can therefore decompose M-step into 3 optimization problems (as in HMMs):

M-step $(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0)$:

$$(\boldsymbol{\pi}_0^{k+1}, \boldsymbol{\Sigma}_0^{k+1}) = \text{argmax}_\theta \sum_{n=1}^N \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\log p_\theta(\boldsymbol{z}_0)]$$

M-step $(\boldsymbol{C}, \boldsymbol{R})$:

$$(\boldsymbol{C}^{k+1}, \boldsymbol{R}^{k+1}) = \text{argmax}_\theta \sum_{n=1}^N \mathbb{E}_{z\sim q^k(z|x^{(n)})}\left[\sum_{t=0}^T \log p_\theta\left(\boldsymbol{x}_t^{(n)}\middle|\boldsymbol{z}_t\right)\right]$$

M-step $(\boldsymbol{A}, \boldsymbol{Q})$:

$$(\boldsymbol{A}^{k+1}, \boldsymbol{Q}^{k+1}) = \text{argmax}_\theta \sum_{n=1}^N \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\sum_{t=1}^T \log p_\theta(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})]$$

- We will address them one by one next

M-step $(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0)$

$$p(Z_0) = \mathcal{N}(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0)$$

$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]$$

$$(\boldsymbol{\pi}_0^{k+1}, \boldsymbol{\Sigma}_0^{k+1}) = \text{argmax}_\theta \sum_{n=1}^N \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\log p_\theta(\boldsymbol{z}_0)]$$

- Since $p_\theta(\boldsymbol{z}_0)$ is Gaussian, we have:

$$\log p_\theta(\boldsymbol{z}_0) \propto -\log\det\boldsymbol{\Sigma}_0 - (\boldsymbol{z}_0 - \boldsymbol{\pi}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{z}_0 - \boldsymbol{\pi}_0)$$

- And now we get this:

$$(\boldsymbol{\pi}_0^{k+1}, \boldsymbol{\Sigma}_0^{k+1}) = \text{argmin}_\theta\, N\log\det\boldsymbol{\Sigma}_0 + \sum_{n=1}^N \mathbb{E}_{z\sim q^k(z|x^{(n)})}[(\boldsymbol{z}_0 - \boldsymbol{\pi}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{z}_0 - \boldsymbol{\pi}_0)]$$

M-step $(\boldsymbol{\pi}_0, \boldsymbol{\Sigma}_0)$

$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]$$

$$(\boldsymbol{\pi}_0^{k+1}, \boldsymbol{\Sigma}_0^{k+1}) = \operatorname{argmin}_\theta N \log\det\boldsymbol{\Sigma}_0 + \sum_{n=1}^N \mathbb{E}_{z\sim q^k(z|x^{(n)})}[(\boldsymbol{z}_0-\boldsymbol{\pi}_0)^\top\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{z}_0-\boldsymbol{\pi}_0)]$$

use the definitions of $\mathbb{E}_k^{(n)}[\boldsymbol{z}_0]$ and $\mathbb{E}_k^{(n)}[\boldsymbol{z}_0\boldsymbol{z}_0^\top]$

$$\left(\boldsymbol{\pi}_0^{k+1}, \boldsymbol{\Sigma}_0^{k+1}\right) = \operatorname{argmin}_\theta N \log\det\boldsymbol{\Sigma}_0 + N\boldsymbol{\pi}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\pi}_0 + \sum_{n=1}^N \operatorname{tr}\left(\mathbb{E}_k^{(n)}[\boldsymbol{z}_0\boldsymbol{z}_0^\top]\boldsymbol{\Sigma}_0^{-1}\right) - 2\sum_{n=1}^N \boldsymbol{\pi}_0^\top\boldsymbol{\Sigma}_0^{-1}\mathbb{E}_k^{(n)}[\boldsymbol{z}_0]$$

- Setting the derivative with respect to $\boldsymbol{\pi}_0$ to 0 yields $\boldsymbol{\pi}_0^{k+1} = \dfrac{\sum_{n=1}^N \mathbb{E}_k^{(n)}[\boldsymbol{z}_0]}{N}$

Use $\boldsymbol{\pi}_0^{k+1}$

$$\boldsymbol{\Sigma}_0^{k+1} = \operatorname{argmin}_\theta N\cdot\log\det\boldsymbol{\Sigma}_0 + \sum_{n=1}^N \operatorname{tr}\left(\mathbb{E}_k^{(n)}[\boldsymbol{z}_0\boldsymbol{z}_0^\top]\boldsymbol{\Sigma}_0^{-1}\right) - N\cdot\operatorname{tr}\left(\boldsymbol{\pi}_0^{k+1}\left(\boldsymbol{\pi}_0^{k+1}\right)^\top\boldsymbol{\Sigma}_0^{-1}\right)$$

$m\cdot\log\det\boldsymbol{\Sigma} + \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})$ is minimized at $\boldsymbol{\Sigma} = \boldsymbol{S}/m$

$$\boldsymbol{\Sigma}_0^{k+1} = \frac{\sum_{n=1}^N \mathbb{E}_k^{(n)}[\boldsymbol{z}_0\boldsymbol{z}_0^\top]}{N} - \boldsymbol{\pi}_0^{k+1}(\boldsymbol{\pi}_0^{k+1})^\top$$

M-step $(\boldsymbol{C}, \boldsymbol{R})$

$$\mathbb{P}(X_t|\boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{C}\boldsymbol{z}_t, \boldsymbol{R})$$

$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]$$

$$(\boldsymbol{C}^{k+1}, \boldsymbol{R}^{k+1}) = \operatorname{argmax}_\theta \sum_{n=1}^N \mathbb{E}_{z\sim q^k(z|x^{(n)})}\left[\sum_{t=0}^T \log p_\theta\left(\boldsymbol{x}_t^{(n)}\middle|\boldsymbol{z}_t\right)\right]$$

$p_\theta\left(\boldsymbol{x}_t^{(n)}\middle|\boldsymbol{z}_t\right)$ is Gaussian

$$(\boldsymbol{C}^{k+1}, \boldsymbol{R}^{k+1}) = \operatorname{argmin}_\theta N(T+1)\log\det\boldsymbol{R} + \sum_{n=1}^N\sum_{t=0}^T \mathbb{E}_{\boldsymbol{z}_t\sim q^k(\boldsymbol{z}_t|x^{(n)})}\left[\left(\boldsymbol{x}_t^{(n)} - \boldsymbol{C}\boldsymbol{z}_t\right)^\top \boldsymbol{R}^{-1}\left(\boldsymbol{x}_t^{(n)} - \boldsymbol{C}\boldsymbol{z}_t\right)\right]$$

use the definitions of $\mathbb{E}_k^{(n)}[\boldsymbol{z}_t]$ and $\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]$

$$\min_\theta N(T+1)\log\det\boldsymbol{R} + \sum_{n=1}^N\sum_{t=0}^T\left\{\operatorname{tr}\left(\left(\boldsymbol{C}\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]\boldsymbol{C}^\top + \boldsymbol{x}_t^{(n)}\left(\boldsymbol{x}_t^{(n)}\right)^\top\right)\boldsymbol{R}^{-1} - 2\boldsymbol{x}_t^{(n)}\mathbb{E}_k^{(n)}[\boldsymbol{z}_t]^\top\boldsymbol{C}^\top\boldsymbol{R}^{-1}\right)\right\}$$

- Setting the derivative with respect to $\boldsymbol{C}$ to 0 yields

$$\boldsymbol{C}^{k+1} = \left(\sum_{n=1}^N\sum_{t=0}^T \boldsymbol{x}_t^{(n)}\,\mathbb{E}_k^{(n)}[\boldsymbol{z}_t]^\top\right)\left(\sum_{n=1}^N\sum_{t=0}^T \mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]\right)^{-1}$$

Use $\boldsymbol{C}^{k+1}$

$$\min_\theta N(T+1)\log\det\boldsymbol{R} + \sum_{n=1}^N\sum_{t=0}^T\left\{\operatorname{tr}\left(\boldsymbol{x}_t^{(n)}\left(\boldsymbol{x}_t^{(n)}\right)^\top\boldsymbol{R}^{-1} - \boldsymbol{x}_t^{(n)}\mathbb{E}_k^{(n)}[\boldsymbol{z}_t]^\top(\boldsymbol{C}^{k+1})^\top\boldsymbol{R}^{-1}\right)\right\}$$

$$\frac{\partial\operatorname{tr}(\boldsymbol{C}^\top\boldsymbol{X})}{\partial\boldsymbol{C}} = \boldsymbol{X}$$
$$\frac{\partial\operatorname{tr}(\boldsymbol{C}^\top\boldsymbol{X}\boldsymbol{C}\boldsymbol{Y})}{\partial\boldsymbol{C}} = \boldsymbol{X}\boldsymbol{C}\boldsymbol{Y} + \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{Y}^\top$$

$m\cdot\log\det\boldsymbol{\Sigma} + \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})$ *is minimized at* $\boldsymbol{\Sigma} = \boldsymbol{S}/m$

$$\boldsymbol{R}^{k+1} = \frac{\sum_{n=1}^N\sum_{t=0}^T \boldsymbol{x}_t^{(n)}\left(\boldsymbol{x}_t^{(n)}\right)^\top - \boldsymbol{C}^{k+1}\sum_{n=1}^N\sum_{t=0}^T \mathbb{E}_k^{(n)}[\boldsymbol{z}_t]\left(\boldsymbol{x}_t^{(n)}\right)^\top}{N(T+1)}$$

M-step $(\boldsymbol{A}, \boldsymbol{Q})$

$$p(Z_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{z}_{t-1}, \boldsymbol{Q})$$

$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]$$
$$\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top] \coloneqq \mathbb{E}_{z\sim q^k(z|x^{(n)})}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]$$

$$(\boldsymbol{A}^{k+1}, \boldsymbol{Q}^{k+1}) = \operatorname{argmax}_\theta \sum_{n=1}^N \mathbb{E}_{z\sim q^k(z|x^{(n)})}\left[\sum_{t=1}^T \log p_\theta(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})\right]$$

$p_\theta(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ is Gaussian

$$(\boldsymbol{A}^{k+1}, \boldsymbol{Q}^{k+1}) = \operatorname{argmin}_\theta NT \log\det\boldsymbol{Q} + \sum_{n=1}^N \sum_{t=1}^T \mathbb{E}_{z\sim q^k(z|x^{(n)})}[(\boldsymbol{z}_t - \boldsymbol{A}\boldsymbol{z}_{t-1})^\top \boldsymbol{Q}^{-1}(\boldsymbol{z}_t - \boldsymbol{A}\boldsymbol{z}_{t-1})]$$

use the definitions of $\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]$ and $\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]$

$$\min_\theta NT \log\det\boldsymbol{Q} + \sum_{n=1}^N \sum_{t=1}^T \left\{\operatorname{tr}\left(\left(\boldsymbol{A}\left(\mathbb{E}_k^{(n)}[\boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}^\top]\right)\boldsymbol{A}^\top + \mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]\right)\boldsymbol{Q}^{-1} - 2\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]\boldsymbol{A}^\top\boldsymbol{Q}^{-1}\right)\right\}$$

- Setting the derivative with respect to $\boldsymbol{A}$ to 0 yields

$$\boldsymbol{A}^{k+1} = \left(\sum_{n=1}^N \sum_{t=1}^T \mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top]\right)\left(\sum_{n=1}^N \sum_{t=1}^T \mathbb{E}_k^{(n)}[\boldsymbol{z}_{t-1}\boldsymbol{z}_{t-1}^\top]\right)^{-1}$$

Use $\boldsymbol{A}^{k+1}$

$$\min_\theta NT \log\det\boldsymbol{Q} + \sum_{n=1}^N \sum_{t=1}^T \left\{\operatorname{tr}\left(\mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top]\boldsymbol{Q}^{-1} - \mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_{t-1}^\top](\boldsymbol{A}^{k+1})^\top\boldsymbol{Q}^{-1}\right)\right\}$$

$m \cdot \log\det\boldsymbol{\Sigma} + \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})$ *is minimized at* $\boldsymbol{\Sigma} = \boldsymbol{S}/m$

$$\boldsymbol{Q}^{k+1} = \frac{\sum_{n=1}^N \sum_{t=0}^T \mathbb{E}_k^{(n)}[\boldsymbol{z}_t\boldsymbol{z}_t^\top] - \boldsymbol{A}^{k+1}\sum_{n=1}^N \sum_{t=0}^T \mathbb{E}_k^{(n)}[\boldsymbol{z}_{t-1}\boldsymbol{z}_t^\top]}{NT}$$

$$\frac{\partial \operatorname{tr}(\boldsymbol{A}^\top\boldsymbol{X})}{\partial \boldsymbol{A}} = \boldsymbol{X}$$
$$\frac{\partial \operatorname{tr}(\boldsymbol{A}^\top\boldsymbol{X}\boldsymbol{A}\boldsymbol{Y})}{\partial \boldsymbol{A}} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{Y} + \boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{Y}^\top$$

Summary: EM for LDSs (iteration: k)

- $\hat{z}_{s|t} := \mathbb{E}[z_s | x_0, \dots, x_t]$
- $\hat{\Sigma}_{s|t} := \mathrm{Cov}(z_s | x_0, \dots, x_t)$
- $L_{t-1} := \hat{\Sigma}_{t-1|t-1} A^\top \hat{\Sigma}^{-1}_{t|t-1}$

E-step

Given θ^k, for each $x^{(n)}_{0:T}$, use Kalman Filter & Smoothing to compute:

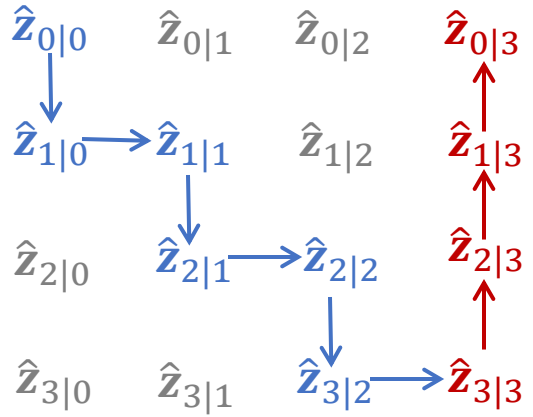
- $\mathbb{E}^{(n)}_k[z_t] := \mathbb{E}_{z \sim q^k(z|x^{(n)})}\left[z_t | x^{(n)}_{0:T}\right]$
- $\mathbb{E}^{(n)}_k[z_t z_t^\top] := \mathbb{E}_{z \sim q^k(z|x^{(n)})}\left[z_t z_t^\top | x^{(n)}_{0:T}\right]$
- $\mathbb{E}^{(n)}_k[z_t z_{t-1}^\top] := \mathbb{E}_{z \sim q^k(z|x^{(n)})}\left[z_t z_{t-1}^\top | x^{(n)}_{0:T}\right]$

add indices n, k

Kalman Filter & Smoothing can compute

- $\mathbb{E}[z_t | x_{0:T}] = \hat{z}_{t|T}$
- $\mathbb{E}[z_t z_t^\top | x_{0:T}] = \hat{\Sigma}_{t|T} + \hat{z}_{t|T}\hat{z}^\top_{t|T}$
- $\mathbb{E}[z_t z_{t-1}^\top | x_{0:T}] = L_{t-1}\hat{\Sigma}_{t|T} + \hat{z}_{t|T}\hat{z}^\top_{t-1|T}$

via forward & backward passes



M-step

Update parameters:

- $\pi_0^{k+1} = \frac{\sum_{n=1}^N \mathbb{E}^{(n)}_k[z_0]}{N}$
- $\Sigma_0^{k+1} = \frac{\sum_{n=1}^N \mathbb{E}^{(n)}_k[z_0 z_0^\top]}{N} - \pi_0^{k+1}(\pi_0^{k+1})^\top$
- $C^{k+1} = \left(\sum_{n=1}^N \sum_{t=0}^T x_t^{(n)} \mathbb{E}^{(n)}_k[z_t]^\top\right)\left(\sum_{n=1}^N \sum_{t=0}^T \mathbb{E}^{(n)}_k[z_t z_t^\top]\right)^{-1}$
- $R^{k+1} = \frac{\sum_{n=1}^N \sum_{t=0}^T x_t^{(n)}\left(x_t^{(n)}\right)^\top - C^{k+1}\sum_{n=1}^N \sum_{t=0}^T \mathbb{E}^{(n)}_k[z_t]\left(x_t^{(n)}\right)^\top}{N(T+1)}$
- $A^{k+1} = \left(\sum_{n=1}^N \sum_{t=1}^T \mathbb{E}^{(n)}_k[z_t z_{t-1}^\top]\right)\left(\sum_{n=1}^N \sum_{t=1}^T \mathbb{E}^{(n)}_k[z_{t-1} z_{t-1}^\top]\right)^{-1}$
- $Q^{k+1} = \frac{\sum_{n=1}^N \sum_{t=0}^T \mathbb{E}^{(n)}_k[z_t z_t^\top] - A^{k+1}\sum_{n=1}^N \sum_{t=0}^T \mathbb{E}^{(n)}_k[z_{t-1} z_t^\top]}{NT}$

Model Parameters:

- $\theta := (\pi_0, \Sigma_0, A, Q, C, R)$

$$p(Z_0) = \mathcal{N}(\pi_0, \Sigma_0)$$
$$p(Z_t | z_{t-1}) = \mathcal{N}(Az_{t-1}, Q)$$
$$p(X_t | z_t) = \mathcal{N}(Cz_t, R)$$