TS : KALMAN FILTERING.

So far, we have considered i.i.d. observations.

+ : likelihood can easily be expressed as a product
– : unrealistic in many situations.

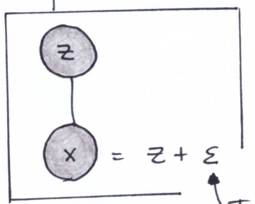
In this chapter, we depart from the i.i.d assumption, and consider an important model for sequential data.

Ex: (i) Rainfall measurements on successive days at a particular location.

(ii) Position of a moving object.

(iii) Daily values of the EUR-RUB exchange rate.

Usually, the quantity of interest (denoted z) is measured using a noisy sensor that returns an observation x representing z plus some noise ε



← What we want (hidden)

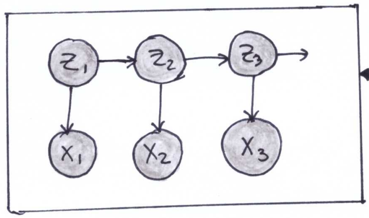
← What we observe

Typically, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

'Best' guess for z is x

Better guess = repeat the measurement n times and average the xs.

This is OK if z is not changing over time.



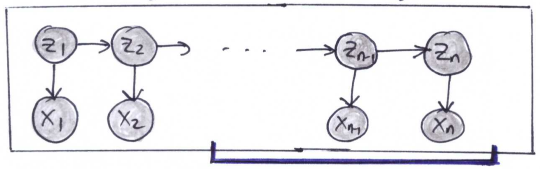
← How would you estimate $z_1, z_2, z_3, \dots$ now?

$\hat{z}_1 = x_1$, $\hat{z}_2 = x_2$, $\hat{z}_3 = x_3$?

↳ Maybe average the xs to reduce the noise.

⚠ By doing so, you introduce a bias.

Idea: Use a moving window of length L



Take the L most recent observations and average them.

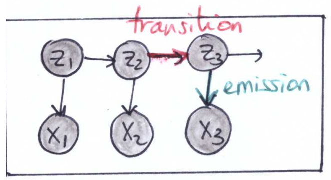
If z is slowly varying + σ^2 large ⇒ Take a long window

If z is quickly varying + σ^2 small ⇒ Take a short window

Situations in between?

In any case, we need a better way to decide on the value of L ⇒ we need a probabilistic model.

- **Model # 1**



- The zs are discrete & evolve according to a Markov Chain (MC).
- Link between z and x is probabilistic (arbitrary e.g. gaussian / Binomial / ...)

$P(z_2 = j \mid z_1 = i) = a_{ij}$ = transition probability

$P(x_2 = k \mid z_2 = j) = \pi_{jk}$ = emission probability

⇒ End up with Hidden Markov Models (HMM) (not studied here)

- **Model # 2**

- The zs are continuous & normally distributed.
Transition probability is also gaussian: $p(z_j \mid z_{j-1}) = \mathcal{N}(z_j \mid A z_{j-1}, \Gamma)$
- Emission probability is Gaussian: $p(x_j \mid z_j) = \mathcal{N}(x_j \mid C z_j, \Sigma)$

variable name   mean   cov.

⇒ End up with equations of Kalman Filtering.

+ : likelihood can still be easily written.

The transition + emission distributions are commonly expressed in an equivalent way in terms of linear equations:

$$\begin{cases} z_1 = \mu_0 + u \,, & u \sim \mathcal{N}(u \mid 0, P_0) \\ z_j = A z_{j-1} + w_j \,, & w \sim \mathcal{N}(w \mid 0, \Gamma) \\ x_j = C z_j + v_j \,, & v \sim \mathcal{N}(v \mid 0, \Sigma) \end{cases}$$

Q: (i) Estimation of model parameters $\theta = \{A, \Gamma, C, \Sigma, \mu_0, P_0\}$

↳ use MLE / EM algorithm

(ii) Predict z_n (and x_n) given $x_1, \dots, x_{n-1}$ (and x_n).

We turn our attention to question (ii), and postpone (i) for later.

⇒ We are interested in the posterior distribution

$$p(z_n \mid x_1, \dots, x_n) =: \hat{\alpha}(z_n)$$

To end up with an efficient algorithm, we should obtain $\hat{\alpha}(z_n)$ easily from $\hat{\alpha}(z_{n-1})$. These should also have the same functional form. Our Gaussian assumption will make everything work.

- $\hat{\alpha}(z_{n-1}) = p(z_{n-1} \mid x_1, \dots, x_{n-1})$
- $\hat{\alpha}(z_{n-1}) \underbrace{p(z_n \mid z_{n-1})}_{\text{transition probability}} = p(z_{n-1} \mid x_1, \dots, x_{n-1})\, p(z_n \mid z_{n-1})$

  $= p(z_{n-1}, z_n \mid x_1, \dots, x_{n-1})$

  (conditioned on z_{n-1}, z_n is independent of $x_1, \dots, x_{n-1}$)
- $\hat{\alpha}(z_{n-1})\, p(z_n \mid z_{n-1}) \underbrace{p(x_n \mid z_n)}_{\text{emission probability}} = p(z_{n-1}, z_n, x_n \mid x_1, \dots, x_{n-1})$

  (conditioned on z_n, x_n is independent of $z_{n-1}, x_1, \dots, x_{n-1}$)

- $\displaystyle\int \hat{\alpha}(z_{n-1})\, p(z_n \mid z_{n-1})\, p(x_n \mid z_n)\, dz_{n-1}$

  $\displaystyle = \int p(z_{n-1}, z_n, x_n \mid x_1, \dots, x_{n-1})\, dz_{n-1}$

  $= p(z_n, x_n \mid x_1, \dots, x_{n-1})$

  $= p(z_n \mid x_1, \dots, x_n)\, \boxed{p(x_n \mid x_1, \dots, x_{n-1})}$

  $= \hat{\alpha}(z_n)\, \boxed{c_n}$

We obtained the recursion equation:

$$(*) \quad \boxed{c_n \hat{\alpha}(z_n) = p(x_n \mid z_n) \int \hat{\alpha}(z_{n-1})\, p(z_n \mid z_{n-1})\, dz_{n-1}} \quad n \geq 2$$

Remarks

(i) The initial z_1 is gaussian. Since z_j is expressed as a linear combination of gaussian variables, it is also gaussian. Likewise, the distribution $\hat{\alpha}(z_n)$ can be seen to be gaussian. We write

$$\hat{\alpha}(z_n) = \mathcal{N}(z_n \mid \mu_n, V_n)$$

Use the recursion equation $(*)$ to express μ_n, V_n in terms of μ_{n-1}, V_{n-1} and the model parameters.

⇒ prediction of z_n using μ_n (mean).

C.I obtained from V_n.

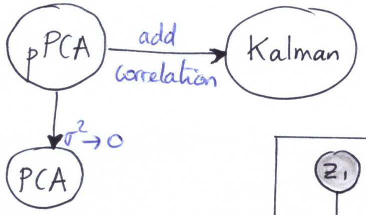
(ii) Kalman filtering ≡ extension of pPCA & Factor Analysis:

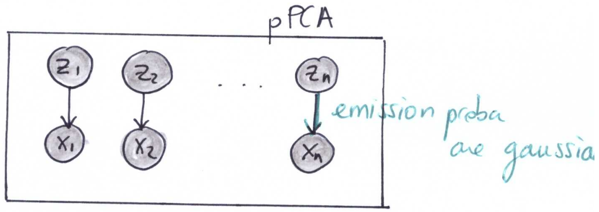
$\{x_n, z_n\}$ = linear - gaussian latent variable model

latent variable.

Recall: $x_n = W z_n + \mu + \varepsilon_n$ (page 17 Chp PCA)

Main difference: the z_n are no longer treated as independent variables: we allow sequential correlation in the data.





OK, time to do the maths & express recurrence equations for μ_n, V_n.

Starting point:

$$(*):\quad c_n\,\hat{\alpha}(z_n) = \underbrace{p(x_n|z_n)}_{\mathcal{N}(x_n|Cz_n,\,\Sigma)} \int \underbrace{\hat{\alpha}(z_{n-1})}_{\mathcal{N}(z_{n-1}|\mu_{n-1},\,V_{n-1})}\; \underbrace{p(z_n|z_{n-1})}_{\mathcal{N}(z_n|Az_{n-1},\,\Gamma)}\, dz_{n-1}\,.$$

→ Evaluation of such an integral is standard.

Toolbox: If
$$p(x) = \mathcal{N}(x|\mu, \Lambda^{-1})$$
$$p(y|x) = \mathcal{N}(y|Ax+b,\ L^{-1})$$

Then
$$p(y) = \mathcal{N}(y|A\mu+b,\ L^{-1} + A\Lambda^{-1}A^t)$$
$$p(x|y) = \mathcal{N}(x\,|\ S\{A^tL(y-b)+\Lambda\mu\},\ S)$$
where
$$S = (\Lambda + A^tLA)^{-1}$$
(left as an exercise)

where $p(y) = \int p(y|x)\,p(x)\,dx$

$$\Rightarrow \int \mathcal{N}(\underset{y}{z_n}\,|\,\underset{b=0}{Az_{n-1}},\ \underset{L^{-1}}{\Gamma})\ \mathcal{N}(\underset{x}{z_{n-1}}\,|\,\underset{\mu}{\mu_{n-1}},\ \underset{\Lambda^{-1}}{V_{n-1}})\ d\underset{x}{z_{n-1}}$$

$$= \mathcal{N}(z_n\,|\,A\mu_{n-1},\ \underbrace{\Gamma + AV_{n-1}A^t}_{\text{Call this } P_{n-1}})$$

$$\Rightarrow \underbrace{c_n}_{p(x_n|x_1,\dots,x_{n-1})}\ \underbrace{\hat{\alpha}(z_n)}_{\mathcal{N}(z_n|\mu_n,V_n) = p(z_n|x_1,\dots,x_n)} = \mathcal{N}(x_n|Cz_n,\Sigma)\,\mathcal{N}(z_n|A\mu_{n-1},P_{n-1})\,.$$

Variable z_n appears on both terms. We need to reshuffle things out to identify the RHS and LHS

$$\overbrace{\mathcal{N}(\underset{y}{x_n}|\underset{A}{C}\underset{x}{z_n},\underset{L^{-1}}{\Sigma})}^{p(y|x)}\ \overbrace{\mathcal{N}(\underset{x}{z_n}|\underset{\mu}{A\mu_{n-1}},\underset{\Lambda^{-1}}{P_{n-1}})}^{p(x)} \quad b=0$$

$$= p(y|x)\,p(x)$$
$$= p(x,y)$$
$$= p(x|y)\,p(y)$$

$$= \overbrace{\mathcal{N}(\underset{y}{x_n}|\underset{A\mu}{CA\mu_{n-1}},\ \underset{L^{-1}}{\Sigma} + \underset{A}{C}\underset{\Lambda^{-1}}{P_{n-1}}\underset{A^t}{C^t})}^{p(y)}$$

$$\times\ \overbrace{\mathcal{N}(\underset{x}{z_n}\,|\,S\{\underset{A^t}{C^t}\underset{L}{\Sigma^{-1}}(\underset{y}{x_n}-0) + \underset{\Lambda}{P_{n-1}^{-1}}\underset{\mu}{A\mu_{n-1}}\},\ S)}^{p(x|y)}$$

where
$$S = (P_{n-1}^{-1} + C^t\Sigma^{-1}C)^{-1}$$

$$= c_n\,\hat{\alpha}(z_n)$$

$$\Rightarrow \boxed{c_n = \mathcal{N}(x_n|CA\mu_{n-1},\ CP_{n-1}C^t + \Sigma)}$$

We simplify terms in the expression of $\mathcal{N}(z_n | \dots, \dots)$ by making use of two convenient matrix inverses identities: (7)

his one is asy to rove, ight?

$$(P^{-1} + B^t R^{-1} B)^{-1} B^t R^{-1} = P B^t (B P B^t + R)^{-1} \qquad (1)$$

&

$$(A + B D^{-1} C)^{-1} = A^{-1} - A^{-1} B (D + C A^{-1} B)^{-1} C A^{-1} \qquad (2)$$

Woodbury

(1a) • $$S = (\underbrace{P_{n-1}^{-1}}_{A} + \underbrace{C^t}_{B} \underbrace{\Sigma^{-1}}_{D^{-1}} \underbrace{C}_{C})^{-1} \overset{(2)}{=} P_{n-1} - \underbrace{P_{n-1} C^t (\Sigma + C P_{n-1} C^t)^{-1}}\, C P_{n-1}$$

(2a) • $$(\underbrace{P_{n-1}^{-1}}_{P^{-1}} + \underbrace{C^t}_{B^t} \underbrace{\Sigma^{-1}}_{R^{-1}} \underbrace{C}_{B})^{-1} \underbrace{C^t}_{B^t} \underbrace{\Sigma^{-1}}_{R^{-1}} \overset{(1)}{=} \overline{P_{n-1} C^t (C P_{n-1} C^t + \Sigma)^{-1}}$$

$$\Rightarrow S\{C^t \Sigma^{-1} x_n + P_{n-1}^{-1} A \mu_{n-1}\}$$

$$= \underbrace{P_{n-1} C^t (\Sigma + C P_{n-1} C^t)^{-1}}_{(2a)} x_n + \{\overbrace{P_{n-1} - P_{n-1} C^t (-)^{-1} C P_{n-1}}^{(1a)}\} \times P_{n-1}^{-1} A \mu_{n-1}$$

$$= A \mu_{n-1} + \underbrace{\boxed{P_{n-1} C^t (\Sigma + C P_{n-1} C^t)^{-1}}}_{\equiv\, K_n} (x_n - C A \mu_{n-1})$$

$$= A \mu_{n-1} + K_n (x_n - C A \mu_{n-1})$$

Our μ_n

Also, $S = P_{n-1} - K_n C P_{n-1}$

$= (I - K_n C) P_{n-1}$

Our V_n

SUMMARY

$\mu_n = A \mu_{n-1} + K_n (x_n - C A \mu_{n-1})$, $\left\{ \begin{array}{l} K_n = P_{n-1} C^t (\Sigma + C P_{n-1} C^t)^{-1} \\ P_{n-1} = \Gamma + A V_{n-1} A^t \end{array} \right.$

$V_n = (I - K_n C) P_{n-1}$

$c_n = \mathcal{N}(x_n \mid C A \mu_{n-1}, C P_{n-1} C^t + \Sigma)$

$n \geq 2$

(8)

$\Rightarrow$ Given μ_{n-1}, V_{n-1} and a new observation x_n, we can evaluate the distribution of z_n with mean μ_n and covariance matrix V_n.
(+ normalization coefficient c_n)

• $K_n = P_{n-1} C^t (\Sigma + C P_{n-1} C^t)^{-1}$ is known as the KALMAN GAIN.

Initial conditions: $\underbrace{c_1}_{p(x_1)} \underbrace{\hat{\alpha}(z_1)}_{p(z_1|x_1)} = \underbrace{p(z_1)}_{\mathcal{N}(z_1 | \mu_0, P_0)} \underbrace{p(x_1|z_1)}_{\mathcal{N}(x_1 | C z_1, \Sigma)}$

$\Rightarrow$ Proceed as before to obtain:
[Main difference: replace term $A\mu_{n-1}$ by μ_0 :]

INITIAL CONDITIONS

$\mu_1 = \mu_0 + K_1 (x_1 - C \mu_0)$

$V_1 = (I - K_1 C) P_0$

$c_1 = \mathcal{N}(x_1 \mid C \mu_0, C P_0 C^t + \Sigma)$

where

$K_1 = P_0 C^t (\Sigma + C P_0 C^t)^{-1}$

Interpretation of $\mu_n = A \mu_{n-1} + K_n (x_n - \underbrace{C A \mu_{n-1}})$

Mean of z_{n-1} updated using the transition matrix A

$z_n = A z_{n-1} + w_n$

Predicted observation obtained by applying the matrix C to the predicted hidden-state me

$\Rightarrow x_n - C A \mu_{n-1} = \text{error}$

$\Rightarrow$ posterior mean μ_n = predicted mean $A \mu_{n-1}$ + correction proportional to the error $(x_n - C A \mu_{n-1})$

We make this precise now

5 Alternative view of Kalman equations: (8a)

- In a filtering problem, we are interested in finding the 'best' predictor of z_n given observations $x_1, \dots, x_n$.
  ↳ denote it by $\hat{z}_n = \varphi^*(x_1, \dots, x_n)$ = function of the data

'Best' is commonly measured by the square loss between the prediction and the true value:

- $\hat{z}_n = \varphi^*(x_1, \dots, x_n) = \underset{\varphi}{\operatorname{argmin}} \; \mathbf{E}\left(z_n - \varphi(x_1, \dots, x_n)\right)^2$

  we know for a long time now that φ^* is given by the CE $\varphi^*(x_1, \dots, x_n) = E(z_n \mid x_1, \dots, x_n)$.

↘ We derived that $p(z_n \mid x_1, \dots, x_n) = \mathcal{N}(z_n \mid \mu_n, V_n)$, page 7

so that our best estimate of z_n is $\boxed{\hat{z}_n = \mathbf{E}(z_n \mid x_1, \dots, x_n) = \mu_n}$. BEST

↘ Moreover, the quality of this estimate is provided by the error covariance matrix,

$$E\left\{(z_n - \hat{z}_n)(z_n - \hat{z}_n)^t \mid x_1, \dots, x_n\right\}$$

$$= E\left\{\left(z_n - E(z_n \mid x_1, \dots, x_n)\right)\left(z_n - E(z_n \mid x_1, \dots, x_n)\right)^t \mid x_1, \dots, x_n\right\}$$

$$= V_n .$$

- We may now interpret equations derived at the bottom of page 7:

  ↘ Denote by $\hat{z}_{n|n}$ the best estimate at time n, (our μ_n) given observations up to and including time n.

  ↘ Denote by $P_{n|n}$ the error covariance matrix of $\hat{z}_{n|n}$ (Our V_n)

(8b)

- A time $(n-1)$, the best estimate of z_{n-1} given observations up to and including time $(n-1)$ is $\mu_{n-1} = \hat{z}_{n-1|n-1}$.

⇒ Predicted (a priori) latent estimate is $\hat{z}_{n|n-1} = A\hat{z}_{n-1|n-1}$

why 'a priori'? Because we are looking for an estimate of z_n based solely on observations $x_1, \dots, x_{n-1}$: x_n is <u>not</u> observed.

⇒ Also, the predicted (a priori) covariance estimate is the covariance matrix of $A\hat{z}_{n-1|n-1} + w_n$, which is $(A P_{n-1|n-1} A^t + \Gamma) =: P_{n|n-1}$ $\quad w_n \sim \mathcal{N}(0, \Gamma)$

<u>Remark</u>: Expression $\hat{z}_{n|n-1} = A\hat{z}_{n-1|n-1}$ makes sense intuitively: the best we can, given we have not observed x_n, is to make use of the equation $z_n = A z_{n-1} + w_n$ (our model), and to set the noise term to 0.

But how would you prove this formally?

Well, the 'best' estimate of z_n, given $x_1, \dots, x_{n-1}$ is given by the conditional mean $E(z_n \mid x_1, \dots, x_{n-1})$, and the predicted covariance matrix is given by

$$E\left\{\left(z_n - E(z_n \mid x_1, \dots, x_{n-1})\right)\left(z_n - E(z_n \mid x_1, \dots, x_{n-1})\right)^t \mid x_1, \dots, x_{n-1}\right\}$$

⇒ We need to derive $p(z_n \mid x_1, \dots, x_{n-1})$:

$$p(z_n \mid x_1, \dots, x_{n-1}) = \int p(z_n, z_{n-1}, x_1, \dots, x_{n-1})\, dz_{n-1}$$

$$= \int p(z_n \mid z_{n-1})\, p(z_{n-1} \mid x_1, \dots, x_n)\, dz_{n-1}$$

$$= \int \mathcal{N}(\underset{y}{z_n} \mid \underset{A}{A}\underset{x}{z_{n-1}}, \underset{L^{-1}}{\Gamma})\, \mathcal{N}(\underset{x}{z_{n-1}} \mid \underset{\mu}{\mu_{n-1}}, \underset{\Lambda^{-1}}{V_{n-1}})\, d\underset{x}{z_{n-1}}$$

$$\Rightarrow p(z_n | x_1,\dots,x_{n-1}) = \mathcal{N}(z_n | A\mu_{n-1}, \Gamma + A V_{n-1} A^t)$$

So that indeed • $\hat z_{n|n-1} = A \hat z_{n-1|n-1}$

• Cov $= P_{n|n-1} = \Gamma + A P_{n-1|n-1} A^t$

But then, what would be your 'best' estimate of z_{n+1}, given $x_1,\dots,x_{n-1}$?

$$\Rightarrow \text{Compute } p(z_{n+1}|x_1,\dots,x_{n-1}) = \iint p(z_{n+1}|z_n)\,p(z_n|z_{n-1})\,p(z_{n-1}|x_1,\dots,x_{n-1})\,dz_{n-1}\,dz_n$$

where

$$\int p(z_{n+1}|z_n)\,p(z_n|z_{n-1})\,dz_n = \int \mathcal{N}(\underset{y}{z_{n+1}}|\underset{A}{A}\underset{x}{z_n},\underset{L^{-1}}{\Gamma})\,\mathcal{N}(\underset{x}{z_n}|\underset{\mu}{A z_{n-1}},\underset{\Lambda^{-1}}{\Gamma})\,d\underset{x}{z_n}$$

$$= \mathcal{N}(z_{n+1}| A^2 z_{n-1}, \Gamma + A\Gamma A^t).$$

$$\Rightarrow p(z_{n+1}|x_1,\dots,x_{n-1}) = \int \mathcal{N}(\underset{y}{z_{n+1}}|\underset{A}{A^2}\underset{x}{z_{n-1}}, \underset{L^{-1}}{\Gamma + A\Gamma A^t})\,\mathcal{N}(\underset{x}{z_{n-1}}|\underset{\mu}{\mu_{n-1}}, \underset{\Lambda^{-1}}{V_{n-1}})\,d\underset{x}{z_{n-1}}$$

$$= \mathcal{N}(z_{n+1}| A^2 \mu_{n-1}, \Gamma + A\Gamma A^t + A V_{n-1} A^t)$$

Thus, • $\hat z_{n+1|n-1} = A^2 \hat z_{n-1|n-1}$ = propagate $\hat z_{n-1|n-1}$ linearly

• $P_{n+1|n-1} = \Gamma + \underbrace{\boxed{A\Gamma A^t}}_{\text{uncertainty increases!}} + A V_{n-1} A^t$

More generally, you can convince yourself that

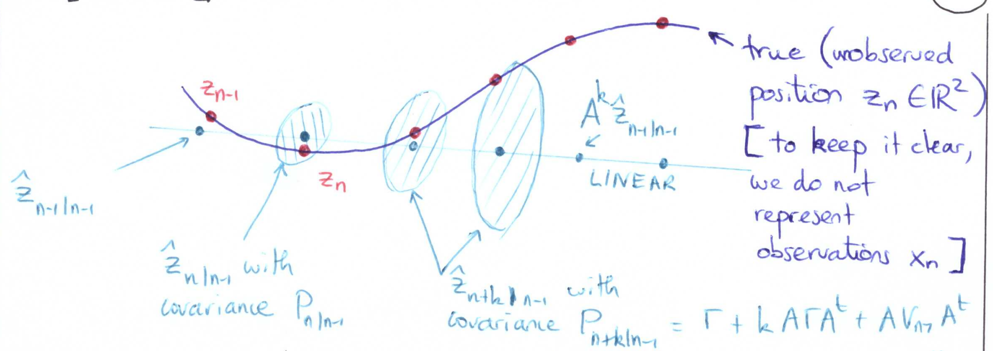
$$p(z_{n+k}|x_1,\dots,x_{n-1}) = \mathcal{N}(z_{n+k}| A^{k+1}\mu_{n-1}, \Gamma + kA\Gamma A^t + AV_{n-1}A^t)$$

$k \geq 0$

$\Rightarrow$ • $\hat z_{n+k|n-1} = A^{k+1}\mu_{n-1}$

• $P_{n+k|n-1} = \Gamma + \underbrace{\boxed{kA\Gamma A^t}}_{\text{increases with } k} + AV_{n-1}A^t$

[Picture]



$\Rightarrow$ uncertainty about z_{n+k} grows $\equiv$ the covariance matrix gets larger.

$\Rightarrow$ To reduce uncertainty, correct the trajectory with subsequent observations $x_n, x_{n+1}, \dots$

• Update prediction, given you observe x_n:

$\Rightarrow$ The residual error (aka INNOVATION) is

$$i_n = x_n - C\hat z_{n|n-1} = x_n - CA\hat z_{n-1|n-1}.$$

Use evolution equation, $x_n = Cz_n + v_n$.

$\Rightarrow$ The residual covariance is $C P_{n|n-1} C^t + \Sigma =: S_n$

The predicted a priori covariance matrix

$\Rightarrow$ The optimal Kalman gain is $K_n = P_{n|n-1} C^t S_n^{-1}$.

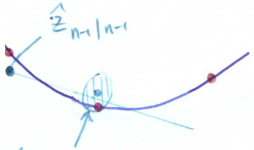
⇒ The updated (a posteriori) latent estimate:

$$\hat{z}_{n|n} = \underbrace{\hat{z}_{n|n-1}}_{\text{a priori estimate}} + \underbrace{K_n}_{\text{optimal 'weight'}} \underbrace{i_n}_{\text{error}}$$

$$= A \hat{z}_{n-1|n-1} \qquad = \text{correction to your linear prediction.}$$

Compare with the expression of μ_n page 7

⇒ The updated (a posteriori) covariance estimate is:



$$P_{n|n} = (I - K_n C) P_{n|n-1}.$$

(V_n)

We are indeed reducing uncertainty since $P_{n|n} < P_{n|n-1}$. look: $K_n C P_{n|n-1} = P_{n|n-1} C^t S_n^{-1} C P_{n|n-1}$ is positive definite.

- A closely related problem is that of SMOOTHING (the FILTERING problem aims at recovering at time n some information about z_n given $x_1, .., x_n$). In the smoothing problem, measurements derived later than time n can be used in obtaining information about z_n.

Ex: the way the human brain tackles the problem of reading written handwriting: words are read sequentially. When a word is difficult to read, words before and after it may be used to attempt to deduce the word.

Formally, we are interested in the 'best' estimate of z_j given $x_1, .., x_n$; $1 \leq j \leq n$.

↳ Under a square loss, the best estimate is given by the conditional mean of z_j given $x_1, .., x_n$; and it is therefore of interest to derive $p(z_j \mid x_1, .., x_n)$.

↳ Section II.1 p. 10

Remarks.

(i) Back to our discussion on the top of page 2:
Suppose that the measurement noise σ^2 is small compared to the rate at which the latent variable is changing; i.e. suppose $\Sigma \approx 0$. Take $C = I$, so that observation x corresponds to z + something small.

Bottom of page 7: $\mu_n = A\mu_{n-1} + K_n(x_n - \overset{C=I}{A}\mu_{n-1})$

$$K_n = P_{n-1} (\underset{0}{\Sigma} + P_{n-1})^{-1} = I$$

$$\Rightarrow \mu_n = x_n$$

⇒ Predict z_n using x_n, in agreement with our intuition.

(ii) What if the latent variable is evolving slowly relative to the observation noise level?

In the extreme case, we may assume

- $A = C = I$
- $\Gamma = 0$
- $P_0 \to \infty$

$\}$ z_j stays constant over time (initial distrib unimportant)

Then it is possible to show that the posterior mean for z_n is determined by the average of the $x_1, .., x_n$.

(iii) Application: tracking a moving object.

From •, the ⊗ are predicting •.

II. LEARNING THE MODEL PARAMETERS

In the previous section we derived the equations of Kalman filter, assuming the parameters are known. In this section, we derive a procedure for estimating $\theta = \{A, \Gamma, C, \Sigma, \mu_0, P_0\}$.

$\Rightarrow$ We do this using ML + EM algorithm (due to the presence of hidden variables).

- First, we need preliminary results:
  - (i) $\rightarrow$ Compute $p(z_j \mid x_1, \ldots, x_n)$, $1 \le j \le n$
  - (ii) $\rightarrow$ Compute $p(z_{j-1}, z_j \mid x_1, \ldots, x_n)$, $2 \le j \le n$

We address (i) + (ii) in the next section.

II.1. Preliminary results.

[To simplify notation, we write $\underline{X} = (x_1, \ldots, x_n)$; so that $p(z_j \mid x_1, \ldots, x_n)$ is rewritten $p(z_j \mid \underline{X})$.]

$\rightarrow$ We are interested in calculating (efficiently) the posterior probability $p(z_j \mid \underline{X}) =: \gamma(z_j)$, for $1 \le j \le n$.

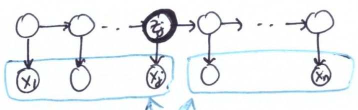
Note that for $j = n$, we have $\gamma(z_n) = \hat{\alpha}(z_n)$ (page 3)

Bayes $\Rightarrow$ $\gamma(z_j) = \dfrac{\boxed{p(\underline{X} \mid z_j)}\, p(z_j)}{p(\underline{X})}$, $j \le n-1$

Conditional independence property (left as an exercise)

$$= \frac{\boxed{p(x_1, \ldots, x_j \mid z_j)\, p(x_{j+1}, \ldots, x_n \mid z_j)}\, p(z_j)}{p(\underline{X})}$$

$$= \frac{p(x_1, \ldots, x_j, z_j)\, p(x_{j+1}, \ldots, x_n \mid z_j)}{p(\underline{X})}$$



independent, conditionally on z_j.

$$=: \frac{\alpha(z_j)\, \beta(z_j)}{p(\underline{X})}.$$

Where we have defined

$$\boxed{\begin{aligned} \alpha(z_j) &= p(x_1, \ldots, x_j, z_j) \\ \beta(z_j) &= p(x_{j+1}, \ldots, x_n \mid z_j) \end{aligned}}$$

Remark: we usually work with scaled versions of $\alpha(z_j)$ and $\beta(z_j)$, for numerical issues. Specifically,

- $\hat{\alpha}(z_j) = p(z_j \mid x_1, \ldots, x_j) = \dfrac{\alpha(z_j)}{p(x_1, \ldots, x_j)}$

  introduced on page 3 already.

  We introduce as well

  $c_j = p(x_j \mid x_1, \ldots, x_{j-1})$ [defined page 4]

  so that

  $p(x_1, \ldots, x_j) = p(x_j \mid x_1 \ldots x_{j-1})\, p(x_{j-1} \mid x_1 \ldots x_{j-2}) \ldots p(x_2 \mid x_1)\, p(x_1)$

  $= \prod_{m=1}^{j} c_m$

  So that $\boxed{\alpha(z_j) = \left(\prod_{m=1}^{j} c_m\right) \hat{\alpha}(z_j)}$

- Likewise, we defined the scaled variables $\hat{\beta}(z_j)$ as:

  $$\boxed{\beta(z_j) = \left(\prod_{m=j+1}^{n} c_m\right) \hat{\beta}(z_j)}$$

Now, $\left(\prod_{m=1}^{n} c_m\right) = p(\underline{X}) = \left(\prod_{m=1}^{j} c_m\right)\left(\prod_{m=j+1}^{n} c_m\right)$

$\Rightarrow \prod_{m=j+1}^{n} c_m = \dfrac{p(\underline{X})}{p(x_1, \ldots, x_j)} = p(x_{j+1}, \ldots, x_n \mid x_1, \ldots, x_j)$

so that $\hat{\beta}(z_j) = \dfrac{\beta(z_j)}{p(x_{j+1}, \ldots, x_n \mid x_1, \ldots, x_j)}$

Recursion equations for $\alpha / \beta / \hat{\alpha} / \hat{\beta}$. (12)

→ We already established on page 4 that

scaled version [$$c_j \hat{\alpha}(z_j) = p(x_j|z_j) \int \hat{\alpha}(z_{j-1}) p(z_j|z_{j-1}) \, dz_{j-1} \quad \text{—— (A)}$$

Equivalently,

$$c_j \frac{\alpha(z_j)}{\left(\prod_{m=1}^{j} c_m\right)} = p(x_j|z_j) \int \frac{\alpha(z_{j-1})}{\left(\prod_{m=1}^{j-1} c_m\right)} p(z_j|z_{j-1}) \, dz_{j-1}$$

unscaled version [$$\alpha(z_j) = \underbrace{p(x_j|z_j) \int \alpha(z_{j-1})}_{} p(z_j|z_{j-1}) \, dz_{j-1} \quad \text{—— (B)}$$

we have a FORWARD message passing: from z_{j-1} to z_j.

→ We establish a similar recursion equation for $\beta / \hat{\beta}$:

$$\beta(z_j) = p(x_{j+1}, \dots, x_n | z_j)$$
$$= \int p(x_{j+1}, \dots, x_n, z_{j+1} | z_j) \, dz_{j+1}$$
(cond. independence)
$$= \int p(x_{j+1}, \dots, x_n | z_{j+1}, \cancel{z_j}) \, p(z_{j+1}|z_j) \, dz_{j+1}$$
$$= \int p(x_{j+2}, \dots, x_n | z_{j+1}, \cancel{x_{j+1}}) \, p(x_{j+1}|z_{j+1}) \, p(z_{j+1}|z_j) \, dz_{j+1}$$
⇓

unscaled version [$$\beta(z_j) = \int \beta(z_{j+1}) \, p(x_{j+1}|z_{j+1}) \, p(z_{j+1}|z_j) \, dz_{j+1} \quad \text{—— (C)}$$

Equivalently,

$$\left(\prod_{m=j+1}^{n} c_m\right) \hat{\beta}(z_j) = \int \left(\prod_{m=j+2}^{n} c_m\right) \hat{\beta}(z_{j+1}) p(x_{j+1}|z_{j+1}) p(z_{j+1}|z_j) \, dz_{j+1}$$

scaled version [$$c_{j+1} \hat{\beta}(z_j) = \int \hat{\beta}(z_{j+1}) p(x_{j+1}|z_{j+1}) p(z_{j+1}|z_j) \, dz_{j+1} \quad \text{—— (D)}$$

We have a BACKWARD message passing: from z_{j+1} to z_j.

For these reasons, • $\alpha / \hat{\alpha}$ are referred to as the FORWARD variables (13)

• $\beta / \hat{\beta}$ —"— the BACKWARD variables.

(You will meet these again when discussing HMM)

Recall from page 10,

$$p(z_j | x_1, \dots, x_n) = \gamma(z_j) = \frac{\alpha(z_j)\beta(z_j)}{p(x_1, \dots, x_n)}$$
$$= \frac{\alpha(z_j)\beta(z_j)}{\left(\prod_{m=1}^{n} c_m\right)}$$
$$= \frac{\alpha(z_j)}{\left(\prod_{m=1}^{j} c_m\right)} \frac{\beta(z_j)}{\left(\prod_{m=j+1}^{n} c_m\right)}$$
$$= \hat{\alpha}(z_j) \hat{\beta}(z_j).$$

$$\Rightarrow \boxed{\gamma(z_j) = \frac{\alpha(z_j)\beta(z_j)}{p(x_1, \dots, x_n)} = \hat{\alpha}(z_j)\hat{\beta}(z_j)}$$

This representation, together with the recurrence equations satisfied by $\hat{\alpha} / \hat{\beta}$, will allow us to derive the distribution of $z_j | x_1, \dots, x_n$.

Starting point: recurrence equation (D)

$$c_{j+1} \hat{\beta}(z_j) = \int \hat{\beta}(z_{j+1}) \, p(x_{j+1}|z_{j+1}) p(z_{j+1}|z_j) \, dz_{j+1} \quad \times \hat{\alpha}(z_j)$$

$$c_{j+1} \underbrace{\hat{\alpha}(z_j) \hat{\beta}(z_j)}_{\gamma(z_j) = \text{gaussian} = \mathcal{N}(z_j | \hat{\mu}_j, \hat{V}_j)} = \int \hat{\alpha}(z_j) \hat{\beta}(z_{j+1}) p(x_{j+1}|z_{j+1}) p(z_{j+1}|z_j) \, dz_{j+1}$$

We are now looking for recurrence relations on $\hat{\mu}_j$ / $\hat{V}_j$. (14)

$$c_{j+1}\,\mathcal{N}(z_j \mid \hat{\mu}_j, \hat{V}_j) = \int \hat{\beta}(z_{j+1})\,\hat{\alpha}(z_j)\,p(x_{j+1} \mid z_{j+1})\,p(z_{j+1} \mid z_j)\,dz_{j+1}$$

where

$$\hat{\alpha}(z_j)\,p(z_{j+1} \mid z_j) = \mathcal{N}(\underset{x}{\underset{\shortparallel}{z_j}} \mid \underset{\mu}{\underset{\shortparallel}{\mu_j}}, \underset{\Lambda^{-1}}{\underset{\shortparallel}{V_j}})\,\mathcal{N}(\underset{y}{\underset{\shortparallel}{z_{j+1}}} \mid A\underset{x}{\underset{\shortparallel}{z_j}}, \underset{L^{-1}}{\underset{\shortparallel}{\Gamma}})$$

(notation from bottom of page 5)

$$= p(x)\,p(y \mid x)$$

$$= p(x \mid y)\,p(y)$$

$$= \mathcal{N}\Big(\underset{x}{\underset{\shortparallel}{z_j}} \,\Big|\, \underset{S}{\underset{\shortparallel}{M_j}}(A^t\underset{L}{\underset{\shortparallel}{\Gamma^{-1}}}\underset{y}{\underset{\shortparallel}{z_{j+1}}} + \underset{\Lambda}{\underset{\shortparallel}{V_j^{-1}}}\underset{\mu}{\underset{\shortparallel}{\mu_j}}), \underset{S}{\underset{\shortparallel}{M_j}}\Big)\,\mathcal{N}\big(\underset{y}{\underset{\shortparallel}{z_{j+1}}} \,\big|\, \underset{\mu}{\underset{\shortparallel}{A\mu_j}}, \underset{L^{-1}+A\Lambda^{-1}A^t}{\underset{\shortparallel}{P_j}}\big)$$

where

$$M_j = (V_j^{-1} + A^t\Gamma^{-1}A)^{-1}$$

Put $\boxed{m_j = M_j(A^t\Gamma^{-1}z_{j+1} + V_j^{-1}\mu_j)}$

$$\Rightarrow \quad \hat{\alpha}(z_j)\,p(z_{j+1} \mid z_j) = \mathcal{N}(z_j \mid m_j, M_j)\,\mathcal{N}(z_{j+1} \mid A\mu_j, P_j)$$

Now, $M_j = (V_j^{-1} + A^t\Gamma^{-1}A)^{-1}$

$= V_j - V_jA^t(\Gamma + AV_jA^t)^{-1}AV_j$ Relation (2) page 7

$= V_j - V_jA^tP_j^{-1}AV_j$ (◆)

$= (I - \underbrace{\boxed{V_jA^tP_j^{-1}}}_{=J_j}A)V_j$

$= (I - J_jA)V_j$

$$\Rightarrow \quad \boxed{\begin{aligned} M_j &= (I - J_jA)V_j \\ J_j &= V_jA^tP_j^{-1} \end{aligned}}$$

(15)

$$\Rightarrow \quad c_{j+1}\,\mathcal{N}(z_j \mid \hat{\mu}_j, \hat{V}_j) = \int \hat{\beta}(z_{j+1})\,\underbrace{p(x_{j+1} \mid z_{j+1})\,\mathcal{N}(z_{j+1} \mid A\mu_j, P_j)}_{=\,c_{j+1}\hat{\alpha}(z_{j+1})}\,\times\mathcal{N}(z_j \mid m_j, M_j)\,dz_{j+1}$$

(third line on page 6)

$$= c_{j+1}\int \gamma(z_{j+1})\,\mathcal{N}(z_j \mid m_j, M_j)\,dz_{j+1}$$

(depends on z_{j+1})

$$= c_{j+1}\int \mathcal{N}(\underset{x}{\underset{\shortparallel}{z_{j+1}}} \mid \underset{\mu}{\underset{\shortparallel}{\hat{\mu}_{j+1}}}, \underset{\Lambda^{-1}}{\underset{\shortparallel}{\hat{V}_{j+1}}})\,\mathcal{N}(\underset{y}{\underset{\shortparallel}{z_j}} \mid \underset{x}{\underset{\shortparallel}{m_j}}, \underset{L^{-1}}{\underset{\shortparallel}{M_j}})\,dz_{j+1}$$

$$m_j = \underset{A}{\underset{\shortparallel}{M_jA^t\Gamma^{-1}}}\,\underset{x}{\underset{\shortparallel}{z_{j+1}}} + \underset{b}{\underset{\shortparallel}{M_jV_j^{-1}\mu_j}}$$

(notation bottom of page 5)

$$= \int p(x)\,p(y \mid x)\,dx$$

$$= p(y)$$

$$= \mathcal{N}\Big(z_j \,\Big|\, \underbrace{M_jA^t\Gamma^{-1}}_{A}\underbrace{\hat{\mu}_{j+1}}_{\mu} + \overbrace{M_jV_j^{-1}\mu_j}^{b},\ \underbrace{M_j}_{L^{-1}} + \underbrace{M_jA^t\Gamma^{-1}}_{A}\underbrace{\hat{V}_{j+1}}_{\Lambda^{-1}}\underbrace{\Gamma^{-1}AM_j}_{A^t}\Big)$$

$$= c_{j+1}\,\mathcal{N}\Big(z_j \,\Big|\, M_j(A^t\Gamma^{-1}\hat{\mu}_{j+1} + V_j^{-1}\mu_j),\ M_j + M_jA^t\Gamma^{-1}\hat{V}_{j+1}\Gamma^{-1}AM_j\Big)$$

(M_j = symm.)

$$\Rightarrow \quad \left[\begin{aligned} \hat{\mu}_j &= M_j(A^t\Gamma^{-1}\hat{\mu}_{j+1} + V_j^{-1}\mu_j) \\ \hat{V}_j &= M_j + M_jA^t\Gamma^{-1}\hat{V}_{j+1}\Gamma^{-1}AM_j. \end{aligned}\right.$$

Let's simplify these expressions.

Note that $M_j A^t \Gamma^{-1} = \big(\underbrace{V_j - V_j A^t P_j^{-1} A V_j}_{= M_j \text{ (relation (♦) page 14)}}\big) A^t \Gamma^{-1}$ (16)

$$= V_j A^t \big(I - P_j^{-1} \underbrace{A V_j A^t}_{= P_j - \Gamma \text{ by definition of } P_j\text{, page 7}}\big) \Gamma^{-1}$$

$$= V_j A^t (P_j^{-1} \Gamma) \Gamma^{-1}$$

$$= V_j A^t P_j^{-1}$$

$$= J_j \quad \text{(defined bottom of page 14)}$$

$\Rightarrow$ • $\hat{\mu}_j = M_j (A^t \Gamma^{-1} \hat{\mu}_{j+1} + V_j^{-1} \mu_j)$

$$= J_j \hat{\mu}_{j+1} + M_j V_j^{-1} \mu_j$$

$$= J_j \hat{\mu}_{j+1} + \underbrace{(I - J_j A) V_j}_{\text{bottom of page 14}} V_j^{-1} \mu_j$$

$$\boxed{\hat{\mu}_j = \mu_j + J_j (\hat{\mu}_{j+1} - A \mu_j)}$$

• $\hat{V}_j = \underbrace{M_j} + \underbrace{M_j A^t \Gamma^{-1}} \hat{V}_{j+1} \underbrace{\Gamma^{-1} A M_j}$

$$= \underbrace{V_j - V_j A^t P_j^{-1} A V_j} + J_j \hat{V}_{j+1} J_j^t$$

$$= V_j + J_j \hat{V}_{j+1} J_j^t - J_j \underbrace{A V_j}$$

$J_j = V_j A^t P_j^{-1}$

$J_j P_j = V_j A^t$

$P_j^t J_j^t = A V_j^t$

$P_j J_j^t = A V_j$ } V_j, P_j symm

$$= V_j + J_j \hat{V}_{j+1} J_j^t - J_j P_j J_j^t$$

$$\boxed{\hat{V}_j = V_j + J_j (\hat{V}_{j+1} - P_j) J_j^t}$$

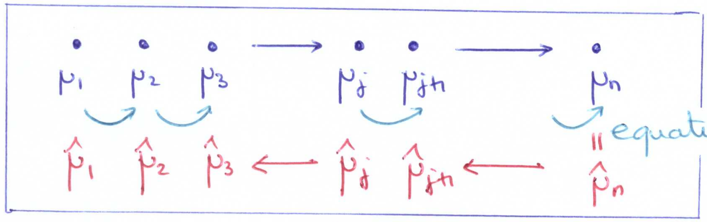
<u>Summary</u>: $\gamma(z_j) = p(x_j \mid x_1, \dots, x_n)$ (17)

$$= \mathcal{N}(z_j \mid \hat{\mu}_j, \hat{V}_j),$$

where • $\hat{\mu}_j = \mu_j + J_j (\hat{\mu}_{j+1} - A \mu_j)$

• $\hat{V}_j = V_j + J_j (\hat{V}_{j+1} - P_j) J_j^t$, $J_j = V_j A^t P_j^{-1}$

To compute $\hat{\mu}_j$ and $\hat{V}_j$, we need the value of μ_j and V_j
$\Rightarrow$ A <u>FORWARD PASS</u> must be completed first.



$\Rightarrow$ Then, using the 'boundary conditions' $\hat{\mu}_n = \mu_n$ and $\hat{V}_n = V_n$, a <u>BACKWARD PASS</u> yields $\hat{\mu}_j$ and $\hat{V}_j$, $j = n-1, \dots, 1$.

<u>Important</u>: Meaning of update $\hat{\mu}_j = \mu_j + J_j (\hat{\mu}_{j+1} - A \mu_j)$.

Well,

- μ_j = our best guess of z_j, computed from $x_1, \dots, x_j$, and independently of $x_{j+1}, \dots, x_n$.

$\Rightarrow$ To get our 'best' guess of z_j given the full dataset, one must somehow correct the estimate μ_j, using the additional information contained in $x_{j+1}, \dots, x_n$: $\hat{\mu}_j = \mu_j +$ some correction

→ If you only observe $x_1, \dots, x_j$, you would predict z_{j+1} using $A\mu_j$ (see page 8b)

→ But you know how to predict z_{j+1} more efficiently, using the full data set: it is given by $\hat{\mu}_{j+1}$.

Wisdom: learn from your errors.

Here, learn from $\hat{\mu}_{j+1} - A\mu_j$.

$\Rightarrow$ $\hat{\mu}_j = \mu_j +$ something proportional to $\hat{\mu}_{j+1} - A\mu_j$

And thus 'something' is exactly given by J_j, we computed it. Good.

$\Rightarrow$ Similarly for the covariance matrix: update V_j using a term involving $\hat{V}_{j+1} - \underset{\substack{\| \\ P_{j+1|j}}}{P_j}$ = error.

→ The next quantity of interest is $p(z_{j-1}, z_j \mid x_1, \dots, x_n) =: \xi(z_{j-1}, z_j)$

We have

$$\xi(z_{j-1}, z_j) = \frac{p(\underline{x} \mid z_{j-1}, z_j)\, p(z_{j-1}, z_j)}{p(\underline{x})}$$

$$= \frac{p(x_1, \dots, x_{j-1} \mid z_{j-1})\, p(x_j \mid z_j)\, p(x_{j+1}, \dots, x_n \mid z_j)\, p(z_j \mid z_{j-1})\, p(z_{j-1})}{p(\underline{x})}$$

α/β were introduced on page 11

unscaled version
$$= \frac{\alpha(z_{j-1})\, p(x_j \mid z_j)\, p(z_j \mid z_{j-1})\, \beta(z_j)}{p(\underline{x})}$$

scaled version
$$= c_j^{-1}\, \hat{\alpha}(z_{j-1})\, p(x_j \mid z_j)\, p(z_j \mid z_{j-1})\, \hat{\beta}(z_j)$$

$$\xi(z_{j-1}, z_j) = \frac{\mathcal{N}(z_{j-1} \mid \mu_{j-1}, V_{j-1})\, p(x_j \mid z_j)\, p(z_j \mid z_{j-1})\, \gamma(z_j)}{c_j\, \hat{\alpha}(z_j)}$$

$c_j\, \hat{\alpha}(z_j) = p(x_j \mid z_j)\, \mathcal{N}(z_j \mid A\mu_{j-1}, P_{j-1})$ (page 5/6)

$$= \frac{\mathcal{N}(z_{j-1} \mid \mu_{j-1}, V_{j-1})\, \mathcal{N}(z_j \mid A z_{j-1}, \Gamma)\, \mathcal{N}(z_j \mid \hat{\mu}_j, \hat{V}_j)}{\mathcal{N}(z_j \mid A\mu_{j-1}, P_{j-1})}$$

Same calculation as on page 14, with j replaced by $j-1$.

$$= \mathcal{N}(z_{j-1} \mid m_{j-1}, M_{j-1})\, \mathcal{N}(z_j \mid A\mu_{j-1}, P_{j-1})$$

$$= \frac{\mathcal{N}(z_{j-1} \mid m_{j-1}, M_{j-1})\, \cancel{\mathcal{N}(z_j \mid A\mu_{j-1}, P_{j-1})}\, \mathcal{N}(z_j \mid \hat{\mu}_j, \hat{V}_j)}{\cancel{\mathcal{N}(z_j \mid A\mu_{j-1}, P_{j-1})}}$$

$$\Rightarrow \boxed{\xi(z_{j-1}, z_j) = \mathcal{N}(z_{j-1} \mid m_{j-1}, M_{j-1})\, \mathcal{N}(z_j \mid \hat{\mu}_j, \hat{V}_j)}$$

From this expression, we may deduce the mean of the vector $\begin{pmatrix} z_{j-1} \\ z_j \end{pmatrix}$, and $\mathrm{cov}(z_{j-1}, z_j)$:

Toolbox: If $p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$
$p(y \mid x) = \mathcal{N}(y \mid Ax + b, L^{-1})$

Then $(x\ y)^t$ is gaussian with

$$\mathrm{Cov}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^t \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^t \end{pmatrix}$$

$$E\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}$$

With the notation at the bottom of page 18,

$x = z_j$ $\qquad$ $y = z_{j-1}$

$\mu = \hat{\mu}_j$ $\qquad$ $A = M_{j-1} A^t \Gamma^{-1}$

$\Lambda^{-1} = \hat{V}_j$ $\qquad$ $b = M_{j-1} V_{j-1}^{-1} \mu_{j-1}$

$\qquad\qquad$ $L^{-1} = M_{j-1}$

$$\Rightarrow \bullet\ \mathbf{E}\begin{pmatrix} z_j \\ z_{j-1} \end{pmatrix} = \begin{pmatrix} \overbrace{\hat{\mu}_j}^{\mu} \\ \underbrace{M_{j-1} A^t \Gamma^{-1}}_{A} \underbrace{\hat{\mu}_j}_{\mu} + \underbrace{M_{j-1} V_{j-1}^{-1} \mu_{j-1}}_{b} \end{pmatrix} = \begin{pmatrix} \hat{\mu}_j \\ \hat{\mu}_{j-1} \end{pmatrix}$$

(recurrence equation, bottom of page 15)

$$\bullet\ \mathrm{Cov}(z_j, z_{j-1}) = \overbrace{\hat{V}_j}^{\Lambda^{-1}} \ \overbrace{\Gamma^{-1} A M_{j-1}}^{A^t} = \hat{V}_j J_{j-1}^t$$

$= J_{j-1}^t$ from top of page 16.

Summary: $p(z_{j-1}, z_j \mid x_1, \dots, x_n) = \xi(z_{j-1}, z_j)$ is a gaussian with

- $\mathbf{E}\begin{pmatrix} z_{j-1} \\ z_j \end{pmatrix} = \begin{pmatrix} \hat{\mu}_{j-1} \\ \hat{\mu}_j \end{pmatrix}$

- $\mathrm{Cov}(z_j, z_{j-1}) = \hat{V}_j J_{j-1}^t$

  equivalently,

  $E(z_j z_{j-1}^t) = \hat{V}_j J_{j-1}^t + \hat{\mu}_j \hat{\mu}_{j-1}^t$

- $\mathrm{Cov}(z_j, z_j) = \hat{V}_j$

  or

  $E(z_j z_j^t) = \hat{V}_j + \hat{\mu}_j \hat{\mu}_j^t$.

II.2. EM algorithm.

Step I: Complete log-likelihood.

$$\mathcal{L}_c = \log p(x_1, \dots, x_n, z_1, \dots, z_n)$$

$$= \log \Big\{ \underbrace{p(z_1)}_{\mathcal{N}(z_1 \mid \mu_0, P_0)} \prod_{j=2}^{n} \underbrace{p(z_j \mid z_{j-1})}_{\mathcal{N}(z_j \mid A z_{j-1}, \Gamma)} \prod_{j=1}^{n} \underbrace{p(x_j \mid z_j)}_{\mathcal{N}(x_j \mid C z_j, \Sigma)} \Big\}$$

$$= \log\{p(z_1; \mu_0, P_0)\} + \sum_{j=2}^{n} \log p(z_j \mid z_{j-1}; A, \Gamma)$$

$$+ \sum_{j=1}^{n} \log p(x_j \mid z_j; C, \Sigma)$$

$$= -\frac{1}{2} \log |P_0| - \frac{1}{2} (z_1 - \mu_0)^t P_0^{-1} (z_1 - \mu_0)$$

$$- \frac{n-1}{2} \log |\Gamma| - \frac{1}{2} \sum_{j=2}^{n} (z_j - A z_{j-1})^t \Gamma^{-1} (z_j - A z_{j-1})$$

$$- \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{j=1}^{n} (x_j - C z_j)^t \Sigma^{-1} (x_j - C z_j)$$

(we omit terms independent of the parameters)

Step II: E-step.

Keeping the observed variables $x_1, \dots, x_n$ fixed, we need to compute the expected value of $\mathcal{L}_c$ with respect to $z \mid x_1, \dots, x_n$.

$\Rightarrow$ we only need information about $z_j \mid x_1, \dots, x_n$ & $z_{j-1}, z_j \mid x_1, \dots, x_n$, which were derived in the previous section.

We denote $\langle z_j \rangle$ for $E(z_j \mid x_1, \dots, x_n)$, given by $\hat{\mu}_j$ (bottom of page 19)

Similarly, $\begin{cases} \langle z_j z_{j-1}^t \rangle = \hat{V}_j J_{j-1}^t + \hat{\mu}_j \hat{\mu}_{j-1}^t \\ \langle z_j z_j^t \rangle = \hat{V}_j + \hat{\mu}_j \hat{\mu}_j^t \end{cases}$.

Step III: M-step.

We treat the green / red / blue terms separately:

$$Q_1 = -\frac{1}{2}\log|P_0| - \frac{1}{2} E_{Z|X}\Big(z_1^t \underline{P_0^{-1} z_1} - z_1^t \underline{P_0^{-1}\mu_0} - \mu_0^t \underline{P_0^{-1} z_1} + \mu_0^t \underline{P_0^{-1}\mu_0} \Big)$$

$$= -\frac{1}{2}\log|P_0| - \frac{1}{2} E_{Z|X} \operatorname{Tr}\big(P_0^{-1} z_1 z_1^t - P_0^{-1}\mu_0 z_1^t - P_0^{-1} z_1 \mu_0^t + P_0^{-1}\mu_0\mu_0^t \big)$$

$$= \frac{1}{2}\Big\{ \log|P_0^{-1}| - \operatorname{Tr}\big\{ P_0^{-1}\big(\langle z_1 z_1^t \rangle - \mu_0 \langle z_1 \rangle^t - \langle z_1 \rangle \mu_0^t + \mu_0 \mu_0^t \big)\big\} \Big\}$$

→ Derivatives with respect to μ_0:

$$\frac{\partial Q_1}{\partial \mu_0} = -2 P_0^{-1} \langle z_1 \rangle + 2 P_0^{-1} \mu_0 = 0$$

$$\Rightarrow \mu_0^{new} = \langle z_1 \rangle$$

computed with 'old' values of the parameters.

$$\frac{\partial}{\partial x} x^t a = \frac{\partial}{\partial x} a^t x = a$$

→ Derivatives with respect to P_0^{-1}:

$$\frac{\partial Q_1}{\partial P_0^{-1}} = \frac{1}{2}\big(P_0 - \langle z_1 z_1^t \rangle + \langle z_1 \rangle \mu_0^t + \mu_0 \langle z_1 \rangle^t - \mu_0 \mu_0^t \big) = 0$$

$$\frac{\partial}{\partial A} \operatorname{Tr}(AB) = B^t$$

$$\frac{\partial}{\partial A} \ln|A| = (A^{-1})^t$$

$$\Rightarrow P_0^{new} = \langle z_1 z_1^t \rangle - \langle z_1 \rangle (\mu_0^{new})^t - \mu_0^{new} \langle z_1 \rangle^t - \mu_0^{new} (\mu_0^{new})^t$$

$$= \langle z_1 z_1^t \rangle - \langle z_1 \rangle \langle z_1 \rangle^t$$

$$\boxed{\begin{aligned} &\text{Summarising:} \quad \mu_0^{new} = \langle z_1 \rangle \\ &\qquad\qquad\qquad P_0^{new} = \langle z_1 z_1^t \rangle - \langle z_1 \rangle \langle z_1 \rangle^t \end{aligned}}$$

$$Q_2 = \frac{n-1}{2}\log|\Gamma^{-1}| - \frac{1}{2} E_{Z|X} \sum_{j=2}^{n} (z_j - A z_{j-1})^t \Gamma^{-1} (z_j - A z_{j-1})$$

$$= \frac{n-1}{2}\log|\Gamma^{-1}| - \frac{1}{2} E_{Z|X} \sum_{j=2}^{n} z_j^t \underline{\Gamma^{-1} z_j} - z_j^t \underline{\Gamma^{-1} A z_{j-1}} - z_{j-1}^t A^t \underline{\Gamma^{-1} z_j} + z_{j-1}^t A^t \underline{\Gamma^{-1} A z_{j-1}}$$

$$= \frac{n-1}{2}\log|\Gamma^{-1}| - \frac{1}{2}\sum_{j=2}^{n} E_{Z|X} \operatorname{Tr}\big(\Gamma^{-1} z_j z_j^t - \Gamma^{-1} A z_{j-1} z_j^t - \Gamma^{-1} z_j z_{j-1}^t A^t + \Gamma^{-1} A z_{j-1} z_{j-1}^t A^t \big)$$

$$Q_2 = \frac{n-1}{2}\log|\Gamma^{-1}| - \frac{1}{2}\sum_{j=2}^{n} \mathrm{Tr}\Big\{\Gamma^{-1}\big(\langle z_j z_j^t\rangle - A\langle z_{j-1} z_j^t\rangle - \langle z_j z_{j-1}^t\rangle A^t + A\langle z_{j-1} z_{j-1}^t\rangle A^t\big)\Big\}$$

→ Derivative with respect to A:

$$\frac{\partial \mathrm{Tr}\left(\Gamma^{-1}A\langle z_{j-1} z_{j-1}^t\rangle A^t\right)}{\partial A} = 2\Gamma^{-1}A\langle z_{j-1} z_{j-1}^t\rangle$$

$$\frac{\partial \mathrm{Tr}\left(\Gamma^{-1}A\langle z_{j-1} z_j^t\rangle\right)}{\partial A} = \Gamma^{-1}\langle z_j z_{j-1}^t\rangle$$

$$\frac{\partial \mathrm{Tr}\left(\Gamma^{-1}\langle z_j z_{j-1}^t\rangle A^t\right)}{\partial A} = \Gamma^{-1}\langle z_j z_{j-1}^t\rangle$$

<u>Toolbox</u>:

$$\frac{\partial}{\partial A}\mathrm{Tr}(BACA^t) = B^tAC^t + BAC$$

$$\frac{\partial}{\partial A}\mathrm{Tr}(BAC) = B^tC^t$$

$$\frac{\partial}{\partial A}\mathrm{Tr}(BA^t) = B$$

$$\Downarrow$$

$$\frac{\partial Q_2}{\partial A} = \sum_{j=2}^{n}\Gamma^{-1}A\langle z_{j-1} z_{j-1}^t\rangle - \Gamma^{-1}\langle z_j z_{j-1}^t\rangle = 0$$

$$\Rightarrow \quad A^{new} = \left(\sum_{j=2}^{n}\langle z_j z_{j-1}^t\rangle\right)\left(\sum_{j=2}^{n}\langle z_{j-1} z_{j-1}^t\rangle\right)^{-1}$$

→ Derivative with respect to Γ^{-1}:

$$\frac{\partial Q}{\partial \Gamma^{-1}} = \frac{n-1}{2}\Gamma - \frac{1}{2}\frac{\partial}{\partial\Gamma^{-1}}\left(\sum_{j=2}^{n}\mathrm{Tr}\Big\{\Gamma^{-1}\big(\langle z_j z_j^t\rangle - A\langle z_{j-1} z_j^t\rangle - \langle z_j z_{j-1}^t\rangle A^t + A\langle z_{j-1} z_{j-1}^t\rangle A^t\big)\Big\}\right)$$

$$\frac{\partial \mathrm{Tr}(AB)}{\partial A} = B^t$$

$$= \frac{n-1}{2}\Gamma - \frac{1}{2}\sum_{j=2}^{n}\big(A\langle z_{j-1} z_{j-1}^t\rangle A^t - A\langle z_{j-1} z_j^t\rangle - \langle z_j z_{j-1}^t\rangle A^t + \langle z_j z_j^t\rangle\big) = 0$$

$$\Rightarrow \quad \Gamma^{new} = \frac{1}{n-1}\sum_{j=2}^{n}\Big\{\langle z_j z_j^t\rangle - A^{new}\langle z_{j-1} z_j^t\rangle - \langle z_j z_{j-1}^t\rangle A_{new}^t + A_{new}\langle z_{j-1} z_{j-1}^t\rangle A_{new}^t\Big\}$$

<u>Summarizing</u>:

$$A^{new} = \left(\sum_{j=2}^{n}\langle z_j z_{j-1}^t\rangle\right)\left(\sum_{j=2}^{n}\langle z_{j-1} z_{j-1}^t\rangle\right)^{-1}$$

$$\Gamma^{new} = \frac{1}{n-1}\sum_{j=2}^{n}\Big\{\langle z_j z_j^t\rangle - A_{new}\langle z_{j-1} z_j^t\rangle - \langle z_j z_{j-1}^t\rangle A_{new}^t + A_{new}\langle z_{j-1} z_{j-1}^t\rangle A_{new}^t\Big\}$$

The blue term can be treated as the red term :

wherever z_j appears $\rightarrow$ replace with x_j

z_{j-1} — " — $\rightarrow$ — " — z_j

A — " — $\rightarrow$ — " — C

& Summation starts with $j=1$.

$$\boxed{\begin{aligned} C^{new} &= \left(\sum_{j=1}^{n} x_j \langle z_j \rangle^t\right)\left(\sum_{j=1}^{n} \langle z_j z_j^t \rangle\right)^{-1} \\ \Sigma^{new} &= \frac{1}{n}\sum_{j=1}^{n}\Big\{ x_j x_j^t - C^{new}\langle z_j\rangle x_j^t \\ &\qquad - x_j \langle z_j\rangle^t C_{new}^t \\ &\qquad + C_{new}\langle z_j z_j^t\rangle C_{new}^t \Big\} \end{aligned}}$$

EM algorithm for Kalman Filtering.

(i) Initialize μ_0^{old}, P_0^{old}, A^{old}, Γ^{old}, C^{old}, Σ^{old}

(ii) Repeat

* Using old parameter values, compute
  - $\rightarrow$ FWD variables $\mu_1, \dots, \mu_n$, $V_1, \dots, V_n$ (p.7/8)
  - $\rightarrow$ then BWD variables $\hat{\mu}_n, \dots, \hat{\mu}_1$, $\hat{V}_n, \dots, \hat{V}_1$ (p.17)
  - $\rightarrow$ Calculate $\langle z_j \rangle$, $\langle z_j z_{j-1}^t \rangle$, ... (page 19)
* Update parameters
  - $\rightarrow$ μ_0^{new}, P_0^{new}, A^{new}, Γ^{new}, C^{new}, Σ^{new} (p.22/24/25)
* $\mu_0^{old} \leftarrow \mu_0^{new}$ $\quad A^{old} \leftarrow A^{new}$ $\quad C^{old} \leftarrow C^{new}$

  $P_0^{old} \leftarrow P_0^{new}$ $\quad \Gamma^{old} \leftarrow \Gamma^{new}$ $\quad \Sigma^{old} \leftarrow \Sigma^{new}$

Summary of Notation + Useful formula

(i) NOTATION.

- $\alpha(z_j) = p(x_1, \dots, x_j, z_j)$ = FWD variable

  $\beta(z_j) = p(x_{j+1}, \dots, x_n \mid z_j)$ = BWD variable.

  $c_j = p(x_j \mid x_1, \dots, x_{j-1})$

  $\alpha(z_j) = \left(\prod_{m=1}^{j} c_m\right)\hat{\alpha}(z_j)$, $\quad \hat{\alpha}(z_j) = p(z_j \mid x_1, \dots, x_j)$

  $\beta(z_j) = \left(\prod_{m=j+1}^{n} c_m\right)\hat{\beta}(z_j)$, $\quad \hat{\beta}(z_j) = \dfrac{\beta(z_j)}{p(x_{j+1}, \dots, x_n \mid x_1, \dots, x_j)}$

  $\gamma(z_j) = p(z_j \mid x_1, \dots, x_n) = \dfrac{\alpha(z_j)\beta(z_j)}{p(x_1, \dots, x_n)} = \hat{\alpha}(z_j)\hat{\beta}(z_j)$

- $\hat{\alpha}(z_j) = \mathcal{N}(z_j \mid \mu_j, V_j)$

  $\gamma(z_j) = \mathcal{N}(z_j \mid \hat{\mu}_j, \hat{V}_j)$

- $\begin{cases} \mu_j = A\mu_{j-1} + K_j(x_j - CA\mu_{j-1}) \\ V_j = (I - K_j C)P_{j-1} \end{cases}$

  where $K_j = P_{j-1}C^t(\Sigma + CP_{j-1}C^t)^{-1}$

  $P_j = \Gamma + A V_j A^t$

  $c_j = \mathcal{N}(x_j \mid CA\mu_{j-1}, CP_{j-1}C^t + \Sigma)$

  $\begin{cases} \hat{\mu}_j = \mu_j + J_j(\hat{\mu}_{j+1} - A\mu_j) \\ \hat{V}_j = V_j + J_j(\hat{V}_{j+1} - P_j)J_j^t \end{cases}$

  where $J_j = V_j A^t P_j^{-1}$

(ii) USEFUL FORMULA.

- If $p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$
  $p(y|x) = \mathcal{N}(y \mid Ax + b, L^{-1})$

Then

↘ $\begin{pmatrix} x \\ y \end{pmatrix}$ is Gaussian with $E\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}$

and $\mathrm{Cov}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^t \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^t \end{pmatrix}$

↘ $p(y) = \mathcal{N}(y \mid A\mu + b, \; L^{-1} + A\Lambda^{-1}A^t)$
$p(x|y) = \mathcal{N}(x \mid S\{A^t L(y-b) + \Lambda\mu\}, S)$

where
$S = (\Lambda + A^t L A)^{-1}.$

- Woodbury:
  $(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}$
- $(P^{-1} + B^t R^{-1} B)^{-1} B^t R^{-1} = PB^t(BPB^t + R)^{-1}$

- Matrix derivatives:

↘ $\frac{\partial}{\partial x} x^t a = \frac{\partial}{\partial x} a^t x = a$

↘ $\frac{\partial}{\partial A} \mathrm{Tr}(AB) = B^t$

$\frac{\partial}{\partial A} \ln |A| = (A^{-1})^t$

↘ $\frac{\partial}{\partial A} \mathrm{Tr}(BACA^t) = B^t A C^t + BAC$

$\frac{\partial}{\partial A} \mathrm{Tr}(BAC) = B^t C^t$

$\frac{\partial}{\partial A} \mathrm{Tr}(BA^t) = B$