10-708 11/06/2020

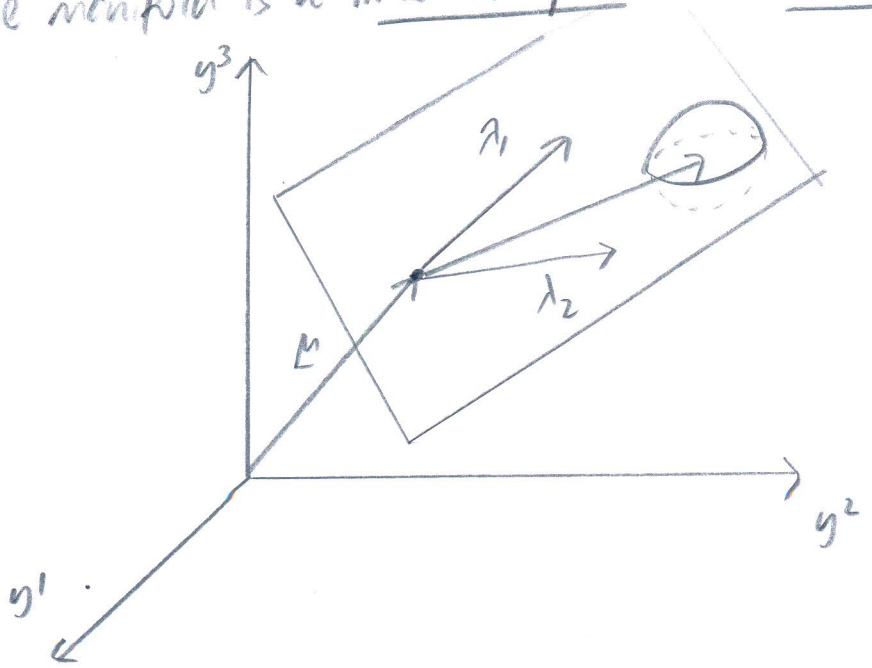
L10 - Review

(#) Focus on
- Resolving queries in notes
- FA derivation
- EM for FA
- KF derivation (set aside RTS smoothing)

- using Jordan (2003) ch 14 + 15

(#) Factor Analysis
- Motivation → 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 centred on a point.
- coordinates of point → components of latent random vector
- If we wish to parameterise cont. manifold; latent variable is cont. random vector.

(#) Factor Analysis (mathem. intuition)
- Assume manifold is a linear subspace → factor analysis



- observed data:
  - in a p-dim. subspace M in $\mathbb{R}^q$ where $p<q$.
  - Given a set of basis vectors $\{\lambda_j\}$, a point in M can be represented as a product $\underline{\Lambda}\underline{x}$
  - $\underline{x}$ - co-ordinate vector
  - $\underline{\Lambda}$ - columns are basis vectors (factor loading mat.)

$$\begin{bmatrix} | & | \\ \underline{\lambda}_1 & \underline{\lambda}_2 \ \dots \\ | & | \end{bmatrix}$$

- Treat co-ordinate vectors as values of continuous r vec $\underline{X}$
- Treat errors $\underline{\epsilon}$ 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 distr centred about that point. (OBS) - not entirely clear
- Resulting density $\longrightarrow$
  - convolution of Gaussian density on manifold M with Gaussian distr 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 preoccup. of density estimation

---

- PCA $\longrightarrow$ FA
  K-means $\rightarrow$ GMM.

(*) Model details
- Supplementary slides and slide notes from Jordan (1003)

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

- Dimensionality lightening:-

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

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

(*) Note $\Psi \in \mathbb{R}^{q \times q}$ is diagonal

(and that in gen. case of $p(\underline{x}) = N(\mu_0, \Sigma_0)$ and $\Psi = \sigma^2 \underline{I} \rightarrow$ PPCA)

Generative model:

- Generate a point within manifold M by sampling $\underline{x} \sim N(0, \underline{I})$ → $\underline{\mu} + \Lambda \underline{x}$
- Add noise to get $\underline{y} = \underline{\mu} + \Lambda \underline{x} + \underline{w}$ by sampling $\underline{w} \sim N(0, \Psi)$

(*) FA: (Gaussian results; Gaussian prop + M.I.L)

- Specify $p(\underline{x})$ and $p(\underline{y}|\underline{x})$ → derive joint $p(\underline{x}, \underline{y})$ → derive $p(\underline{x}|\underline{y})$
- We can use $p(\underline{x}|\underline{y})$ for inference, or as a means to parameter estimation.

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

⇒ Compute $\underline{\mu}_Y = \mathbb{E}[\underline{Y}]$ and $\underline{\Sigma}_{YY} = \mathrm{Var}[\underline{Y}]$

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

(A4) :- For 100% clarity :-

- $\Lambda \mathbb{E}[\underline{x}\underline{w}^T]$ and $\mathbb{E}[\underline{w}\underline{x}^T]\Lambda^T$ 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 ⇒ expectation of product is product of expectations

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

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

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

(A5) / $\underline{\Sigma}_{XY} = \mathrm{cov}(\underline{x}, \underline{y}) = \mathbb{E}[(X - \mu_X)(\underline{Y} - \underline{\mu}_Y)^T]$

(A6)

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

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

$$= \mathbb{E}[\underline{X}\underline{X}^T\Lambda^T + \underline{X}\underline{W}^T] = \underbrace{\mathbb{E}[\underline{X}\underline{X}^T]}_{=I}\Lambda^T + \mathbb{E}[\underline{X}]\underbrace{\mathbb{E}[\underline{W}^T]}_{=0} = \Lambda^T$$

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

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

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

- $\mathbb{E}[Y] = \mathbb{E}[\underline{\mu} + \underline{\Lambda}\underline{X}] = \underline{\mu} + \underline{\Lambda}\mathbb{E}[\underline{X}]$ $\quad \mathbb{E}[\underline{Y}|\underline{X}] = \underline{\mu} + \underline{\Lambda}\underline{X}$

$= \underline{\mu}$

- $\mathrm{Var}[Y] = \mathrm{Var}(\underline{\mu} + \underline{\Lambda}\underline{X}) + \mathbb{E}[\underline{\Psi}]$

$= \mathrm{Var}(\underline{\Lambda}\underline{X}) + \underline{\Psi} = \mathbb{E}[(\underline{\Lambda}\underline{X} - \underbrace{\mathbb{E}[\underline{\Lambda}\underline{X}]}_{=0})(\underline{\Lambda}\underline{X} - \underbrace{\mathbb{E}[\underline{\Lambda}\underline{X}]}_{=0})^T]$

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

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

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

- Yielding some results (via diff means).

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

- Just remember (via a):- $\underline{X} \sim N(\underline{0}, \underline{I})$

$\underline{Y} \sim N(\underline{\mu} + \underline{\Lambda}\underline{z}, \underline{\Psi})$ i.e. $\underline{Y}|\underline{x} \sim N(\underline{\mu} + \underline{\Lambda}\underline{x}, \underline{\Psi})$

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

$\mathbb{E}[\underline{X}|\underline{y}] = \underline{\Lambda}^T(\underline{\Lambda}\underline{\Lambda}^T + \underline{\Psi})^{-1}(\underline{y} - \underline{\mu})$ (i)

$= (\underline{I} + \underline{\Lambda}^T\underline{\Psi}^{-1}\underline{\Lambda})^{-1}\underline{\Lambda}^T\underline{\Psi}^{-1}(\underline{y} - \underline{\mu})$ (ii) (via MIL) (14.24).

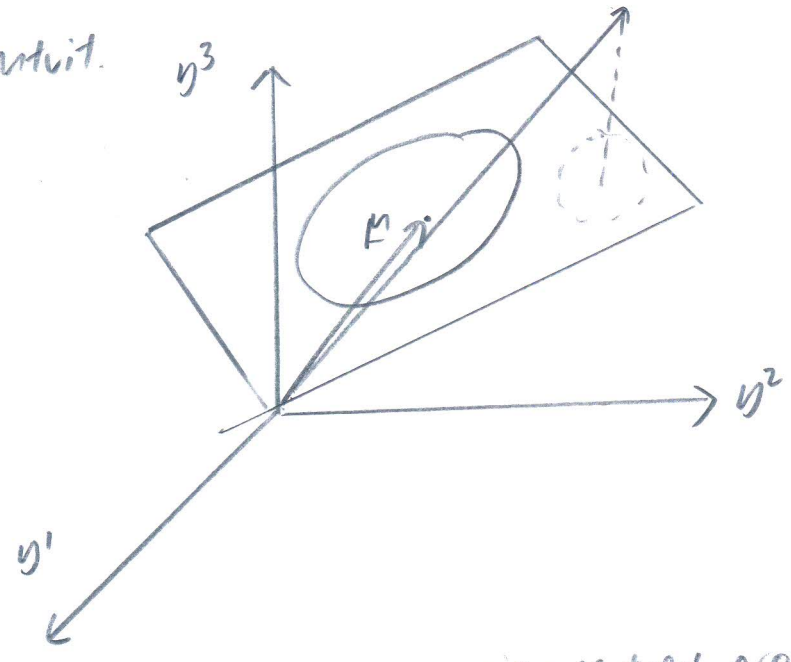
(i) inv of $(q \times q)$ matrix

(ii) inv of $(p \times p)$ matrix.

$$\text{var}(\underline{x}|\underline{y}) = \underline{I} - \underline{\Lambda}^T(\underline{\Lambda}\underline{\Lambda}^T + \underline{\Psi})^{-1}\underline{\Lambda}$$

$$= (\underline{I} + \underline{\Lambda}^T\underline{\Psi}^{-1}\underline{\Lambda})^{-1} \quad \text{(via MIL)} \quad (14.26)$$

(*) Geometric intuit.



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

(6/5/1): Don't fully understand geometric intuitions 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

(?) -- likelihood remains invariant to orthogonal transformations of $\underline{\Lambda}$:- (i.e. $\underline{R}^T\underline{R} = \underline{R}\underline{R}^T = \underline{I}$)

- Showing this :- $\underline{R}$ -orthogonal mat. $\tilde{\underline{\Lambda}} = \underline{\Lambda}\underline{R}$

$$\tilde{\underline{\Lambda}}\tilde{\underline{\Lambda}}^T = \underline{\Lambda}\underline{R}(\underline{\Lambda}\underline{R})^T = \underline{\Lambda}\underline{R}\underline{R}^T\underline{\Lambda}^T = \underline{\Lambda}\underline{\Lambda}^T$$

⇒ likelihood, which depends on Λ only through product $\Lambda\Lambda^T$ is <u>not changed</u> if Λ is postmultiplied by an orthog mat. $\underline{R}$.

(iii) - nonlinear coupling: parameters Λ and Ψ are coupled by determinant and inverse.

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

(*) Incomplete data log-like.:- (using marg density $p(y)$)

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

where $D = \{y_n : n=1,\ldots,N\}$ (iid)

$$- \hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} y_n \quad (14.29)$$

- omit references to mean μ from here on i.e. estimate μ via (14.29); use these centred variables where we subtract from data vectors y_n; use these centred variables when we specify data vectors y_n.

(*) Complete data log-likelihood:-

- complete data = $\{(x_n, y_n); n=1,\ldots,N\}$ - D_c (*) - note decoupling of Λ and Ψ
- product of joint Gaussians

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

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

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

- Rewrite quadratic forms using trace, using cyclic prop

$$= -\frac{N}{2}\log|\Psi| - \frac{1}{2}\underbrace{\sum_{n=1}^{N}tr(x_nx_n^T)}_{c} - \frac{1}{2}\sum_{n=1}^{N}tr\left[\underbrace{(y_n-\Lambda x_n)(y-\Lambda x_n)^T}_{S}\Psi^{-1}\right] \quad (*)$$

(15.2)

$$= -\frac{N}{2}\log|\Psi| - \frac{N}{2}tr(S\Psi^{-1})$$

c : 0 (due to mean $x = 0$?)

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

E.M for FA

E. step 14.2.3

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

$$\langle \underline{S}\rangle = \frac{1}{N}\sum_{n=1}^{N}\left\langle (y_n - \underline{\Lambda}\underline{x}_n)(y_n - \underline{\Lambda}\underline{x}_n)^T\right\rangle$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left\langle (y_n y_n^T - y_n \underline{x}_n^T \underline{\Lambda}^T - \underline{\Lambda}\underline{x}_n y_n^T + \underline{\Lambda}\underline{x}_n\underline{x}_n^T\underline{\Lambda})\right\rangle$$

- Note $\underline{x}_n$ is r.v (or treated as)

$$= \frac{1}{N}\sum_{n=1}^{N} y_n y_n^T - y_n\langle \underline{x}_n^T\rangle \underline{\Lambda}^T - \underline{\Lambda}\langle \underline{x}_n\rangle y_n^T + \underline{\Lambda}\langle \underline{x}_n\underline{x}_n^T\rangle\underline{\Lambda}$$

- Note that this merits use of C.E. $\langle x_n\rangle$ and $\langle \underline{x}_n\underline{x}_n\rangle$
- (*) expected sufficient statistics required for E-step are C.E. $\langle \underline{x}_n\rangle$ and $\langle \underline{x}_n\underline{x}_n^T\rangle$
- (*) obtained from our derivation of $p(\underline{x}|y)$ and (mean and var of this distn.)

$$\langle x_n\rangle = \mathbb{E}[x_n|y_n]$$

$$\langle \underline{x}_n\underline{x}_n^T\rangle = \mathrm{var}(x_n|y_n) + \mathbb{E}[x_n|y_n]\mathbb{E}[x_n|y_n]^T$$

(015 3) (?) - what form are you not clocking?

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

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

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

($\underline{\tilde{S}}$: contains $\underline{\Lambda}$)

$$= -\frac{N}{2}\underline{\Psi}^{-1}\frac{\partial}{\partial \underline{\Lambda}}\langle \underline{S}\rangle \quad (*)$$

(0/S2 ?)

$$= -\frac{N}{2}\underline{\Psi}^{-1}\frac{\partial}{\partial \underline{\Lambda}}\left(\frac{1}{N}\sum_{n=1}^{N}\left(\underline{y}_n\underline{y}_n^T - \underline{y}_n\langle \underline{x}_n^T\rangle \underline{\Lambda}^T - \underline{\Lambda}\langle \underline{x}_n\rangle \underline{y}_n^T + \underline{\Lambda}\langle \underline{x}_n\underline{x}_n^T\rangle \underline{\Lambda}^T\right)\right)$$

$$= -\frac{N}{2}\underline{\Psi}^{-1}\frac{\partial}{\partial \underline{\Lambda}}\left(\frac{1}{N}\sum_{n=1}^{N}\underline{y}_n\underline{y}_n^T - 2\underline{y}_n\langle \underline{x}_n^T\rangle \underline{\Lambda}^T + \underline{\Lambda}\langle \underline{x}_n\underline{x}_n^T\rangle \underline{\Lambda}^T\right)$$

$$= -\frac{N}{2}\underline{\Psi}^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} -2\underline{y}_n\langle \underline{x}_n^T\rangle + 2\underline{\Lambda}\langle \underline{x}_n\underline{x}_n^T\rangle\right)$$

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

$$\Rightarrow \sum_{n=1}^{N}\underline{\Psi}^{-1}\underline{y}_n\langle \underline{x}_n^T\rangle - \sum_{n=1}^{N}\underline{\Psi}^{-1}\underline{\Lambda}\langle \underline{x}_n\underline{x}_n^T\rangle = \underline{0}$$

$$\Rightarrow \sum_{n=1}^{N}\underline{\Psi}^{-1}\underline{y}_n\langle \underline{x}_n^T\rangle = \sum_{n=1}^{N}\underline{\Psi}^{-1}\underline{\Lambda}\langle \underline{x}_n\underline{x}_n^T\rangle \qquad \text{pre mult. by } \underline{\Psi}$$

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

(*) these are normal equat. from LR!

- For details on why it is not exactly equal (i.e. transpose) → see Jordan (2003).
- But let's ignore it

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

- noting :-

$$\frac{\partial}{\partial \underline{\Lambda}} tr\left[y_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^T \underline{\Psi}^{-1}\right] = \frac{\partial}{\partial \underline{\Lambda}} tr\left[\underline{\Psi}^{-1} y_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^T\right] = \frac{\partial}{\partial \underline{\Lambda}} tr\left[\left(\underline{\Psi}^{-1} y_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^T\right)^T\right]$$

as $tr(ABC) = tr(CAB) = tr(BCA)$ and $tr(A^T) = tr(A)$

$$= \frac{\partial}{\partial \underline{\Lambda}} tr\left[\underbrace{\left(\underline{\Psi}^{-1} y_n \langle \underline{x}_n^T \rangle\right)^T}_{\underline{B}} \underbrace{\underline{\Lambda}}_{\underline{A}}\right]$$

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

$$\Rightarrow \frac{\partial}{\partial \underline{\Lambda}} tr\left[y_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^T \underline{\Psi}^{-1}\right] = \underline{\Psi}^{-1} y_n \langle \underline{x}_n^T \rangle$$

- AND;

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

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

where B, C are symmetric ;

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

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

↘ - concluding M-step deriv. (Jordan 2003)

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

$$\langle \underline{S} \rangle = \frac{1}{N} \sum_{n=1}^{N} \left(y_n y_n^T - y_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^{(t+1)T} - \underline{\Lambda}^{(t+1)} \langle \underline{x}_n \rangle y_n^T + \underline{\Lambda}^{(t+1)} \langle \underline{x}_n \underline{x}_n^T \rangle \underline{\Lambda}^{(t+1)T} \right)$$

$$= \frac{1}{N} \left(\Sigma y_n y_n^T - \Sigma y_n \langle \underline{x}_n^T \rangle \underline{\Lambda}^{(t+1)T} - \underline{\Lambda}^{(t+1)} \Sigma \langle \underline{x}_n \rangle y_n^T + \underline{\Lambda}^{(t+1)} \Sigma \langle \underline{x}_n \underline{x}_n^T \rangle \underline{\Lambda}^{(t+1)T} \right)$$

derivation is tedious to write; note that

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

- not sure about this next part:- ?
- $\underline{\Psi}$ is a diagonal matrix
- to calculate derivative of $Q(\underline{\Psi}|\underline{\theta}^{(t)})$ wrt $\underline{\Psi}$, take usual matrix derivative, but retaining only diagonal terms.
- take derivative wrt $\underline{\Psi}^{-1}$, set to 0, retaining only diag.:- yields:- (O/S 1) (?)

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

(*) From slides of lecture:- (EX S. 2019 L10)

$$\frac{\partial}{\partial \underline{\Psi}^{-1}} \langle l_c(\underline{\theta}|\mathcal{D}_c) \rangle = \frac{\partial}{\partial \underline{\Psi}^{-1}} \left(-\frac{N}{2}\log|\underline{\Psi}| - \frac{1}{2}\sum_{n=1}^{N} \text{tr}\left[\langle \underline{x}_n \underline{x}_n^T \rangle\right] - \frac{N}{2}\text{tr}\left[\langle \underline{S} \rangle \underline{\Psi}^{-1}\right] \right)$$

$$= \frac{N}{2}\underline{\Psi} - \frac{N}{2}\langle \underline{S} \rangle$$

$$\Rightarrow \quad \underline{\Psi}^{(t+1)} = \langle \underline{S} \rangle \qquad (?)$$

(?): lecture slides and Jordan (2003) don't match → you need to figure out which is correct. — continued →

(*) Another issue which requires resolution is in previous (O/S 2) (?):- (digression)

i.e. where EX lecture slides state:-

$$\frac{\partial}{\partial \underline{\Lambda}} \langle l_c(\underline{\theta}|\mathcal{D}_c) \rangle = -\frac{N}{2}\underline{\Psi}^{-1}\frac{\partial}{\partial \underline{\Lambda}}\langle \underline{S} \rangle$$

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

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

- some dimensionality verif:-

$y_n \in \mathbb{R}^q$, $\underline{\Lambda} \in \mathbb{R}^{q \times p}$, $\underline{\Psi} \in \mathbb{R}^{q \times q}$

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

(as expected (cond. Gaussian)

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

(as expected) ✓

Hence;

$$\langle \underline{\underline{S}} \rangle = \frac{1}{N}\left(\sum_{n=1}^{N} y_n y_n^T - \underline{\underline{\Lambda}}^{(t+1)} \langle x_n \rangle y_n^T \right)$$

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

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

- completing the M-step for $\underline{\underline{\Psi}}$.