

110-Review

## (A) Focus on

- resolving queries in notes
- FA derivation
- EM for FA
- KF derivation (set aside RTS smoothing)

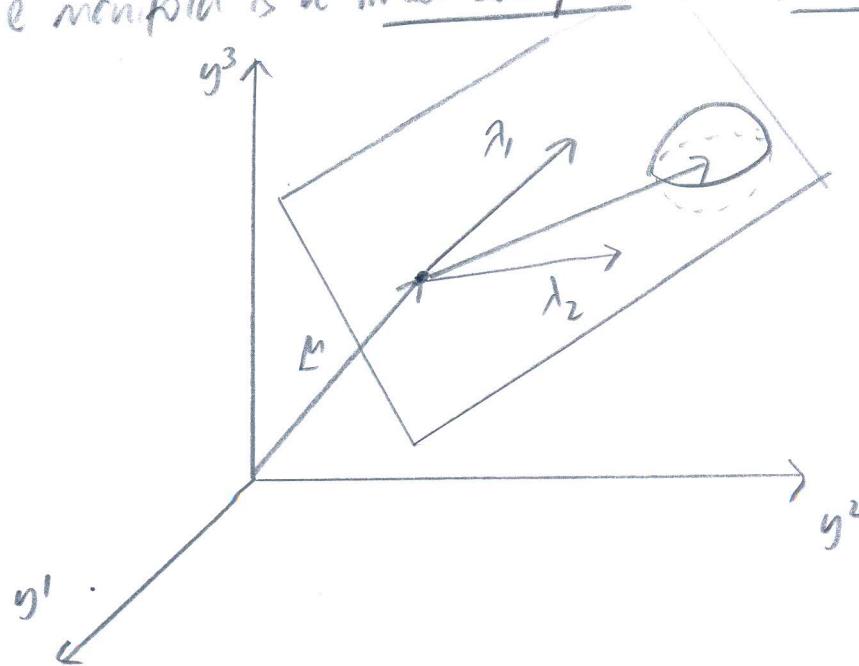
- using Jordan (2003)  
ch 14 + 15

## (B) Factor Analysis

- motivation  $\rightarrow$  lectures were not fully clear.
- latent variable with latent variable is a continuous random vector
- density estimation issue
- measured data vector high dim; but we may have reason to believe data lie in low-dim manifold.
- 2 stage generative process.
  - 1) point in manifold generated via a simple marg. p.d.
  - 2) observed data generated from a simple density that is centered on a point.
- coordinates of point  $\rightarrow$  components of latent random vector
- If we wish to parameterize cont. manifold; latent variable is cont. random vector.

## (C) Factor analysis (mathem. intuition)

- Assume manifold is a linear subspace  $\rightarrow$  factor analysis



- observed data:

- in a p-dim. subspace  $M$  in  $\mathbb{R}^q$  where  $p < q$ .
  - given a set of basis vectors  $\{\underline{\lambda}_j\}$ , a point in  $M$  can be represented as a product  $\underline{\lambda}\underline{x}$
  - $\underline{x}$  - w-ordinate vector
  - $\underline{\lambda}$  - columns are basis vectors  
(factor loading mat.)
  - treat w-ordinate vectors as values of continuous r vec  $\underline{X}$
  - treat  $\underline{X}$  with a p.d.
- ~~
- FA:  $\underline{X}$  is a Gaussian r vec
  - Given a point in  $M$ ; observed  $\underline{Y}$  assumed to be gen acc. to a Gaussian distributed about that point. (0/5) • not entirely clear
  - resulting density  $\rightarrow$
  - convolution of Gaussian density on manifold  $M$  with Gaussian distn extending into  $\mathbb{R}^q \rightarrow$  'thick subspace' in  $\mathbb{R}^q$
- (\*) Jordan: degeneracy in FA an issue if you are concerned with identifiability of basis vectors; but often not preoccupation of density estimation

- PCA  $\rightarrow$  FA

- K-means  $\rightarrow$  GMM.

(\*) model details

- Supplementing slides and slide notes from Jordan (2003)

$\underline{x} \in \mathbb{R}^p$   $\underline{y} \in \mathbb{R}^q$   $p < q$  (vectors)

- dimensionality tightening:-

$$p(\underline{x}) = N(\underline{x} | \underline{0}, \underline{I}) \quad \underline{x} \sim N(\underline{0}, \underline{I}) \quad \underline{\lambda} \in \mathbb{R}^{q \times p} \quad \underline{\varphi} \in \mathbb{R}^{q \times q}$$

$$p(y|\underline{x}) = N(y | \mu + \Lambda \underline{x}, \underline{\varphi}) \quad y|\underline{x} \sim N(\mu + \Lambda \underline{x}, \underline{\varphi})$$

(\*) note  $\Psi \in \mathbb{R}^{q \times q}$  is diagonal  
 (add that in gen case of  $p(\underline{x}) = n(\underline{x}_0, \Sigma_0)$  and  $\Psi = \sigma^2 \mathbb{I} \rightarrow \text{PPCA}$ )

generative model:

- generate a point within manifold  $M$   $\xrightarrow{\text{by sampling } \underline{x} \sim N(0, \mathbb{I})}$

- add noise to get  $y = \mu + \Lambda \underline{x} + \underline{w}$  by sampling  $\underline{w} \sim N(0, \Psi)$

(\*) FA:

- specify  $p(\underline{x})$  and  $p(y|\underline{x})$   $\xrightarrow{\text{Gaussian results}}$  derive joint  $p(\underline{x}, y) \rightarrow$  derive  $p(x|y)$   
 + M.L  
 - we can use  $p(x|y)$  for inference; or as a means to parameter estimation.

specifying joint  $p(\underline{x}, y)$  using Gaussian properties requires specifying  
 marginal  $p(y)$  (or at least its mean and variance).

$\Rightarrow$  compute  $\mu_y = \mathbb{E}[y]$  and  $\Sigma_{yy} = \text{Var}[y]$

(\*) Inflating capitalisation  
 $X, \underline{x}, y, \mathbb{I}$   
 (soz.)

(A4):- for 100% clarity:-

-  $\mathbb{E}[\underline{x}\underline{w}^T]$  and  $\mathbb{E}[\underline{w}\underline{x}^T]\mathbb{I}^T$  are both 0.

- Because  $\mathbb{E}[\underline{x}\underline{w}^T] = \mathbb{E}[\underline{x}]\mathbb{E}[\underline{w}^T] = 0$

as  $\underline{x}$  and  $\underline{w}$  are independent  $\Rightarrow$  expectation of product  
 is product of expectations

note that:-  $\mathbb{E}[\underline{x}\underline{x}^T] = \text{cov}(\underline{x}) = \mathbb{I}$

$\mathbb{E}[\underline{w}\underline{w}^T] = \text{cov}(\underline{w}) = \Psi$

Hence  $y \sim N(\mu, \Lambda\Lambda^T + \Psi)$

$$(A5) / \Sigma_{xy} = \text{cov}(\underline{x}, y) = \mathbb{E}[(\underline{x} - \mu_x)(y - \mu_y)^T]$$

$$(A6) = \mathbb{E}[\underline{x}(\mu + \Lambda\underline{x} + \underline{w} - \mu)^T]$$

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

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

$$\underset{=0}{\approx}$$

(\*) See Jordan (2003) for how some result on marginals using iterated expectations and variance.

$$\cdot \mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$$

$$\cdot \text{Var}[Y] = \text{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[(\text{Var}(Y|X))]$$

$$\cdot \mathbb{E}[Y] = \mathbb{E}[\mu + \Delta X] = \mu + \Delta \mathbb{E}[X] \quad \mathbb{E}[Y|X] = \mu + \Delta X \\ = \mu$$

$$\cdot \text{Var}[Y] = \text{Var}(\mu + \Delta X) + \mathbb{E}[\underline{\text{Var}}]$$

$$= \text{Var}(\Delta X) + \underline{\text{Var}} = \mathbb{E}[(\Delta X - \overbrace{\mathbb{E}[\Delta X]}^{\sim 0})(\Delta X - \overbrace{\mathbb{E}[\Delta X]}^{\sim 0})^T]$$

$$= \mathbb{E}[(\Delta X)(\Delta X)^T] + \underline{\text{Var}}$$

$$= \mathbb{E}[(\Delta X X^T \Delta)] + \underline{\text{Var}}$$

$$= \Delta \mathbb{E}[XX^T] \Delta^T + \underline{\text{Var}} = \Delta \Delta^T + \underline{\text{Var}}$$

- yielding same results (via diff. means).

(\*) ok, Jordan does distinguish between  $X, Z$ ;  $Y, y$  (i.e. random vectors and realisations).

- just remember :-  $X \sim N(0, I)$

(invaria)  
 $Y \sim N(\mu + \Delta Z, \Sigma)$  i.e.  $Y|Z \sim N(\mu + \Delta Z, \Sigma)$

(\*) once  $p(y)$  and  $p(z, y)$  specified; or means and variances; can derive  $p(x|y)$  i.e.  $\mathbb{E}[X|y]$  and  $\text{Var}[X|y]$

$$\mathbb{E}[X|y] = \Delta^T (\Delta \Delta^T + \Sigma)^{-1} (y - \mu) \quad (i)$$

$$= (I + \Delta^T \Sigma^{-1} \Delta)^{-1} \Delta^T \Sigma^{-1} (y - \mu) \quad (ii) \text{ (via MLE)} \quad (ii, vi)$$

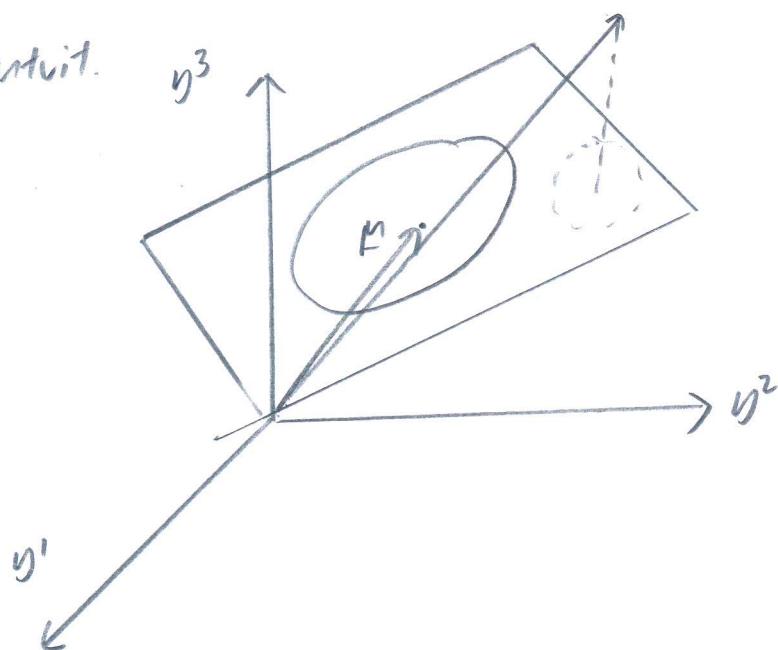
(i)  $n \times q \times q$  matrix

(ii)  $n \times p \times p$  matrix.

$$\text{var}(x|y) = I - D^T(DD^T + y)^{-1}D$$

$$= (I + D^Ty^{-1}D)^{-1} \quad (\text{via MIL}) \quad (14.26)$$

(\*) Geometric intuit.



- Before observing  $y$ ; distri of  $X$  is Gaussian centred around origin of latent variable subspace. (solid ellipse is distri of  $X$ )
- After obs.  $y=y$  obtain an update distri for  $X$  with (14.24) determining mean and (14.26) updated covariance matrix of update distri. (distri  $X$  depicted as dotted ellipse)
- Essentially; we project  $y$  onto latent subspace (not just point proj.) but an uncertainty estimate also.

6/5 i): don't fully understand geometric intuition here  
 ↳ needs some more time / another crack

(\*) MLE of FA via EM (14.2) Jordan (2003)

Jordan (2003) on this section informs most of the slides

0) - likelihood remains invariant to orthogonal transformations of  $D$ : - i.e.  $R^T R = R R^T = I$

- showing this: -  $R$  - orthogonal mat.  $\tilde{D} = DR$

$$\tilde{D}\tilde{D}^T = DR(DR)^T = DR R^T D^T = D D^T$$

$\Rightarrow$  likelihood, which depends on  $\Lambda$  only through product  $\Lambda\Lambda^T$  is not changed if  $\Lambda$  is postmultiplied by an orthog mat.  $R$ .

(ii) nonlinear coupling: parameters  $\Lambda$  and  $\Psi$  are coupled by determinant and inverse.

(iii) lecture slides already specify incomplete cd comp. data log like.;

(iv) incomplete data log-like :- (using neg density p(y))

$$l(\theta | D) = -\frac{N}{2} \log |\Lambda\Lambda^T + \Psi| - \frac{1}{2} \left\{ \sum_{n=1}^N (\mathbf{y}_n - \boldsymbol{\mu})^T (\Lambda\Lambda^T + \Psi)^{-1} (\mathbf{y}_n - \boldsymbol{\mu}) \right\} \quad (14.27)$$

where  $D = \{\mathbf{y}_n : n=1, \dots, N\}$  (IID)

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{N} \sum_{n=1}^N \mathbf{y}_n \quad (14.29)$$

- omit ref to mean  $\boldsymbol{\mu}$  from here on i.e. estimate  $\boldsymbol{\mu}$  via (14.29);  
subtract from data vectors  $\mathbf{y}_n$ ; use these centred variables when we specify data vectors  $\mathbf{y}_n$ .

(v) complete data log-likelihood:-

- complete data =  $\{(x_n, y_n) : n=1, \dots, N\} = D_c$        $\textcircled{D}$  - note decoupling of  $\Lambda$  and  $\Psi$   
- product of joint Gaussians

$$\begin{aligned} l_c(\theta | D_c) &= \sum_{n=1}^N \log p(x_n, y_n) = \sum_{n=1}^N \log p(x_n) + \log p(y_n | x_n) \\ &= -\frac{N}{2} \log |\Lambda| - \frac{1}{2} \sum_{n=1}^N x_n^T x_n - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_{n=1}^N (\mathbf{y}_n - \Lambda x_n)^T \Psi^{-1} (\mathbf{y}_n - \Lambda x_n) \\ &= -\frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_{n=1}^N x_n^T x_n - \frac{1}{2} \sum_{n=1}^N (\mathbf{y}_n - \Lambda x_n)^T \Psi^{-1} (\mathbf{y}_n - \Lambda x_n) \end{aligned}$$

- Rewrite quadratic forms  
using trace, using cyclic prop

$$\begin{aligned} &= -\frac{N}{2} \log |\Psi| - \underbrace{\frac{1}{2} \sum_{n=1}^N \text{tr}(x_n x_n^T)}_{\textcircled{1}} - \underbrace{\frac{1}{2} \sum_{n=1}^N \text{tr}[(\mathbf{y}_n - \Lambda x_n)(\mathbf{y}_n - \Lambda x_n)^T \Psi^{-1}]}_{\textcircled{2}} \\ &\quad \textcircled{S} \quad \textcircled{3} \\ &= -\frac{N}{2} \log |\Psi| - \frac{N}{2} \text{tr}(S \Psi^{-1}) \quad : 0 \text{ (due to mean } \mathbf{y} = \mathbf{0} \text{?)} \end{aligned}$$

0/32

- where  $S := \frac{1}{N} \sum_{n=1}^N (y_n - \Delta x_n)(y_n - \Delta x_n)^T$  (form of sample cov. matrix; but dep. on  $\Delta$ ).

## E-step for FA

### E-step 14.2.3

- $\langle \cdot \rangle$  - condit. exp. in this context.
- take condit. expectation of complete data log-likelihood, conditioning on observed data  $y$  and current param vector  $\theta^{(t)}$ .
- $Q(\theta | \theta^{(t)}) = \langle \ell_C(\theta | \theta_C) \rangle = -\frac{N}{2} \log |\Sigma| - \frac{N}{2} \text{tr}(\langle S \rangle \Sigma^{-1})$
- we have to compute cond. exp.  $\langle S \rangle$
- it requires substituting r.v.  $x_n$  for  $\tilde{x}_n$  in defn. of  $S$  and treat  $S$  as a random quantity

$$\begin{aligned}\langle S \rangle &= \frac{1}{N} \sum_{n=1}^N \langle (y_n - \Delta x_n)(y_n - \Delta x_n)^T \rangle \\ &= \frac{1}{N} \sum_{n=1}^N \langle (y_n y_n^T - y_n x_n^T \Delta^T - \Delta x_n y_n^T + \Delta x_n x_n^T \Delta) \rangle \quad \text{- note } x_n \text{ is r.v.} \\ &= \frac{1}{N} \sum_{n=1}^N y_n y_n^T - \langle x_n \rangle \Delta^T - \Delta \langle x_n \rangle y_n^T + \Delta \langle x_n x_n^T \rangle \Delta\end{aligned}$$

- Note that this merits use of C.E.  $\langle x_n \rangle$  and  $\langle x_n x_n^T \rangle$

(\*) expected sufficient statistics required for E-step are C.E.

$$\langle x_n \rangle \text{ and } \langle x_n x_n^T \rangle$$

(\*\*) obtained from our derivation of  $p(x|y)$  and (mean and var of this dist.)

$$\langle x_n \rangle = E[x_n | y_n]$$

$$\langle x_n x_n^T \rangle = \text{var}(x_n | y_n) + E[x_n | y_n] E[x_n | y_n]^T$$

6/53 (?) - what form are you not clocking?

$$14.2.4 \text{ M-step} = Q(\theta | \theta^{(t)})$$

- take derivatives of  $\langle \ell_c(\theta | D_c) \rangle$  wrt parameters; solve to get ML estimates

$$\frac{\partial}{\partial \underline{\Omega}} \langle \ell_c(\theta | D_c) \rangle = \left( \frac{\partial}{\partial \underline{\Omega}} \left( -\frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_{n=1}^N \text{tr} [\langle X_n X_n^T \rangle] - \frac{N}{2} \text{tr} [\langle S \rangle \Psi^{-1}] \right) \right)$$

$\underbrace{\text{contains } \Delta}$

$$\textcircled{1} \quad \frac{\partial}{\partial \underline{\Omega}} \langle S \rangle = -\frac{N}{2} \Psi^{-1} \frac{\partial}{\partial \underline{\Omega}} \langle S \rangle$$

$$\textcircled{2} \quad = -\frac{N}{2} \Psi^{-1} \frac{\partial}{\partial \underline{\Omega}} \left( \frac{1}{N} \sum_{n=1}^N (y_n y_n^T - y_n \langle X_n^T \rangle \Omega^T - \Omega \langle X_n \rangle y_n^T + \Omega \langle X_n X_n^T \rangle \Omega^T) \right)$$

$$= -\frac{N}{2} \Psi^{-1} \frac{\partial}{\partial \underline{\Omega}} \left( \frac{1}{N} \sum_{n=1}^N y_n y_n^T - 2 y_n \langle X_n^T \rangle \Omega^T + \Omega \langle X_n X_n^T \rangle \Omega^T \right)$$

$$= -\frac{N}{2} \Psi^{-1} \left( \frac{1}{N} \sum_{n=1}^N -2 y_n \langle X_n^T \rangle + 2 \Omega \langle X_n X_n^T \rangle \right)$$

- setting  $\frac{\partial}{\partial \underline{\Omega}} \langle \ell_c(\theta | D_c) \rangle = 0$

$$\Rightarrow \sum_{n=1}^N \Psi^{-1} y_n \langle X_n^T \rangle - \sum_{n=1}^N \Psi^{-1} \Omega \langle X_n X_n^T \rangle = 0$$

$$\Rightarrow \sum_{n=1}^N \Psi^{-1} y_n \langle X_n^T \rangle = \sum_{n=1}^N \Psi^{-1} \Omega \langle X_n X_n^T \rangle \quad \text{pre mult. by } \Psi$$

$$\Rightarrow \hat{\Omega}^{(t+1)} = \left( \sum_{n=1}^N y_n \langle X_n^T \rangle \right) \left( \sum_{n=1}^N \langle X_n X_n^T \rangle \right)^{-1}$$

(\*) these are normal equiv. form LR!

- for details on why it is not exactly equal (i.e. transpose) → see Jordan (2003).

- But that's cosmetic

$$\frac{\partial}{\partial \underline{\Psi}} \langle \ell_c(\theta | D_c) \rangle = \frac{\partial}{\partial \underline{\Psi}} \left( -\frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_{n=1}^N \text{tr} [\langle X_n X_n^T \rangle] - \frac{N}{2} \text{tr} [\langle S \rangle \Psi^{-1}] \right)$$

- noting :-

$$\begin{aligned}\frac{\partial}{\partial \underline{\Lambda}} \text{tr} [\underline{y}_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^T \underline{\psi}^{-1}] &= \frac{\partial}{\partial \underline{\Lambda}} \text{tr} [\underline{\psi}^{-1} \underline{y}_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^T] = \frac{\partial}{\partial \underline{\Lambda}} \text{tr} [(\underline{\psi}^{-1} \underline{y}_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^T)^T] \\ &= \frac{\partial}{\partial \underline{\Lambda}} \text{tr} \left[ \underbrace{(\underline{\psi}^{-1} \underline{y}_n \langle \underline{x}_n^T \rangle)}_{B'} \right]^T \underline{\Lambda} \end{aligned}$$

as  $\text{tr}(ABC) = \text{tr}(CAB)$   
 $= \text{tr}(BCA)$   
and  $\text{tr}(A^T) = \text{tr}(A)$

As  $\frac{\partial}{\partial A} \text{tr}[BA] = B^T$ ;

$$\Rightarrow \frac{\partial}{\partial \underline{\Lambda}} \text{tr} [\underline{y}_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^T \underline{\psi}^{-1}] = \underline{\psi}^{-1} \underline{y}_n \langle \underline{x}_n^T \rangle$$

- AND;

$$\frac{\partial}{\partial \underline{\Lambda}} \text{tr} [\underline{\Lambda} \langle \underline{x}_n \underline{x}_n^T \rangle \underline{\Lambda}^T \underline{\psi}^{-1}] = \frac{\partial}{\partial \underline{\Lambda}} \text{tr} \left[ \underbrace{\langle \underline{x}_n \underline{x}_n^T \rangle}_{B} \underbrace{\underline{\Lambda}^T}_{A^T} \underbrace{\underline{\psi}^{-1} \underline{\Lambda}}_{C} \right]$$

As  $\frac{\partial}{\partial A} \text{tr}[BA^TCA] = 2CAB$

where B,C are symmetric;

$$\frac{\partial}{\partial \underline{\Lambda}} \text{tr} [\underline{\Lambda} \langle \underline{x}_n \underline{x}_n^T \rangle \underline{\Lambda}^T \underline{\psi}^{-1}] = 2\underline{\psi}^{-1} \underline{\Lambda} \langle \underline{x}_n \underline{x}_n^T \rangle$$

Hence  $\frac{\partial \underline{\Lambda}}{\partial \underline{\Lambda}} = \sum_{n=1}^N \underline{\psi}^{-1} \underline{y}_n \langle \underline{x}_n^T \rangle - \sum_{n=1}^N \underline{\psi}^{-1} \underline{\Lambda} \langle \underline{x}_n \underline{x}_n^T \rangle$  as required.

↳ including M-step deriv. (Jordan 2003)

$S = f(\underline{\Lambda})$ ; substitute  $\underline{\Lambda}^{(t+1)}$  into exp for  $\langle S \rangle$

$$\begin{aligned}\langle S \rangle &= \frac{1}{N} \sum_{n=1}^N \left( \underline{y}_n \underline{y}_n^T - \underline{y}_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^{(t+1)T} - \underline{\Lambda}^{(t+1)} \langle \underline{x}_n \rangle \underline{y}_n^T + \underline{\Lambda}^{(t+1)} \langle \underline{x}_n \underline{x}_n^T \rangle \underline{\Lambda}^{(t+1)T} \right) \\ &= \frac{1}{N} \left( \sum_n \underline{y}_n \underline{y}_n^T - \sum_n \underline{y}_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^{(t+1)T} - \underline{\Lambda}^{(t+1)} \sum_n \langle \underline{x}_n \rangle \underline{y}_n^T + \underline{\Lambda}^{(t+1)} \sum_n \langle \underline{x}_n \underline{x}_n^T \rangle \underline{\Lambda}^{(t+1)T} \right)\end{aligned}$$

derivation is tedious to write; note that

$$\underline{\Lambda}^{(t+1)} \sum_n \langle \underline{x}_n \underline{x}_n^T \rangle \underline{\Lambda}^{(t+1)T} = \sum_n \underline{y}_n \langle \underline{x}_n^T \rangle \left\{ (\sum_n \langle \underline{x}_n \rangle) (\sum_n \langle \underline{x}_n \underline{x}_n^T \rangle)^{-1} \right\}^T = \sum_n \underline{y}_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^{(t+1)T}$$

- not sure about this next part:-

?

-  $\underline{\Psi}$  is a diagonal matrix

- to calculate derivative of  $Q(\underline{\Psi} | \theta^{(t)})$  wrt  $\underline{\Psi}$ , take usual matrix derivative;  
but retaining only diagonal terms.

0/51 ⑦

- take derivative wrt  $\underline{\Psi}^{-1}$ , set to 0, retaining only diag.: - yields:-

$$\underline{\Psi}^{(t+1)} = \text{diag}(\langle \underline{\Sigma} \rangle)$$

(\*) From slides of lecture:- (EX 5.2019 v10)

$$\begin{aligned}\frac{\partial}{\partial \underline{\Psi}^{-1}} \langle \underline{\psi}(\theta | \underline{\theta}_c) \rangle &= \frac{\partial}{\partial \underline{\Psi}^{-1}} \left( -\frac{N}{2} \log |\underline{\Psi}| - \frac{1}{2} \sum_{n=1}^N \underline{x}_n \underline{x}_n^T - \frac{N}{2} \text{tr}(\langle \underline{\Sigma} \rangle \underline{\Psi}^{-1}) \right) \\ &= \frac{N}{2} \underline{\Psi} - \frac{N}{2} \langle \underline{\Sigma} \rangle \\ \Rightarrow \quad \underline{\Psi}^{(t+1)} &= \langle \underline{\Sigma} \rangle\end{aligned}$$

⑦

⑦: lecture slides and Jordan (2003) don't match  $\rightarrow$  you need to figure out which is correct. → continued

(\*) Another issue which requires resolution is in previous ① 0/52 ⑦ :- (digression)

i.e. where EX lecture slides state:-

$$\frac{\partial}{\partial \underline{\Lambda}} \langle \underline{\psi}(\theta | \underline{\theta}_c) \rangle = -\frac{N}{2} \underline{\Psi}^{-1} \frac{\partial}{\partial \underline{\Lambda}} \langle \underline{\Sigma} \rangle$$

we will insert Jordan's (2003) derivation here:-

$$- Q(\underline{\Lambda} | \theta^{(t)}) = -\frac{1}{2} \sum_{n=1}^N \text{tr} \{ (\underline{y}_n \underline{y}_n^T - \underline{y}_n \langle \underline{x}_n \rangle \underline{x}_n^T - \underline{A} \langle \underline{x}_n \rangle \underline{y}_n^T + \underline{A} \langle \underline{x}_n \underline{x}_n^T \rangle \underline{A}^T) \underline{\Psi}^{-1} \}$$

- some dimensionality verif:-

$$\underline{y}_n \in \mathbb{R}^q \quad \langle \underline{x}_n \rangle = \underbrace{(\underline{I} + \underline{A}^T \underline{\Psi}^{-1} \underline{A})^{-1}}_{(p \times p)} \underbrace{\underline{A}^T \underline{\Psi}^{-1} (\underline{y} - \underline{\mu})}_{(p \times 1)} \text{ so } \langle \underline{x}_n \rangle \in \mathbb{R}^p$$

as expected  
(cond. Gaussian)

$$\underline{\Psi} \in \mathbb{R}^{q \times q} \quad \langle \underline{x}_n \underline{x}_n^T \rangle = (\underline{I} + \underline{A}^T \underline{\Psi}^{-1} \underline{A})^{-1} \text{ so } \langle \underline{x}_n \underline{x}_n^T \rangle \in \mathbb{R}^{p \times p}$$

as expected ✓

Hence;

$$\langle \Sigma \rangle = \frac{1}{N} \left( \sum_{n=1}^N y_n y_n^\top - \Omega^{(t+1)} \langle x_n \rangle y_n^\top \right)$$

and as  $\Psi^{(t+1)} = \text{diag} \langle \Sigma \rangle$  ;

$$\Psi^{(t+1)} = \frac{1}{N} \text{diag} \left\{ \sum_{n=1}^N y_n y_n^\top - \Omega^{(t+1)} \sum_{n=1}^N \langle x_n \rangle y_n^\top \right\}$$

completing the M-step for  $\Psi$ .



Q10-Review

(also 1/2 S. 2019 v11)

- clear up doubts, ambiguities.

- Kalman filtering (Jordan 2003) ch15 15.4

- model: SSM

- inference algo: KF

(\*) inference problem for SSM is exactly the same as it was for HMMs.

- i.e. calculate posterior probability of latent states given output sequence(\*\*) we wish to calculate  $p(x_t | y_0, y_1, \dots, y_t)$  (lecture slides use  $p(x_t | y_1, \dots, y_t)$ )(\*\*) inference n SSM  $\rightarrow$  find a recursion linking these conditional means- d variances at neighbouring moments in time.

(\*\*) AS Gaussians; charact. by mean, cov.

(\*) important notation (A3):-

- indicate condit. mean/covariances with emphasis on part. output sequence  
being conditioned on-  $\hat{x}_{t|t}$  - mean of  $x_t$  conditioned on partial sequence  $y_0, \dots, y_t$ -  $P_{t|t}$  - covariance matrix of  $x_t$ 

$$\hat{x}_{t|t} := E[x_t | y_0, \dots, y_t]$$

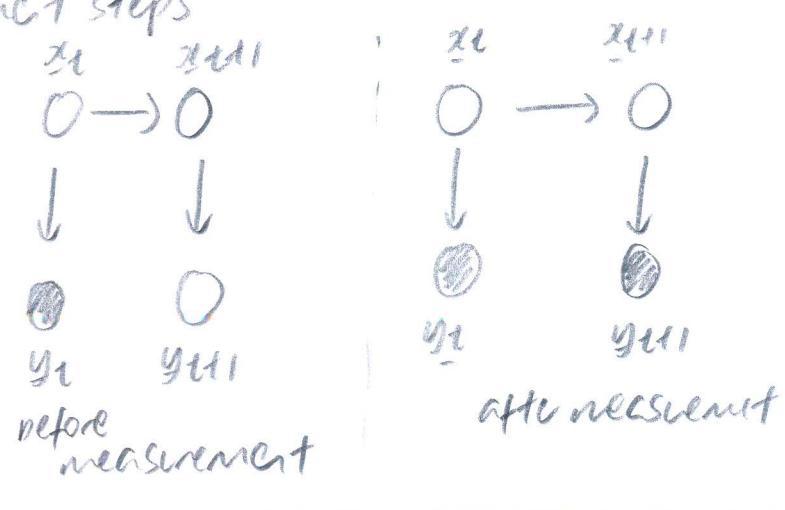
$$P_{t|t} := E[(x_t - \hat{x}_{t|t})(x_t - \hat{x}_{t|t})^\top | y_0, \dots, y_t]$$

- also: we will require distri of  $x_t$  conditioned on  $y_0, \dots, y_{t-1}$  (meas/obs)  
i.e.  $\hat{x}_{t|t-1}$  and  $P_{t|t-1}$ 

(\*) Clarity on time update, measurement steps

- LHS: conditioned on  $y_0, \dots, y_t$ 

- Assume already calc.

 $p(x_t | y_0, \dots, y_t)$  (i.e.  $\hat{x}_{t|t}, P_{t|t}$ )- we want to carry this distri  
forward into frequent on RHS  
(conditioning on  $y_0, \dots, y_{t+1}$ )

(\*) Decompose transformation :-

Time update  $p(x_t | y_0, \dots, y_t) \rightarrow p(x_{t+1} | y_0, \dots, y_t)$

Measurement update  $p(x_{t+1} | y_0, \dots, y_t) \rightarrow p(x_{t+1} | y_0, \dots, y_{t+1})$

Time update: propagate distribution forward in time; evaluating new mean and covariance ( $\hat{x}_{t+1|t}$ ,  $P_{t+1|t}$ ) from old ✓  
mean and cov( $\hat{x}_{t|t}$ ,  $P_{t|t}$ ); but with no new measurements

Measurement update: incorporate new measurement  $y_{t+1}$ , update p.d. for  $x_{t+1}$

- overall result: transformation from  $\hat{x}_{t|t}$  and  $P_{t|t}$   $\rightarrow \hat{x}_{t+1|t+1}$  and  $P_{t+1|t+1}$

(\*) Time update equations:- (dynamical model)

- SSM equations:-

$$x_{t+1} = Ax_t + Gw_t \quad w_t \sim N(0, Q) \quad (i) \quad (15.1)$$

$$y_t = Cx_t + v_t \quad v_t \sim N(0, R) \quad (ii) \quad (15.2)$$

- note:-  $y_t | x_t \sim N(Cx_t, R)$  and  $x_{t+1} | x_t \sim N(Ax_t, GQG^T)$

(\*) take conditional expectations on both sides of (i)

-  $w_t$  is independent of conditioning variables  $y_0, \dots, y_t$   
(process noise is independent of observations)

$$\hat{x}_{t+1|t} = E[(Ax_t + Gw_t) | y_0, \dots, y_t] = A\hat{x}_{t|t} + G\underbrace{E[w_t | y_0:t]}_{=0}$$

$$\Rightarrow \hat{x}_{t+1|t} = A\hat{x}_{t|t} \quad (15.10)$$

(\*) take condit. covariance of both sides of (ii)

-  $\hat{x}_{t|t}$  - constant in condit. distri

-  $w_t$  - zero mean

-  $w_t$  and  $x_t$  are independent (process noise and state are independent)

$$E[x_t w_t] = E[x_t] E[w_t]$$

$$P_{t+1|t} = E[(x_{t+1} - \hat{x}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})^T | y_{0:t}]$$

subs. (15.9), (15.10)

$$= E[(Ax_t + Gw_t - A\hat{x}_{t|t})(Ax_t + Gw_t - A\hat{x}_{t|t})^T | y_{0:t}]$$

expanding and collecting terms:-

$$\begin{aligned} P_{t+1|t} &= E[Ax_t x_t^T A^T - Ax_t \hat{x}_{t|t}^T A^T - A\hat{x}_{t|t} x_t^T A^T + A\hat{x}_{t|t} \hat{x}_{t|t}^T A^T | y_{0:t}] \\ &\quad + G E[w_t w_t^T | y_{0:t}] G^T \\ &\quad + A E[x_t w_t^T | y_{0:t}] G^T + G E[w_t x_t^T | y_{0:t}] A^T - G E[w_t | y_{0:t}] \hat{x}_{t|t}^T A^T \\ &\quad - A \hat{x}_{t|t} E[w_t^T | y_{0:t}] G^T \end{aligned}$$

note: (i) :-  $E[x_t w_t^T | y_{0:t}] = E[x_t | y_{0:t}] \underbrace{E[w_t^T | y_{0:t}]}_{=0} = 0$   
via independence

(ii) :-  $E[w_t | y_{0:t}] = 0$

$$\Rightarrow P_{t+1|t} = A E[(x_t - \hat{x}_{t|t})(x_t - \hat{x}_{t|t})^T | y_{0:t}] A^T + \underbrace{G E[w_t w_t^T | y_{0:t}] G^T}_{=Q \text{ as } w_t \text{ has 0 mean}} \\ = A P_{t|t} A^T + G Q G^T \quad (15.13)$$

- 
- (\*) we have established the conditional distn of  $x_{t+1}$  given  $y_0, \dots, y_t$  }
  - (\*) now we establish conditional distn of  $y_{t+1}$  " "
  - (\*) use this to establish joint conditional of  $x_{t+1}, y_{t+1}$  " "
- 

- SSM equations:-  $y_{t+1} = (x_{t+1} + v_{t+1}) \quad v_{t+1} \sim N(0, R)$   
(observation model)

- Similar to previous; take conditional expectations wrt  $y_0, \dots, y_t$

$$\begin{aligned} \bullet E[y_{t+1}|y_0, \dots, y_t] &= E[(x_{t+1} + v_{t+1})|y_0, \dots, y_t] \\ &= (E[x_{t+1}|y_0:t] + E[v_{t+1}]) \\ &\stackrel{\sim}{=} 0 \\ &= \hat{x}_{t+1|t} \end{aligned}$$

$$\bullet E[(y_{t+1} - \hat{y}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})^T | y_0:t]$$

$$= E[((x_{t+1} + v_{t+1} - \hat{y}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})^T | y_0:t)] \quad \text{- suppressing conditioning}\\ \text{- brevity}$$

$$\begin{aligned} &= (E[x_{t+1}x_{t+1}^T] - (E[x_{t+1}\hat{x}_{t+1|t}^T] + E[v_{t+1}\hat{x}_{t+1|t}^T] - E[v_{t+1}\hat{x}_{t+1|t}^T] \\ &\quad - E[\hat{y}_{t+1|t}x_{t+1}^T] + E[\hat{y}_{t+1|t}\hat{x}_{t+1|t}^T]) \end{aligned}$$

?

(O/S)

$$\bullet E[(y_{t+1} - \hat{y}_{t+1|t})(y_{t+1} - \hat{y}_{t+1|t})^T | y_0, \dots, y_t] \quad \text{- suppressing condit.}$$

$$= E[(x_{t+1} + v_{t+1} - \hat{x}_{t+1|t})(x_{t+1} + v_{t+1} - \hat{x}_{t+1|t})^T]$$

$$= E[(x_{t+1}x_{t+1}^T - x_{t+1}\hat{x}_{t+1|t}^T - \hat{x}_{t+1|t}x_{t+1}^T + \hat{x}_{t+1|t}\hat{x}_{t+1|t}^T]$$

$$+ E[v_{t+1}v_{t+1}^T]$$

$$\left. \quad + C E[x_{t+1}v_{t+1}^T] + E[v_{t+1}x_{t+1}^T]C^T - E[v_{t+1}\hat{x}_{t+1|t}^T]C^T - C E[\hat{x}_{t+1|t}v_{t+1}^T] \right\}$$

all terms (see before)

$$= C \mathbb{E}[(x_{t+1} - \hat{x}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})^T] C^T + R$$

$$= P_{t+1|t} C^T + R$$

(\*) Joint distribution (cond. on past outputs  $y_0:t$ )

$$p(x_{t+1}, y_{t+1} | y_0, \dots, y_t) = N(\mu_{t+1}, V_{t+1})$$

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

(\*) Similar to FA;

Use arrow → compute condit. distri. of  $x_{t+1}$  given  $y_{t+1}$  (i.e. find posterior)  
when  $x_{t+1}$  and  $y_{t+1}$  have a joint Gaussian distri.

(\*) Difference: joint distri. is itself also a conditional.

$$(*) \text{Conditional distri.:- (measurement update)} \quad \mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (y - \mu_2)$$

$$p(x_{t+1}|y_0, \dots, y_{t+1}) = N(x_{t+1} | \mu_{1|2}, V_{1|2}) \quad V_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$

$$\hat{x}_{t+1|t+1} = \mu_{1|2} = P_{t+1|t} C^T (C P_{t+1|t} C^T + R)^{-1} (y_{t+1} - \hat{c}_{t+1|t})$$

$$P_{t+1|t+1} = V_{1|2} = P_{t+1|t} - P_{t+1|t} C^T (C P_{t+1|t} C^T + R)^{-1} C P_{t+1|t}$$

(\*) Don't be put off by dense symbology → a straightforward application of condit. Gaussians.

(\*) Kalman gain matrix

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

(\*) Filtering equations:

At time  $t$ , assume we have the mean estimate  $\hat{x}_{t|t}$  and covariance est  $P_{t|t}$ .

- using this; we recursively calculate  $\hat{x}_{\text{fullt}}$  and  $P_{\text{fullt}}$  :-

time update :-  $\hat{x}_{\text{fullt}} = A\hat{x}_{\text{elt}}$

$$P_{\text{fullt}} = AP_{\text{elt}}A^T + GQG^T$$

measured update :-  $\hat{x}_{\text{fullt+1}} = \hat{x}_{\text{fullt}} + K_{\text{tt}}(y_{\text{tt}} - G\hat{x}_{\text{fullt}})$

$$P_{\text{fullt+1}} = P_{\text{fullt}} - K_{\text{tt}}P_{\text{fullt}}G^T \quad (\text{from slides})$$

(?) Q15.2: lecture slides state that  $K_t$  can be recomputed, as independent of data. Referring to?

- slides use different notation for  $K_t$  and  $K_{\text{tt}}$ ; while in Jordan, they are the same expression.
- what am I missing here?

(?) Omitted review from Jordan:-

- i) VMS map.
- ii) information filter
- iii) RTS smoothing

(?) Following things to cement understanding:-

- i) probabilistic story (via appl. of Gaussians)
- ii) graphical models story
- iii) physical model e.g. airplane position.
- iv) other primitive anchors.