EM Algorithm and JFA Theorems Proofs

Mengke HU

ASPITRG Group, ECE Department
Drexel University

mengke.hu@gmail.com

February 24, 2013

Outline

Maximum Likelihood Estimator

$$\Theta^* = \arg\max_{\Theta} \ln P(\mathbf{Y}|\Theta) \tag{1}$$

- $\mathbf{Y}$ is observations
- Θ is parameter set that we need to estimate
- **Problem:** Unobservable random variable $\mathbf{X}$ lay behind $\mathbf{Y}$
- **Solution:** Expectation-Maximization (EM) Algorithm

Incomplete data and Complete data

- 2 sample space $\mathcal{X}$ and $\mathcal{Y}$,
  Unobservable data $\mathbf{X} \in \mathcal{X}$ and observable data $\mathbf{Y} \in \mathcal{Y}$
- Many-to-one mapping $y : \mathcal{X} \to \mathcal{Y}$, i.e. $\mathbf{Y} = y(\mathbf{X})$
- *Incomplete data*: $\mathbf{Y}$
- *Complete data*: $\mathbf{X}$

Iterative Algorithm

- Vector parameter $\Theta_{1\times r} \in \Omega$, Ω is an $r-$dimensional convex set
- We have: $\Theta^{(0)} \to \Theta^{(1)} \to \Theta^{(2)} \ldots$
- $\Theta^{(p)} \in \Omega$: current value of Θ after p cycles of iterative algorithm
- Iterative Algorithm: a rule applicable to any starting point on the list, s.t. $\mathbf{M} : \Theta^{(p)} \to \Theta^{(p+1)}, \Theta^{(p)} \in \Omega, \Theta^{(p+1)} \in \Omega$

$$\Theta^{(p+1)} = \mathbf{M}(\Theta^{(p)})$$

Outline

EM algorithm

EM: exponential family case

- **Sufficient Statistics** $T(X_1 \cdots X_n)$**:**
  $P(X_1 \cdots X_n | T(X_1 \cdots X_n), \Theta)$ is independent of Θ
- $p(\mathbf{X}, \mathbf{Y}|\Theta)$ is element in the regular exponential-family:

$$p(\mathbf{X}, \mathbf{Y}|\Theta) = p(\mathbf{X}, y(\mathbf{X})|\Theta) = b(\mathbf{X}) \frac{\exp(\Theta T(\mathbf{X})^T)}{a(\Theta)}$$

- By the fact $\int \int p(\mathbf{X}, \mathbf{Y}|\Theta) d\mathbf{X} d\mathbf{Y} = 1$

$$a(\Theta) = \int_{\mathcal{X}} b(\mathbf{X}) \exp(\Theta T(\mathbf{X})^T) d\mathbf{X}$$

EM algorithm

EM: exponential family case

1. **E-step:** Estimate the complete-data sufficient statistics $T(\mathbf{X})$

$$T^{(p)}(\mathbf{X}) = \mathbb{E}[T(\mathbf{X})|\mathbf{Y}, \Theta^{(p)}]$$

2. **M-step:** $\Theta^{(p+1)}$ (maximize the likelihood) as solution of the following equations:

$$\mathbb{E}[t(\mathbf{X})|\Theta] = T^{(p)}(\mathbf{X})$$

EM algorithm brief proof

EM: exponential family case

likelihood

$$\begin{aligned}
\mathcal{L}(\mathbf{Y}|\Theta) &= \ln p(\mathbf{Y}|\Theta) = \ln \frac{p(\mathbf{X},\mathbf{Y}|\Theta)}{p(\mathbf{X}|\mathbf{Y},\Theta)} \\
&= \ln p(\mathbf{X},\mathbf{Y}|\Theta) - \ln p(\mathbf{X}|\mathbf{Y},\Theta) \\
&= \ln b(\mathbf{X}) \frac{\exp(\Theta T(\mathbf{X})^T)}{a(\Theta)} - \ln b(\mathbf{X}) \frac{\exp(\Theta T(\mathbf{X})^T)}{a(\Theta|\mathbf{Y})} \\
&= -\ln a(\Theta) + \ln a(\Theta|\mathbf{Y})
\end{aligned}$$

EM algorithm brief proof

EM: exponential family case

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\mathbf{Y}|\Theta)}{\partial \Theta} &= \frac{\partial(-\ln a(\Theta) + \ln a(\Theta|\mathbf{Y})))}{\partial \Theta} \\
&= -\frac{1}{a(\Theta)}\frac{\partial a(\Theta)}{\partial \Theta} + \frac{1}{a(\Theta|\mathbf{Y})}\frac{\partial a(\Theta|\mathbf{Y})}{\partial \Theta} \\
&= \frac{1}{a(\Theta|\mathbf{Y})}\frac{\partial}{\partial \Theta}\left(\int_{\mathcal{X}(y)} b(\mathbf{X}) \exp(\Theta T(\mathbf{X})^T) d\mathbf{X}\right) \\
&- \frac{1}{a(\Theta)}\frac{\partial}{\partial \Theta}\left(\int_{\mathcal{X}} b(\mathbf{X}) \exp(\Theta T(\mathbf{X})^T) d\mathbf{X}\right) \\
&= \mathbb{E}[T(\mathbf{X})|\mathbf{Y}, \Theta] - \mathbb{E}(T(\mathbf{X})|\Theta) = 0
\end{aligned}
$$

EM algorithm

EM: exponential family case revise

$$\frac{\partial \mathcal{L}(\mathbf{Y}|\Theta)}{\partial \Theta} = \mathbb{E}[T(\mathbf{X})|\mathbf{Y}, \Theta] - \mathbb{E}(T(\mathbf{X})|\Theta) = 0$$

1. **E-step:** Estimate the complete-data sufficient statistics $T(\mathbf{X})$

$$T^{(p)}(\mathbf{X}) = \mathbb{E}[T(\mathbf{X})|\mathbf{Y}, \Theta^{(p)}]$$

2. **M-step:** $\Theta^{(p+1)}$ as solution of the following equations:

$$\mathbb{E}[t(\mathbf{X})|\Theta] = T^{(p)}(\mathbf{X})$$

EM algorithm

Problem needed to be shown

- **Monotone none-decreasing:**
  Why $\mathcal{L}(\mathbf{Y}|\Theta^{(p+1)}) = \mathcal{L}(\mathbf{Y}|\mathbf{M}(\Theta^{(p)})) \geq \mathcal{L}(\mathbf{Y}|\Theta^{(p)})$?
- **Bounded:** Why $\mathcal{L}(\mathbf{Y}|(\Theta^*)^{(p+1)}) = \mathcal{L}(\mathbf{Y}|(\Theta^*)^{(p)})$
- **Note** 1: Monotone + Bounded $\Rightarrow$ Limit point exists
- **Note** 2: Let $E = \left\{\mathcal{L}(\mathbf{Y}|\Theta^{(p)})\right\}_p$ and $\mathcal{L}(\mathbf{Y}|\Theta^*)$ is limit point of E.
  $\mathcal{L}(\mathbf{Y}|\Theta^*)$ is not necessary in E
- We will prove this 2 problem in general EM form

Outline

General EM algorithm (GEM)

General formula

1. **E-step**:
$$\mathcal{Q}(\Theta|\Theta^{(p)}) = \mathbb{E}_{p(\mathbf{X}|\mathbf{Y},\Theta^{(p)})}[\ln p(\mathbf{X},\mathbf{Y}|\Theta)]$$

2. **M-step**:
$$\Theta^{(p+1)} = \arg\max_{\Theta} \mathcal{Q}(\Theta|\Theta^{(p)})$$

Outline

GEM

Proof of likelihood none-decreasing

We need to show:

$$\mathcal{L}(\mathbf{Y}|\Theta^{(p+1)}) \geq \mathcal{L}(\mathbf{Y}|\Theta^{(p)})$$

Observe:

$$\begin{aligned}
&\mathcal{Q}(\Theta^{(p+1)}|\Theta^{(p)}) = \mathbb{E}_{p(\mathbf{X}|\mathbf{Y},\Theta^{(p)})}[\ln p(\mathbf{X},\mathbf{Y}|\Theta^{(p+1)})] \\
&= \mathbb{E}_{p(\mathbf{X}|\mathbf{Y},\Theta^{(p)})}[\ln p(\mathbf{X}|\mathbf{Y},\Theta^{(p+1)})] + \mathbb{E}_{p(\mathbf{X}|\mathbf{Y},\Theta^{(p)})}[\ln p(\mathbf{Y}|\Theta^{(p+1)})] \\
&= \mathcal{H}(\Theta^{(p+1)}|\Theta^{(p)}) + \mathcal{L}(\mathbf{Y}|\Theta^{(p+1)})
\end{aligned}$$

Similarly

$$\mathcal{Q}(\Theta^{(p)}|\Theta^{(p)}) = \mathcal{H}(\Theta^{(p)}|\Theta^{(p)}) + \mathcal{L}(\mathbf{Y}|\Theta^{(p)})$$

GEM

Proof of likelihood none-decreasing

From M-step:

$$\mathcal{Q}(\Theta^{(p)}|\Theta^{(p)}) \leq \mathcal{Q}(\Theta^{(p+1)}|\Theta^{(p)})$$

Also we can easily prove

$$\mathcal{H}(\Theta^{(p+1)}|\Theta^{(p)}) = -\mathcal{D}(p(\mathbf{X}|\mathbf{Y},\Theta^{(p)})||p(\mathbf{X}|\mathbf{Y},\Theta^{(p+1)})) + \mathcal{H}(\Theta^{(p)}|\Theta^{(p)})$$

Therefore:

$$\begin{aligned}\Delta\mathcal{L} &= \mathcal{L}(\mathbf{Y}|\Theta^{(p+1)}) - \mathcal{L}(\mathbf{Y}|\Theta^{(p)}) \\ &= \mathcal{D}(p(\mathbf{X}|\mathbf{Y},\Theta^{(p)})||p(\mathbf{X}|\mathbf{Y},\Theta^{(p+1)})) \\ &\geq 0\end{aligned}$$

Outline

GEM

Proof of Boundedness

Theorem

Suppose $\exists \Theta^*$ *s.t.* $\mathcal{L}(\mathbf{Y}|\Theta^*) \geq \mathcal{L}(\mathbf{Y}|\Theta), \ \ \forall \Theta \in \Omega$, *then*

$$\mathcal{Q}((\Theta^*)^{(p+1)}|(\Theta^*)^{(p)}) = \mathcal{Q}((\Theta^*)^{(p)}|(\Theta^*)^{(p)}) \tag{2}$$

$$\mathcal{L}(\mathbf{Y}|(\Theta^*)^{(p+1)}) = \mathcal{L}(\mathbf{Y}|(\Theta^*)^{(p)}) \tag{3}$$

- (2) is obviously by M-step expression
- (3) holds, since $\mathcal{D}(p(\mathbf{X}|\mathbf{Y}, (\Theta^*)^{(p)})||p(\mathbf{X}|\mathbf{Y}, (\Theta^*)^{(p+1)})) = 0$

GEM

Another View

Observe:

$$\mathcal{D}(p(\mathbf{X},\mathbf{Y}|\Theta^{(p)})||p(\mathbf{X},\mathbf{Y}|\Theta^{(p+1)})) =$$

$$\iint p(\mathbf{X},\mathbf{Y}|\Theta^{(p)}) \ln \frac{p(\mathbf{X},\mathbf{Y}|\Theta^{(p)})}{p(\mathbf{X},\mathbf{Y}|\Theta^{(p+1)})} d\mathbf{X} d\mathbf{Y}$$

$$= \iint p(\mathbf{Y}|\Theta^{(p)}) p(\mathbf{X}|\mathbf{Y},\Theta^{(p)}) \left(\ln p(\mathbf{X},\mathbf{Y}|\Theta^{(p)}) - \ln p(\mathbf{X},\mathbf{Y}|\Theta^{(p+1)}) \right) d\mathbf{X} d\mathbf{Y}$$

$$= \int p(\mathbf{Y}|\Theta^{(p)}) \left(\mathbb{E}_{p(\mathbf{X}|\mathbf{Y},\Theta^{(p)})} [\ln p(\mathbf{X},\mathbf{Y}|\Theta^{(p)}) - \ln p(\mathbf{X},\mathbf{Y}|\Theta^{(p+1)})] \right) d\mathbf{Y}$$

$$= \int p(\mathbf{Y}|\Theta^{(p)}) (\mathcal{Q}(\Theta^{(p)}|\Theta^{(p)}) - \mathcal{Q}(\Theta^{(p+1)}|\Theta^{(p)})) d\mathbf{Y}$$

$$\geq 0$$

1. **E-step**: We need to calculate the posterior distribution to obtain the expectation:

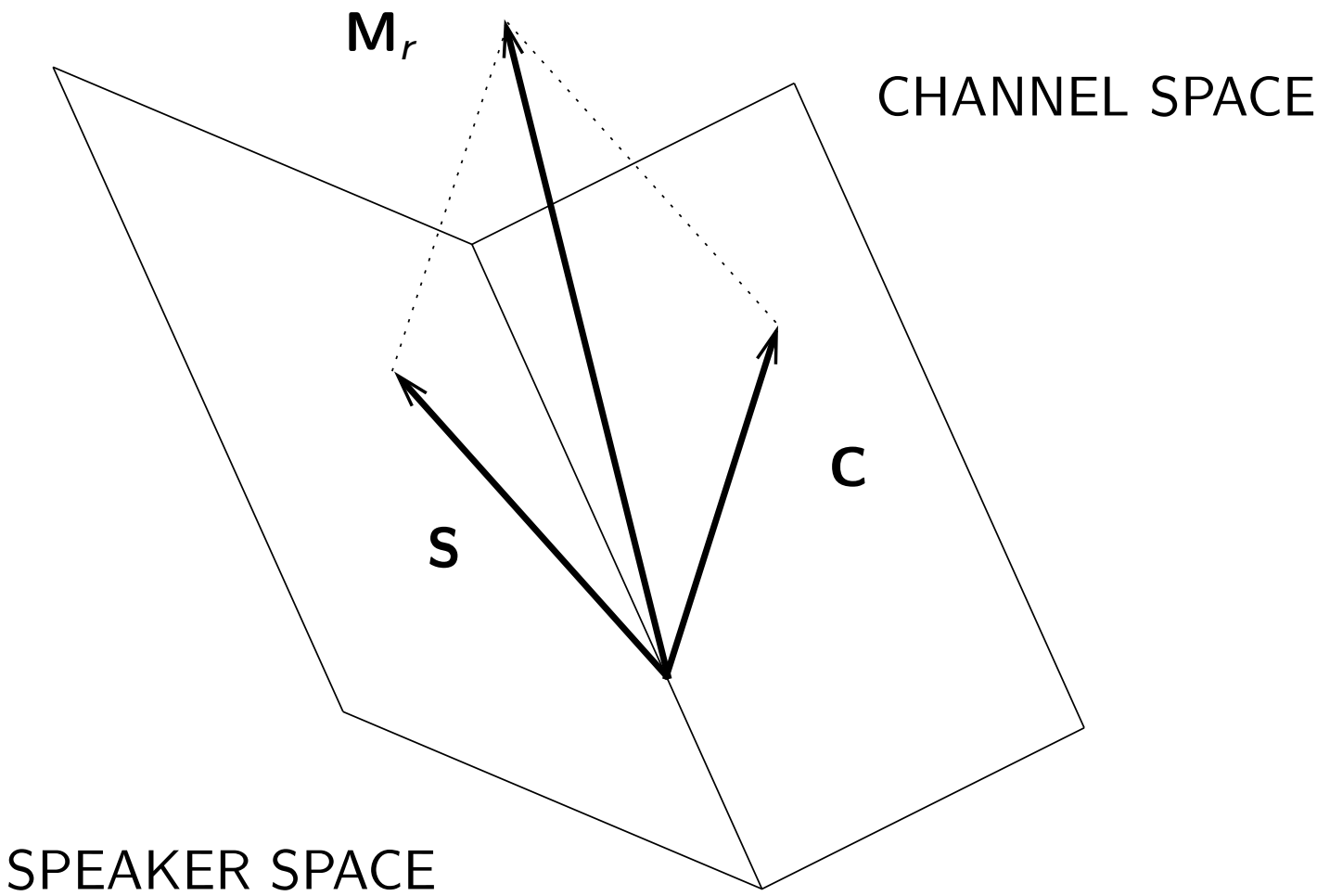
$$\mathcal{Q}(\Theta|\Theta^{(p)}) = \mathbb{E}_{p(\mathbf{X}|\mathbf{Y},\Theta^{(p)})}[\ln p(\mathbf{X},\mathbf{Y}|\Theta)]$$

2. **M-step**:

$$\Theta^{(p+1)} = \arg\max_{\Theta} \mathcal{Q}(\Theta|\Theta^{(p)})$$

Outline

JFA model Review

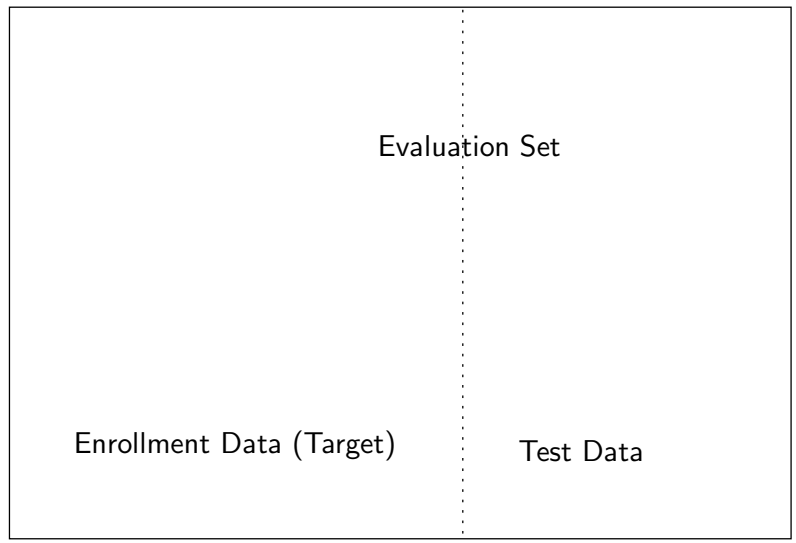


JFA model Review

$$\mathbf{Y}_{t,c,r,s} = \mathbf{M}_{c,r,s} + \mathbf{W}_{t,c,r,s}$$
$$\mathbf{M}_{c,r,s} = \mathbf{M}_{c,s} + \mathbf{u}_{c,s}\mathbf{x}_{r,s}$$
$$\mathbf{M}_{c,s} = \mathbf{m}_c + \mathbf{v}_{c,s}\mathbf{y}_s + \mathbf{d}_{c,s}\mathbf{z}_{c,s}$$

$(t, c, r, s) = (\mathit{frameID}, \mathit{componentID}, \mathit{recordingID}, \mathit{speakerID})$ $\mathbf{Y}$ is observation, i.e. features.
$\mathbf{Y}_{t,c,r,s}, \mathbf{W}_{t,c,r,s}$ are IID $\mathcal{N}(\mathbf{m}_c, \Sigma_c)$
$\mathbf{m}_c, \Sigma_c$

JFA Baseline Dataset



JFA Baseline Procedure

1. Train UBM (Training set)
2. Alignment (Determine which mixture component that each MFCC belongs to.)
3. Train JFA world model (mainly on Training set, re-fix on Enrollment Set)
4. Adapt JFA target model (Enrollment Set)
5. Testing (Test Data)

Outline

JFA model latent variable representation

The model can be written as:

$$\mathbf{Y}_{t,c,r,s} = \mathbf{m}_c + \mathbf{v}_{c,s}\mathbf{y}_s + \mathbf{d}_{c,s}\mathbf{z}_{c,s} + \mathbf{u}_{c,s}\mathbf{x}_{r,s} + \mathbf{W}_{t,c,r,s}$$

By Kronecker product:

$$\mathbb{1}_{rt\times 1} \otimes \begin{pmatrix} \vdots \\ (\mathbf{m}_{c,s})_{F\times 1} \\ \vdots \end{pmatrix}_{CF\times 1} = (\mathbf{m}_s)_{CFrt\times 1}$$

JFA model latent variable representation

Then we have

$$\mathbf{Y}_s = \mathbf{m}_s + \mathbf{v}_s\mathbf{y}_s + \mathbf{d}_s\mathbf{z}_s + \mathbf{u}_s\mathbf{x}_{r,s}$$

That is

$$\mathbf{Y}_s = \mathbf{m}_s + \mathbf{V}_x\mathbf{X} \tag{4}$$

where

$$\mathbf{V} = \begin{pmatrix} \mathbf{u}_s & 0 & 0 & \mathbf{v}_s & \mathbf{d}_s \\ 0 & \ddots & 0 & \vdots & \vdots \\ 0 & 0 & \mathbf{u}_s & \mathbf{v}_s & \mathbf{d}_s \end{pmatrix}_{rCF \times (rR_c + R_s + CF)}$$

and

$$\mathbf{X}_s = \begin{pmatrix} \vdots \\ \mathbf{x}_{r,s} \\ \vdots \\ \mathbf{y}_s \\ \mathbf{z}_s \end{pmatrix}$$

JFA: E-step

Joint distribution: distribution of complete data

$$\mathbf{Y}_s = \mathbf{m}_s + \mathbf{V}_x \mathbf{X} \tag{5}$$

Distribution of complete data is Gaussian (in exponential family)

$$(\mathbf{Y}, \mathbf{X}) \sim \mathcal{N}\left(\begin{bmatrix} m_Y \\ m_X \end{bmatrix}; \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} \right)$$

$$\Sigma_{YY} = \mathbf{V}^T \mathbf{V} + \begin{pmatrix} \Sigma_c & 0 & 0 \\ 0 & \Sigma_c & 0 \\ 0 & 0 & \Sigma_c \end{pmatrix}$$

$$\Sigma_{YX} = \Sigma_{XY}^T = \mathbf{V}$$

JFA: E-step

Incomplete data distribution and Posterior distribution

Incomplete data distribution

$$(\mathbf{Y}|\mathbf{X}) \sim \mathcal{N}(m_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - m_X);\ \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})$$

Posterior Distribution

$$(\mathbf{X}|\mathbf{Y}) \sim \mathcal{N}(m_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - m_Y);\ \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX})$$

JFA: M-step

Parameter set:

$$\Lambda = (\mathbf{m}, \mathbf{v}, \mathbf{u}, \mathbf{d}, \Sigma)$$

Overall likelihood on the training set:

$$\mathcal{L}(\Lambda|\mathbf{Y}) = \sum_s \ln p_\Lambda(\mathbf{Y}_s)$$

JFA: M-step

Theorem

Jensen's Inequality: *Function f is convex, then* $f(\int xp(s)dx) \leq \int f(x)p(x)dx$ *Note:* ln *is concave*

$$\begin{aligned}
\max_{\Lambda} \mathcal{L}(\Lambda|\mathbf{Y}) &\Longleftrightarrow \max_{\Lambda}(\mathcal{L}(\Lambda|\mathbf{Y}) - \mathcal{L}(\Lambda_0|\mathbf{Y})) \\
&= \max(\sum_s \ln p_\Lambda(\mathbf{Y}) - \sum_s \ln p_{\Lambda_0}(\mathbf{Y})) \\
&= \max(\sum_s \ln \int \frac{p_\Lambda(\mathbf{X},\mathbf{Y})}{p_{\Lambda_0}(\mathbf{X},\mathbf{Y})} p_{\Lambda_0}(\mathbf{X}|\mathbf{Y})d\mathbf{X}) \\
&\geq \max(\sum_s \int \ln \frac{p_\Lambda(\mathbf{X},\mathbf{Y})}{p_{\Lambda_0}(\mathbf{X},\mathbf{Y})} p_{\Lambda_0}(\mathbf{X}|\mathbf{Y})d\mathbf{X}) \\
&= \max(\mathcal{Q}(\Lambda|\Lambda_0) - \mathcal{Q}(\Lambda_0|\Lambda_0)) \\
&\Longleftrightarrow \max \mathcal{Q}(\Lambda|\Lambda_0)
\end{aligned}$$

JFA: M-step

$$\max\Big(\sum_s \int \ln \frac{p_\Lambda(\mathbf{X},\mathbf{Y})}{p_{\Lambda_0}(\mathbf{X},\mathbf{Y})} p_{\Lambda_0}(\mathbf{X}|\mathbf{Y}) d\mathbf{X}\Big)$$

$$= \max\Big(\sum_s \int \ln \frac{p_\Lambda(\mathbf{Y}|\mathbf{X})\mathcal{N}(\mathbf{X};\mathbf{0},\mathbf{I})}{p_{\Lambda_0}(\mathbf{Y}|\mathbf{X})\mathcal{N}(\mathbf{X};\mathbf{0},\mathbf{I})} p_{\Lambda_0}(\mathbf{X}|\mathbf{Y}) d\mathbf{X}\Big)$$

$$= \max\Big(\sum_s \int \ln \frac{p_\Lambda(\mathbf{Y}|\mathbf{X})}{p_{\Lambda_0}(\mathbf{Y}|\mathbf{X})} p_{\Lambda_0}(\mathbf{X}|\mathbf{Y}) d\mathbf{X}\Big)$$

$$= \max\Big(\sum_s \int \ln p_\Lambda(\mathbf{Y}|\mathbf{X}) p_{\Lambda_0}(\mathbf{X}|\mathbf{Y}) d\mathbf{X} - \sum_s \int \ln p_{\Lambda_0}(\mathbf{Y}|\mathbf{X}) p_{\Lambda_0}(\mathbf{X}|\mathbf{Y}) d\mathbf{X}\Big)$$

$$\Longleftrightarrow \max \sum_s \int \ln p_\Lambda(\mathbf{Y}|\mathbf{X}) p_{\Lambda_0}(\mathbf{X}|\mathbf{Y}) d\mathbf{X}$$

JFA: M-step

$$\max \sum_s \int \ln p_\Lambda(\mathbf{Y}|\mathbf{X}) p_{\Lambda_0}(\mathbf{X}|\mathbf{Y}) d\mathbf{X}$$
$$= \max \sum_s \mathbb{E}_{p_{\Lambda_0}(\mathbf{Y}|\mathbf{X})}[\ln p_\Lambda(\mathbf{Y}|\mathbf{X})]$$
$$= \max \sum_s \mathcal{A}_\Lambda(s)$$

Outline

Min Divergence: M-step

Theorem

Let $\lambda = \{\mu_y, \mu_z, \mathbf{K}_{xx}, \mathbf{K}_{yy}, \mathbf{K}_{zz}, \Sigma\}$, *s.t.* $\mathbf{y}'_s \sim \mathcal{N}(\mu_y; \mathbf{K}_{yy})$, $\mathbf{x}'_{r,s} \sim \mathcal{N}(\mathbf{0}; \mathbf{K}_{xx})$ $\mathbf{z}'_s \sim \mathcal{N}(\mu_z; \mathbf{K}_{zz})$

$$\mathbf{M}_{r.s} = \mathbf{m} + \mathbf{v}\mathbf{y}_s + \mathbf{u}\mathbf{x}_{r,s} + \mathbf{d}\mathbf{z}_s = \mathbf{v}\mathbf{y}'_s + \mathbf{u}\mathbf{x}'_{r,s} + \mathbf{d}\mathbf{z}'_s$$

We can find Λ *fix the observation data better by choosing* λ *s.t.*

$$\min \sum_s \mathcal{D}(p_{\Lambda_0}(\mathbf{X}_s|\mathbf{Y}_s)||p_\lambda(\mathbf{X}_s)) \tag{6}$$

$$\max \sum_s \mathbb{E}[\ln p_\Sigma(\mathbf{Y}_s|\mathbf{M}_{r,s})] \tag{7}$$

Min Divergence: M-step

Proof

$$\max \sum_s \ln p_\Lambda(\mathbf{Y}_s) \Longleftrightarrow \max \sum_s \ln \frac{p_\Lambda(\mathbf{Y}_s)}{p_{\Lambda_0}(\mathbf{Y}_s)}$$

Need to prove the following theorem first:

Theorem

$$\ln \frac{p_\Lambda(\mathbf{Y}_s)}{p_{\Lambda_0}(\mathbf{Y}_s)} \geq -\mathbb{E}[\ln p_{\Sigma_0}(\mathbf{Y}_s|\mathbf{M})] + \mathbb{E}[\ln p_\Sigma(\mathbf{Y}_s|\mathbf{M})] + \mathcal{D}(p_{\lambda_0}(\mathbf{M}|\mathbf{Y}_s)||p_{\lambda_0}(\mathbf{M})) - \mathcal{D}(p_{\lambda_0}(\mathbf{M}|\mathbf{Y}_s)||p_\lambda(\mathbf{M}))$$

Min Divergence: M-step

Proof

$$
\begin{aligned}
&- \mathbb{E}[\ln p_{\Sigma_0}(\mathbf{Y}_s|\mathbf{M})] + \mathbb{E}[\ln p_{\Sigma}(\mathbf{Y}_s|\mathbf{M})] \\
&+ \mathcal{D}(p_{\lambda_0}(\mathbf{M}|\mathbf{Y}_s)||p_{\lambda_0}(\mathbf{M})) - \mathcal{D}(p_{\lambda_0}(\mathbf{M}|\mathbf{Y}_s)||p_{\lambda}(\mathbf{M})) \\
&= \int \ln \left(\frac{p_{\Sigma}(\mathbf{Y}_s) p(\mathbf{M}|\lambda)}{p_{\Sigma_0}(\mathbf{Y}_s) p(\mathbf{M}|\lambda_0)} \right) p(\mathbf{M}|\mathbf{Y}, \lambda_0) d\mathbf{M} \\
&\leq \ln \int \left(\frac{p_{\Sigma}(\mathbf{Y}_s) p(\mathbf{M}|\lambda)}{p_{\Sigma_0}(\mathbf{Y}_s) p(\mathbf{M}|\lambda_0)} \right) p(\mathbf{M}|\mathbf{Y}, \lambda_0) d\mathbf{M} \\
&= \ln \frac{p_{\Lambda}(\mathbf{Y}_s)}{p_{\Lambda_0}(\mathbf{Y})}
\end{aligned}
$$

Min Divergence: M-step

Proof

Also

$$\mathcal{D}(p_{\Lambda_0}(\mathbf{X}|\mathbf{Y}_s)||p_\lambda(\mathbf{X})) =$$

$$\mathcal{D}(p_{\lambda_0}(\mathbf{M}|\mathbf{Y})||p_\lambda(\mathbf{M})) + \int \mathcal{D}(p_{\lambda_0}(\mathbf{X}|\mathbf{M})||p_\lambda(\mathbf{X}|\mathbf{M}))p_{\lambda_0}(\mathbf{M}|\mathbf{Y})d\mathbf{M}$$

Min Divergence: M-step

Update Parameters

$$\mathbf{m} = \mathbf{m}_0 + \mathbf{v}_0\mu_y + \mathbf{d}_0\mu_z \tag{8}$$

$$\mathbf{u} = \mathbf{u}_0\mathbf{K}_{xx}^{\frac{1}{2}} \tag{9}$$

$$\mathbf{v} = \mathbf{v}_0\mathbf{K}_{yy}^{\frac{1}{2}} \tag{10}$$

$$\mathbf{d} = \mathbf{d}_0\mathbf{K}_{zz}^{\frac{1}{2}} \tag{11}$$

$$\Sigma = \mathbf{N}^{-1}\sum_s\sum_r\left(\mathbf{S}_r(\mathbf{m}_0) - 2diag(\mathbf{F}_r\mathbb{E}[\mathbf{O}_r^T]) + diag(\mathbb{E}[\mathbf{O}_r\mathbf{O}_r^T]\mathbf{N}_r)\right) \tag{12}$$

Min Divergence: M-step

$$\mu_y = \frac{1}{S}\sum_s \mathbb{E}[\mathbf{y}_s] \tag{13}$$

$$\mu_z = \frac{1}{S}\sum_s \mathbb{E}[\mathbf{z}_s] \tag{14}$$

$$\mathbf{K}_{xx} = \frac{1}{r}\sum_s\sum_r \mathbb{E}[\mathbf{x}_{r,s}\mathbf{x}_{r,s}^T] \tag{15}$$

$$\mathbf{K}_{yy} = \frac{1}{S}\sum_s \mathbb{E}[\mathbf{y}_s\mathbf{y}_s^T] - \mu_y\mu_y^* \tag{16}$$

$$\mathbf{K}_{zz} = diag\left(\frac{1}{S}\sum_s \mathbb{E}[\mathbf{z}_s\mathbf{z}_s^*] - \mu_z\mu_z^T\right) \tag{17}$$

$$\mathbf{N} = \sum_s\sum_r \mathbf{N}_{r,s} \tag{18}$$

$$r = \sum_s r(s) \tag{19}$$