10-708 08/06/2020

L10 - sequential models

(*) explore other properties of graphical models with Gaussian distri.
(*) should already be familiar with mixture models (graphically)
- factor analysis just relates both Y and X continuous - cont./discrete)

(*) can better understand embeddings
(*) HMMs as a time-sequenced MM
(*) HMM inference skipped; EX focused on variable elim; message passing interpretations

EX: what is used for inference on HMM?
(A): - viterbi algorithm? (MAP assignment)
- forward-backwards/α-β recursion/BP → param estimation
- inference tasks: - Baum-welch

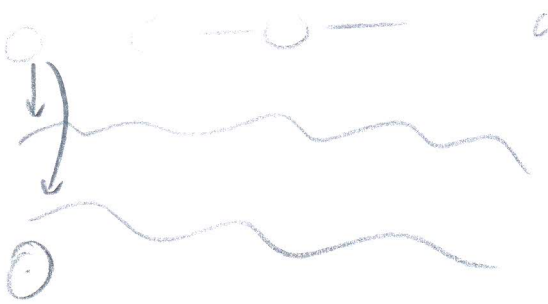
(*) today's extrapolation is to FA, SSMS (see diagram).
- Algorithms prev. studied all were counterparts
- spirit → break problem down locally into a subproblem
(*) factorial HMM, switching SSM

EX: ML is not about equations, maths; use a story/motivating problem to anchor the mathematics in a solid mental model!

(A) factorial HMM ⟶ regime-switching time-series?
- 3 dealers with dishonest casino

SSM ⟶ aircraft radar measurements (obs.)
physical locations (states)

switching SSM ⟶



- envisage S as doing generating which aircraft appears on your screen/radar.
- multiple aircrafts

- EX: think of what kind of stories can be told with these equations (i.e. corresp. with real). This is art of modelling; way more important than the mathematics. Latter is a language to express ideas, rather than an end in itself.

(*) MVG
- covered in previous lectures

(*) M.I.L.
- inversion of matrices in terms of sub-partitions.
EX: Remembered this

(*) Matrix algebra
- convert matrix → no. via trace
(A1): Review results.
$tr(\underline{x}^T \underline{A} \underline{x}) = \underline{x}^T \underline{A} \underline{x}$

(*) Factor Analysis
- as an example; imagine $\underline{x} \in \mathbb{R}^2$ (e.g. a plane).
- imagine different orientations of the plane → can be described with 3D co-ord.
- Plane is a 'manifold' subspace
- $\underline{y}$ corresponds to points in 3d space e.g. $\underline{y} \in \mathbb{R}^3$
(*) relation between $\underline{x}$ and $\underline{y}$
- orientation of subspace affects how points in manifold (subspace) are assigned co-ordinates in 3d space.
$\underline{y} = \underline{\mu} + \underline{\Lambda} \underline{x}$ (convert $\underline{x} \in \mathbb{R}^2 \rightarrow \underline{y} \in \mathbb{R}^3$ (projection?))

(*) Λ - to factor loading matrix
$\underline{\Psi}$ - diagonal

(*) (PG) - Geometric story review - important ✓
- idea of a latent space (low dimensional) (but not all the same)
- similar ideas in PCA

(*) Marginal data distn

- We have a latent factor $\underline{x}$ in a low-dim space that is unobserved
- We do observe $\underline{y}$, whose components/points we do know.

$\underline{w}$ - noise term (like ϵ)

(X) $p(\underline{x}) = N(\underline{x} | \underline{0}, \underline{I})$

↓

(Y) $p(\underline{y}|\underline{x}) = N(\underline{y} | \underline{\mu} + \underline{\Lambda}\underline{x}, \underline{\Psi})$ $\quad \underline{y} = \underline{\mu} + \underline{\Lambda}\underline{x} + \underline{w}$ $\underline{w} \sim N(0, \underline{\Psi})$

GOAL: Infer $p(\underline{x}|\underline{y})$ i.e. unobserved given observations.

- latent space is not necessarily low dimensional (e.g. radar → physical; 2D → 3D)

EX: A procedure for achieving goal

- We know $p(\underline{x})$ and $p(\underline{y}|\underline{x})$ is Gaussian.
- Hence $p(\underline{x}, \underline{y})$ is jointly Gaussian, $p(\underline{x}|\underline{y})$ is Gaussian

$$\begin{pmatrix} \underline{x} \\ \underline{y} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\underline{x}} \\ \mu_{\underline{y}} \end{pmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right)$$

- (A3): use Gaussian results to derive $p(\underline{y})$ and $p(\underline{x}|\underline{y})$

(A11): This is the key idea for inference in factor analysis

(*) being sloppy with $\underline{x}, \underline{y}, \underline{X}, \underline{Y}$ (all same) ~~it~~ i.e. capitalised.

$\mu_{\underline{x}} = 0$

$\underline{\mu} = \mathbb{E}[\underline{y}] = \mathbb{E}[\underline{\mu} + \underline{\Lambda}\underline{x} + \underline{w}] = \underline{\mu} + \underline{\Lambda}\underbrace{\mathbb{E}[\underline{x}]}_{=0} + \underbrace{\mathbb{E}[\underline{w}]}_{=0} = \underline{\mu}$

$\Sigma_{xx} = \underline{I}$

$\Sigma_{YY} = \mathrm{Var}[\underline{Y}] = \mathbb{E}[(\underline{Y} - \underline{\mu})(\underline{Y} - \underline{\mu})^T]$

$= \mathbb{E}[(\underline{\mu} + \underline{\Lambda}\underline{x} + \underline{w} - \underline{\mu})(\underline{\mu} + \underline{\Lambda}\underline{x} + \underline{w} - \underline{\mu})^T]$

$= \mathbb{E}[(\underline{\Lambda}\underline{x} + \underline{w})(\underline{\Lambda}\underline{x} + \underline{w})^T] = \mathbb{E}[(\underline{\Lambda}\underline{x}\underline{x}^T\underline{\Lambda}^T + \underline{\Lambda}\underline{x}\underline{w}^T + \underline{w}\underline{x}^T\underline{\Lambda}^T + \underline{w}\underline{w}^T)]$

$= \mathbb{E}[(\underline{\Lambda}\underline{x}\underline{x}^T\underline{\Lambda}^T + \underline{\Lambda}\mathbb{E}[\underline{x}\underline{w}^T] + \mathbb{E}[\underline{w}\underline{x}^T]\underline{\Lambda}^T + \underline{w}\underline{w}^T]$ (A4) - review ✓

$= \underline{\Lambda}\mathbb{E}[\underline{x}\underline{x}^T]\underline{\Lambda}^T + \mathbb{E}[\underline{w}\underline{w}^T] = \underline{\Lambda}\underline{\Lambda}^T + \underline{\Psi}$

$\Sigma_{xy} = \mathbb{E}[(\underline{x} - \underline{\mu}_x)(\underline{y} - \underline{\mu}_y)^T]$

(15) Review calc.

$= \mathbb{E}[(\underline{x} - \underline{\mu}_x)(\underline{\mu} + \underline{\Lambda}\underline{x} + \underline{w} - \underline{\mu})^T]$

$= \mathbb{E}[\underline{x}(\underline{\Lambda}\underline{x} + \underline{w})^T]$ $\quad \underline{\mu}_x = \underline{0}$

$= \mathbb{E}[\underline{x}\underline{x}^T]\underline{\Lambda}^T + \mathbb{E}[\underline{x}]\mathbb{E}[\underline{w}^T]$

$= \underline{\Lambda}^T$

(*) FA joint distri

- Yielding:

model $p(\underline{x}) = N(\underline{0}, \underline{I})$ $\quad p(\underline{y}|\underline{x}) = N(\underline{\mu} + \underline{\Lambda}\underline{x}, \underline{\Psi})$

cov. $\text{cov}(\underline{x}, \underline{y}) = \mathbb{E}[(\underline{x} - \underline{0})(\underline{y} - \underline{\mu})^T] = \mathbb{E}[\underline{x}(\underline{\mu} + \underline{\Lambda}\underline{x} + \underline{w} - \underline{\mu})^T]$

$= \mathbb{E}[\underline{x}\underline{x}^T\underline{\Lambda}^T + \underline{x}\underline{w}^T]$

$= \mathbb{E}[\underline{x}\underline{x}^T]\underline{\Lambda}^T + \mathbb{E}[\underline{x}]\underline{w}^T$

$\mathbb{E}[\underline{x}] = 0$ (16)

$= \underline{\Lambda}^T$

Joint distri :- $p\begin{bmatrix}\underline{x} \\ \underline{y}\end{bmatrix} = N\left(\begin{bmatrix}\underline{x} \\ \underline{y}\end{bmatrix} \middle| \begin{bmatrix}\underline{0} \\ \underline{\mu}\end{bmatrix}, \begin{bmatrix}\underline{I} & \underline{\Lambda}^T \\ \underline{\Lambda} & \underline{\Lambda}\underline{\Lambda}^T + \underline{\Psi}\end{bmatrix}\right)$

- (*) Assume noise uncorrelated with data or latent variables

---

(*) Inference in FA

- (17) - Review derivation of post.
- EX: Posterior distri that is derived → satisfactory?

(W): Yes, ~~cond~~ given that invertibility is satisfied

- Apply Gaussian condit. formulae.

- set $\Sigma_{11} = I$, $\Sigma_{12} = \Sigma_{12}^T = \Lambda^T$ $\Sigma_{22} = (\Lambda\Lambda^T + \Psi)$

- posterior of latent x given obs y

$p(x|y) \sim N(x | m_{1|2}, V_{1|2})$

$m_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y - \mu_2)$ $\quad V_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

$= \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(y - \mu)$ $\quad = I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda$ (*)

- MIL: $(E - FH^{-1}G)^{-1} = E^{-1} + E^{-1}F(H - GE^{-1}F)^{-1}GE^{-1}$ (A7)

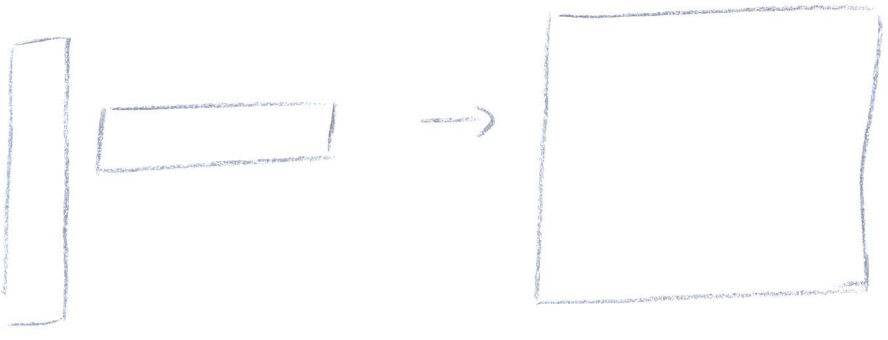
$\Rightarrow V_{1|2} = (I + \Lambda^T\Psi^{-1}\Lambda)^{-1}$ $\quad m_{1|2} = V_{1|2}\Lambda^T\Psi^{-1}(y - \mu)$

(*) computationally $\longrightarrow$ examine $\Lambda\Lambda^T$ and Ψ

- inverting Ψ is trivial as diagonal
- $\Lambda\Lambda^T$ (matrix product of factor loadings)

$E = I$
$F = \Lambda^T$
$G = \Lambda$
$H = \Psi$

- (*) Almost one-to-one correspondence



(*) - $I - I\Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda I$

- MIL allows us to re-express $V_{1|2}$ and $m_{1|2}$ in a different form

- Have different computational implications

(A8): ex exp. of computational savings of MIL through projection, dimensionality of Ψ requires clarity on your part. (inversion of smaller matrix).

methodologically * focus on this

i) get joint distr for Gaussians.
ii) compute condit. mean/covariance (according for comp.)
iii) using MIL to reduce dimensionality

(*) FA - constrained cov. Gaussian

- Review slides (A9)

(*) Geometric interp.

- Review (A10)

(*) Estimating F.A.

- Last Earlier we did inference (?) - derivation of $p(z|y)$ ..
- focus on estimation of params ; given $\{y_n\}_{n=1}^N$
  - loading matrix Λ
  - manifold center μ
  - variance Ψ

Ex: What statistical paradigm is appropriate for estimation?
(i.e. what procedure is suitable for the construction of estimators?)

- under MLE (at a cursory level):-

$$[\Lambda^*, \mu^*, \Psi^*] = \underbrace{\theta}_{\text{argmax}} \ell(\theta; y) = \underbrace{\theta}_{\text{argmax}} p(y)$$

θ (?)

(*) EM for Factor Analysis

- Incomplete data log like. (marginal) of Ψ

$$\ell(\theta, D) = -\frac{N}{2}\log|\Lambda\Lambda^T + \Psi| - \frac{1}{2}\sum_{n=1}^{N}(y_n-\mu)^T(\Lambda\Lambda^T+\Psi)^{-1}(y_n-\mu)$$

$$= -\frac{N}{2}\log|\Lambda\Lambda^T+\Psi| - \frac{1}{2}\text{tr}\left[(\Lambda\Lambda^T+\Psi)^{-1}S\right]$$

where $S = \sum_{n=1}^{N}(y_n-\mu)(y_n-\mu)^T$

- Estimating μ : $\hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} y_n$

However, est. Λ and Ψ tricky as there is a non-linear coupling in log-like

(A11)

complete log-like.

$$\ell_c(\theta; D) = \sum_{n=1}^{N} \log p(x_n, y_n) = \sum_{n=1}^{N} \log p(x_n) + \log p(y_n|x_n)$$

$$= -\frac{N}{2}\log|I| - \frac{1}{2}\sum_{n=1}^{N} x_n^T x_n - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{n=1}^{N}(y_n - \Lambda x_n)^T \Psi^{-1}(y_n - \Lambda x_n)$$

$$= -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{n=1}^{N} tr[x_n x_n^T] - \frac{N}{2} tr[S\Psi^{-1}]$$

where $S = \sum_{n=1}^{N}(y_n - \mu)(y_n - \mu)^T$

(*) E-step for factor analysis

(*) Imagine x is observed; can do inference on x^* given y; and can always compute sufficient statistics of x. (?)

(*) $p(x|y)$ - $\langle x \rangle, \langle xx^T \rangle$

(*) M-step for factor analysis

(A11) - Review derivation of EM for FA (Jordan 2003)

- Counterpart of MM with A_n continuous latent r.v.s.

Summary

1. MM has discrete latent state
   FA has continuous latent state
   } some topology graphically

2. MM $\rightarrow p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x,y)}{\sum_x p(x,y)}$ — Inference

3. FA $\rightarrow p(x), p(y|x) \longrightarrow p(x|y)$

   generative

   $\hookrightarrow p(x,y) \longrightarrow p(x|y)$ + M.I.L

4. Param est.

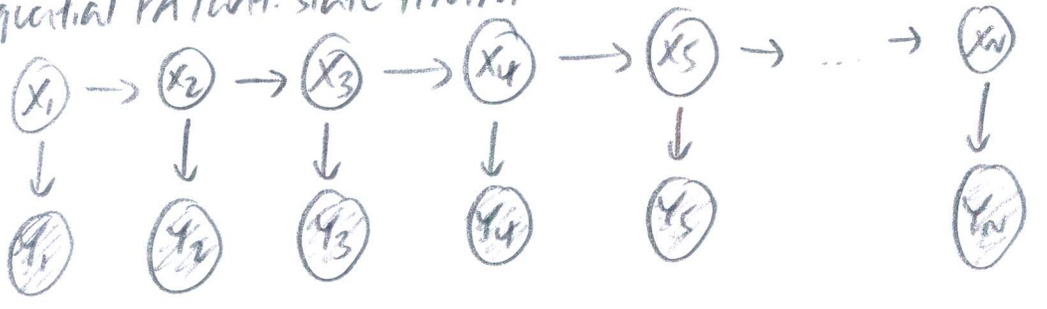
- ML estimation
- Hidden r.v.
- Benefit from inference sol. to comp. exp/suff. stat } - use EM

(*) Model variance and identifiability

(All) - even

(*) SSMs (HMM counterpart). or LDSs

- sequential FA / cont. state HMM



$x_t = A x_{t-1} + G w_t \quad w_t \sim N(0, Q) \quad x_0 \sim N(0, \Sigma_0)$

$y_t = C x_{t-1} + v_t \quad v_t \sim N(0, R)$

- An LDS

- In general: $x_t = f(x_{t-1}) + G w_{t-1}$ f - arbitrary dyn. model

$y_t = g(x_{t-1}) + v_t$ g - arb. obs. model

(*) use HMMs as a reference point for similarity/difference

(*) LDS - 2D tracking

(*) inference problem 1

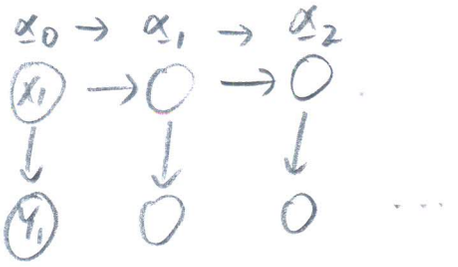
- filtering: given $y_1, y_2, y_3 \dots y_t$ estimate x_t - $P(x_t | y_{1:t})$

- given all previous observations, estimate current latent state (the pos.)

- Kalman filter (algo) → exact online inference/sequential Bayesian inference.

- Gaussian analog of forward algorithm for HMMs, in continuous space → in an LDS

$p(x_t = i | y_{1:t}) = \alpha_t^i \propto p(y_t | x_t = i) \sum_j p(x_t = i | x_{t-1} = j) \alpha_{t-1}^j$



- A non-trivial connection
- HMM as collection of sequential MMS

- forward algorithm is a recursive algorithm

(*) inference problem 2 (not emphassed)

- smoothing → given $y_1 \dots, y_T$, estimate x_t $(t<T)$
  - Rauch-Tung-Striebel smoothe
  - ↳ exact offline inference in an LDS
  - ↳ gaussian analog of forwards-backwards (alpha-gamma recursion)

(#) Kalman filtering derivation

- given $y_{1:t}$ i.e. $y_1, y_2, y_3, y_3, \dots y_t$
- desire $p(x_t | y_{1:t})$
- you already have $p(x_{t-1} | y_{t-1:t})$
  - $p(x_t | x_{t-1})$
  - $p(y_t | x_t)$ (cond.)
- due to Gaussian property :- only need mean and cov. of $p(x_t | y_{1:t})$

$\mathbb{E}[x_t | y_{1:t}] = \mu_{t|t}$ $\quad \mathbb{E}[(x_t - \mu_t)(\cdot)^T] = P_{t|t}$ (A.12) (2)

- inference → need to compute cond. mean and covariance.
- Kalman filtering is a recursive procedure to update belief state
- split into
  - i) prediction step
  - ii) update step

(*) (A13)

- Review SSM, KF derivation slides
- Continued next lecture

10-708

10/06/2020

L11 - Variational inference - loopy belief prop.

(*) continue L10 - derivation of KF for SSMs

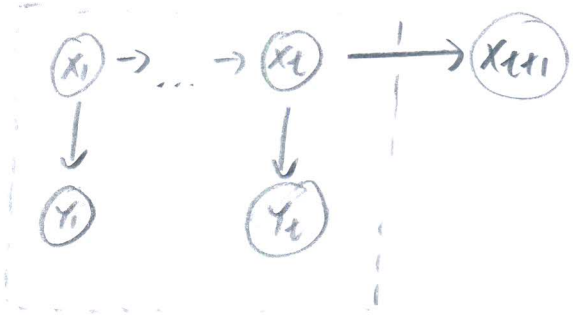
(**): solidify analogy/mental model to crystallise → true understanding.

Kalman filtering derivation

- recall filtering problem
- inference of $p(x_t|y_{1:t})$ (i.e. cond. distn of last hidden state given all observations up to now)
- use diagram as reference point for visualising queries.
- we require:-

  $p(x_t|y_{1:t})$ from a recursive procedure involving $p(x_{t+1}|y_{1:t+1})$, y_{t+1}

- Break up problem into 2 steps:-



1) time update:

- compute $p(x_{t+1}|y_{1:t})$

  from prior belief $p(x_t|y_{1:t})$ and

  dyn. model $p(x_{t+1}|x_t)$

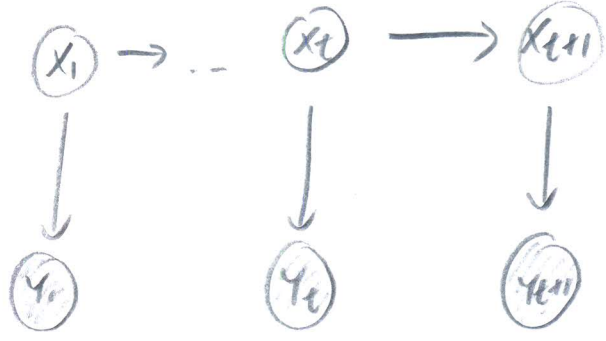
2) measurement update:

- compute new belief $p(x_{t+1}|y_{1:t+1})$

  from prediction $p(x_{t+1}|y_{1:t})$,

  observation y_{t+1}

  obs. model $p(y_{t+1}|x_{t+1})$



(*) Make use of Gaussian properties everywhere

- inference problem → dealing with new means and covariances

(**): same technique as in F.A.

(*) Schematic of the strategy:-

M.I.L.

(Gaussian manip.)

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

$z_1 \rightarrow z_1 z_2 \longrightarrow z_1 | z_2$

$\uparrow$

$z_2 | z_1 \quad (\mu_1, \mu_2)$ (*) $\qquad \mu_2 | \mu_1$ (*)

corresp.

$z_1 \sim x_t | y_{1:t} \qquad x_{t+1} = Ax_t + Gw_t = f(x_t) \longrightarrow p(x_{t+1}, y_{t+1} | y_{1:t}) \longrightarrow p(x_{t+1} | y_{1:t}, y_{t+1})$

$\uparrow$

$p(y_{t+1} | x_{t+1})$

$\uparrow$

$y_{t+1} = Cx_{t+1} + v_{t+1}$

$p(x_{t+1} | y_{1:t})$

(*) μ_1

(A1) - tidy this up (makes some sense, but want full clarity)

- Ex: conditioning is not scary; just some constants in eq.; focus on LHS of cond.

(*) Compute means and covariances within the schematic (A2) review

---

(*) Predict step

- dynamical model $x_{t+1} = Ax_t + Gw_t \quad w_t \sim N(0, Q)$

$$\hat{x}_{t+1|t} = \mathbb{E}[x_{t+1} | y_1, \ldots, y_t] = A\hat{x}_{t|t}$$

$$P_{t+1|t} = \mathbb{E}\left[(x_{t+1} - \hat{x}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})^T | y_1, \ldots, y_t \right]$$

$$= \mathbb{E}\left[(Ax_t + Gw_t - \hat{x}_{t+1|t})(Ax_t + Gw_t - \hat{x}_{t+1|t})^T | y_1, \ldots, y_t \right]$$

(A1)

$$= AP_{t|t}A + GQG^T$$

- $\hat{x}_{t+1|t}$ and $P_{t+1|t}$ are conditional means and covariances (*)

(A3) - reflect on

observation model $y_{t+1} = C x_{t+1} + v_t \quad v_t \sim N(0, R)$

$$\mathbb{E}[y_{t+1} | y_1, \dots, y_t] = \mathbb{E}[C x_{t+1} + v_{t+1} | y_1, \dots, y_t] = C \hat{x}_{t+1|t}$$

$$\mathbb{E}[(y_{t+1} - \hat{y}_{t+1|t})(y_{t+1} - \hat{y}_{t+1|t})^T | y_1, \dots, y_t] = C P_{t+1|t} C^T + R$$

(*) update step

- From previous slide:-

we have joint $p(x_{t+1}, y_{t+1} | y_{1:t}) = N(m_{t+1}, V_{t+1})$ where:-

$$m_{t+1} = \begin{pmatrix} \hat{x}_{t+1|t} \\ C \hat{x}_{t+1|t} \end{pmatrix} \qquad V_{t+1} = \begin{pmatrix} P_{t+1|t} & P_{t+1|t} C^T \\ C P_{t+1|t} & C P_{t+1|t} C^T + R \end{pmatrix}$$

- Now use conditional Gaussian formulae for partitioned matrices to get $p(x_{t+1} | y_{t+1}, y_{1:t})$

(*) Kalman Filter

- Yields closed form measurement and time updates:-

measurement update :- $\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + K_{t+1}(y_{t+1} - C \hat{x}_{t+1|t})$

$$P_{t+1|t+1} = P_{t+1|t} - K C P_{t+1|t} \qquad (?)$$

- K_{t+1} - Kalman gain matrix

Time updates: $\hat{x}_{t+1|t} = A \hat{x}_{t|t}$

$$P_{t+1|t} = A P_{t|t} A + G Q G^T$$

$K_{t+1} = P_{t+1|t} C^T (C P_{t+1|t} C^T + R)^{-1}$

(A3) - Review notation and intuition

Ex. Be sensitive to conditioning in terms of info being used at each point.

(*) Example of KF in 1D

- Just apply KF equations

(PQ) - intuition behind these

(*) KF intuition (ON)

- KF update of mean:- $\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(z_t - C\hat{x}_{t|t-1})$ - innov.

$$= \frac{(\sigma_z + \sigma_x) z_t + \sigma_z \hat{x}_{t|t-1}}{\sigma_z + \sigma_x + \sigma_z}$$

(ON) new belief is convex combo of updates from prior and observation, weighted by Kalman gain matrix

$$K_t = P_{t|t-1} C^T (C P_{t|t-1} C^T + R)^{-1}$$

- observation unreliable, σ_z (i.e. R) large so K_t is small → more att. to prediction

- old prior is unreliable, large σ_x, or process is unpredictable, large σ_x, we pay more attention to observation.

(Q): where do A, G, C matrices come from?

- A parameter estimation (as opposed to filtering, which is an inference problem)
  - inference problem: i.e. compute p.d. of latent var, given data in this context.
- inference assumes learning/ param est. has already occurred.

(*) Complexity of one KF step., Inf prob 2, RTS smoother, RTS deriv.

- Review not in our time

(*) Learning SSMs

(*) use EM. Inference from KF or RTS filtering forms E-step!

(*) Form complete log-like.

(*) E-step → compute suffi. statistics using inference results to give you a posterior

M-step → MLE

EX: You can & should derive in your spare time.

(*) Key principle → inference can be treated as a subroutine for learning in a partially obs. setting

(*) Nonlinear systems

- use derivatives, Taylor exp. approxim. / linearization.

(*) C.S thinking → you can approximate functions to the order you can afford

(*) EX A lot of seemingly difficult, obtuse algebra can be viewed/reduced through an understanding of principles motivating them (often recur).