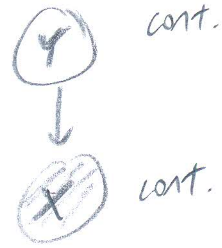
10-708 08/06/2020

L10 - sequential models

(*) explore other properties of graphical models with Gaussian distri.

(*) should already be familiar with mixture models (graphically)

- factor analysis just includes 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?

Ⓐ: - 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 have counterparts

- spirit → break problem down locally into a subproblem

(*) factorial HMM, switching SSM

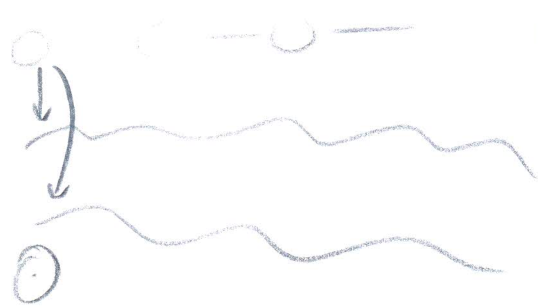
EX: ML is not about equations, maths; use a story/motivating problem to anchor the mathematics in a solid mental model.

Ⓐ. factorial HMM ⟶ regime-switching time-series?

- 3 dealers with dishonest casino

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

switching SSM ⟶



- Envisage S as clock 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 $\longrightarrow$ 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 $\longrightarrow$ can be described with 3D co-ord.
- Plane is a 'manifold' subspace
- y corresponds to points in 3d space e.g. $y \in \mathbb{R}^3$

(*) Relation between $\underline{x}$ and y

- orientation of subspace affects how points in manifold (subspace) are assigned co-ordinates in 3d space.

$y = \underline{\mu} + \underline{\Lambda} \underline{x}$ (convert $\underline{x} \in \mathbb{R}^2 \to y \in \mathbb{R}^3$ (projection?))

(*) $\underline{\Lambda}$ - to factor loading matrix

$\underline{\Psi}$ - diagonal

(*) (A2) - Geometric story review - important ✓

- idea of a latent space (lower dimensional) (but not all the same

- similar ideas in PCA

(*) Marginal data distn

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

$\underline{w}$ - noise term (like $\underline{\epsilon}$)

$$X \rightarrow Y$$

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

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

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

- latent space is not necessarily lower dimensional (e.g. radar $\underset{2D}{} \rightarrow$ physical $\underset{3D}{}$)

EX: A procedure for achieving goal.

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

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

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

E11: This is the key idea for inference in factor analysis

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

$$\mu_x = 0$$

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

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

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

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

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

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

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

$\Sigma_{xy} = \mathbb{E}[(\underline{X}-\mu_x)(\underline{Y}-\mu_Y)^T]$ (15) Review calc.

$= \mathbb{E}[(\underline{X}-\mu_x)(\underline{\mu}+\Lambda\underline{X}+\underline{W}-\underline{\mu})^T]$

$= \mathbb{E}[\underline{X}(\Lambda\underline{X}+\underline{W})^T]$ $\mu_x = \underline{0}$

$= \mathbb{E}[\underline{X}\underline{X}^T]\Lambda^T + \mathbb{E}[\underline{X}]\mathbb{E}[\underline{W}^T]$

$= \underline{\Lambda}^T$

(*) FA joint distn

- Yielding:

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

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

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

$= \mathbb{E}[\underline{x}\underline{x}^T]\Lambda^T + \underbrace{\mathbb{E}[\underline{x}]\underline{W}^T}_{=0}$ (16)

$= \Lambda^T$

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

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

---

(*) Inference in FA

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

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

- Apply Gaussian condit. formulae.

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

- posterior of latent x given obs y

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

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

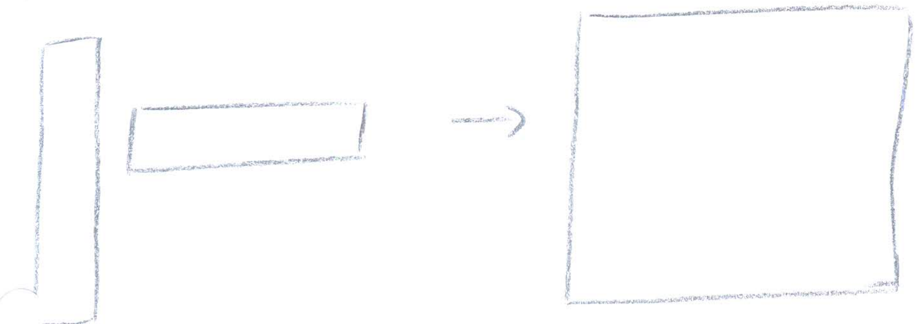
$$= \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(y - \mu) \qquad = I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda \ (*)$$

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

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

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

(*) computationally $\longrightarrow$ examine $\Lambda\Lambda^T$ and Ψ
- inverting Ψ is trivial as diagonal
- $\Lambda\Lambda^T$ (matrix product of factor loadings)



- (*) 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 distri for Gaussians.
ii) compute condit. mean/covariance (accounting for comp.)
iii) using MIL to reduce dimensionality

(*) FA - constrained cov. Gaussian

- Review slides A9

(*) Geometric interp.

- Review A10

(*) Estimating F.A.

- Earlier we did inference ? - derivation of $p(x|y)$ ..
- Focus on estimation of params, given $\{y_n\}_{n=1}^{N}$

  - loading matrix $\underline{\Lambda}$
  - manifold center $\underline{\mu}$
  - variance $\underline{\Psi}$

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):-

$$\underbrace{[\underline{\Lambda}^*, \underline{\mu}^*, \underline{\Psi}^*]}_{\underline{\theta}} = \underset{\underline{\theta}}{\operatorname{argmax}}\ l(\underline{\theta}; y) = \underset{\underline{\theta}}{\operatorname{argmax}}\ p(y)$$?

(*) EM for Factor Analysis

- incomplete data log like. (marginal) of y

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

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

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

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

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

complete log-like.

A1

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

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

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

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. (?)

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

(*) M-step for factor analysis

A1 - Review derivation of EM for FA (Jordan 2003)

- A counterpart of MM with continuous latent r.v.s.

Summary

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

Inference

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

3. FA $\rightarrow \underbrace{p(x), p(y|x)}_{\text{generative}} \longrightarrow p(x|y)$

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

4. Param est.

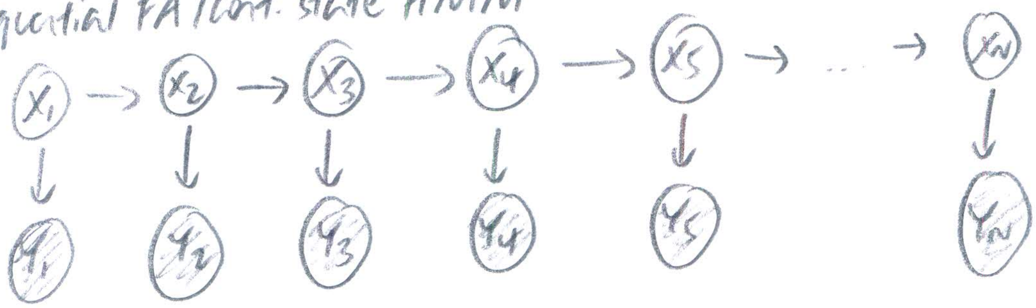
- ML estimation
- Hidden r.v. } - use EM
- Benefit from inference sol. to comp. exp/sufficient

(*) Model variance and identifiability

(All) - review

(*) 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.

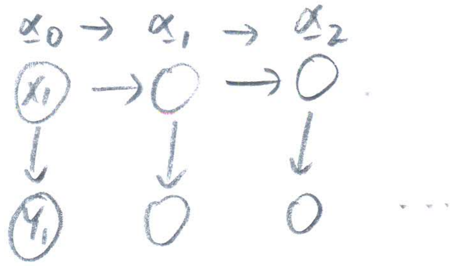
- filtering: given $y_1, y_2, y_3 \ldots y_t$ estimate x_t - $P(x_t | y_{1:t})$

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

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

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

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



- A non-trivial connection

- HMM as collection of sequential MMs.

- Forward algorithm is a recursive algorithm

(*) Inference problem 2 (not emphasised)

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

(*) Kalman filtering derivation

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

$\mathbb{E}[x_t | y_{1:t}] = \mu_{t|t} \qquad \mathbb{E}[(x_t - \mu_t) \ ? \ ] = P_{t|t}$ A12 ?

- Inference $\longrightarrow$ 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

(*) (113)

- Review SSM, KF derivation slides
- continued next lecture