Week 3: The EM algorithm

Maneesh Sahani
maneesh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit
University College London

Term 1, Autumn 2005
Mixtures of Gaussians

Data: \( \mathcal{Y} = \{y_1 \ldots y_N\} \)

Latent process:
\( s_i \overset{iid}{\sim} \text{Discrete}[\pi] \)

Component distributions:
\( y_i \mid (s_i = m) \sim \mathcal{P}_m[\theta_m] = \mathcal{N}[\mu_m; \Sigma_m] \)

Marginal distribution:
\[
P(y_i) = \sum_{m=1}^{k} \pi_m P_m(y; \theta_m)
\]

Log-likelihood:
\[
\log p(\mathcal{Y} \mid \{\mu_m\}, \{\Sigma_m\}, \pi) = \sum_{i=1}^{n} \log \sum_{m=1}^{k} \pi_m \left| 2\pi \Sigma_m \right|^{-1/2} \exp \left[ -\frac{1}{2} (y_i - \mu_m)^T \Sigma_m^{-1} (y_i - \mu_m) \right]
\]
EM for MoGs

- Evaluate responsibilities
  \[ r_{im} = \frac{P_m(y)\pi_m}{\sum_{m'} P_{m'}(y)\pi_{m'}} \]

- Update parameters
  \[ \mu_m \leftarrow \frac{\sum_i r_{im}y_i}{\sum_i r_{im}} \]
  \[ \Sigma_m \leftarrow \frac{\sum_i r_{im}(y_i - \mu_m)(y_i - \mu_m)^T}{\sum_i r_{im}} \]
  \[ \pi_m \leftarrow \frac{\sum_i r_{im}}{N} \]
The Expectation Maximisation (EM) algorithm

The EM algorithm finds a (local) maximum of a latent variable model likelihood. It starts from arbitrary values of the parameters, and iterates two steps:

**E step:** Fill in values of latent variables according to posterior given data.

**M step:** Maximise likelihood as if latent variables were not hidden.

- Useful in models where learning would be easy if hidden variables were, in fact, observed (e.g. MoGs).
- Decomposes difficult problems into series of tractable steps.
- No learning rate.
- Framework lends itself to principled approximations.
Jensen’s Inequality

For $\alpha_i \geq 0$, $\sum \alpha_i = 1$ and any $\{x_i > 0\}$

$$
\log \left( \sum_i \alpha_i x_i \right) \geq \sum_i \alpha_i \log(x_i)
$$

Equality if and only if $\alpha_i = 1$ for some $i$ (and therefore all others are 0).
The Free Energy for a Latent Variable Model

Observed data $\mathcal{Y} = \{y_i\}$; Latent variables $\mathcal{X} = \{x_i\}$; Parameters $\theta$.

**Goal:** Maximize the log likelihood (i.e. ML learning) wrt $\theta$:

$$\mathcal{L}(\theta) = \log P(\mathcal{Y}|\theta) = \log \int P(\mathcal{X}, \mathcal{Y}|\theta) d\mathcal{X},$$

Any distribution, $q(\mathcal{X})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen’s inequality:

$$\mathcal{L}(\theta) = \log \int q(\mathcal{X}) \frac{P(\mathcal{X}, \mathcal{Y}|\theta)}{q(\mathcal{X})} d\mathcal{X} \geq \int q(\mathcal{X}) \log \frac{P(\mathcal{X}, \mathcal{Y}|\theta)}{q(\mathcal{X})} d\mathcal{X} \overset{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,

$$\int q(\mathcal{X}) \log \frac{P(\mathcal{X}, \mathcal{Y}|\theta)}{q(\mathcal{X})} d\mathcal{X} = \int q(\mathcal{X}) \log P(\mathcal{X}, \mathcal{Y}|\theta) d\mathcal{X} - \int q(\mathcal{X}) \log q(\mathcal{X}) d\mathcal{X}$$

$$= \int q(\mathcal{X}) \log P(\mathcal{X}, \mathcal{Y}|\theta) d\mathcal{X} + \mathcal{H}[q],$$

where $\mathcal{H}[q]$ is the entropy of $q(\mathcal{X})$. So:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{X}, \mathcal{Y}|\theta) \rangle_{q(\mathcal{X})} + \mathcal{H}[q]$$
The E and M steps of EM

The lower bound on the log likelihood is given by:

$$\mathcal{F}(q, \theta) = \langle \log P(X, Y|\theta) \rangle_{q(X)} + H[q],$$

EM alternates between:

**E step:** optimize $\mathcal{F}(q, \theta)$ wrt distribution over hidden variables holding parameters fixed:

$$q^{(k)}(X) := \arg\max_q \mathcal{F}(q(X), \theta^{(k-1)}).$$

**M step:** maximize $\mathcal{F}(q, \theta)$ wrt parameters holding hidden distribution fixed:

$$\theta^{(k)} := \arg\max_\theta \mathcal{F}(q^{(k)}(X), \theta) = \arg\max_\theta \langle \log P(X, Y|\theta) \rangle_{q^{(k)}(X)}.$$

The second equality comes from the fact that the entropy of $q(X)$ does not depend directly on $\theta$. 
EM as Coordinate Ascent in $\mathcal{F}$
The E Step

The free energy can be re-written

\[
\mathcal{F}(q, \theta) = \int q(\mathcal{X}) \log \frac{P(\mathcal{X}, \mathcal{Y} | \theta)}{q(\mathcal{X})} \, d\mathcal{X}
\]

\[
= \int q(\mathcal{X}) \log \frac{P(\mathcal{X} | \mathcal{Y}, \theta) P(\mathcal{Y} | \theta)}{q(\mathcal{X})} \, d\mathcal{X}
\]

\[
= \int q(\mathcal{X}) \log P(\mathcal{Y} | \theta) \, d\mathcal{X} + \int q(\mathcal{X}) \log \frac{P(\mathcal{X} | \mathcal{Y}, \theta)}{q(\mathcal{X})} \, d\mathcal{X}
\]

\[
= \mathcal{L}(\theta) - \text{KL}[q(\mathcal{X}) \| P(\mathcal{X} | \mathcal{Y}, \theta)]
\]

The second term is the Kullback-Leibler divergence.

This means that, for fixed \( \theta \), \( \mathcal{F} \) is bounded above by \( \mathcal{L} \), and achieves that bound when \( \text{KL}[q(\mathcal{X}) \| P(\mathcal{X} | \mathcal{Y}, \theta)] = 0 \).

But \( \text{KL}[q \| p] \) is zero if and only if \( q = p \).

So, the E step simply sets

\[
q^{(k)}(\mathcal{X}) = P(\mathcal{X} | \mathcal{Y}, \theta^{(k-1)})
\]

and, after an E step, the free energy equals the likelihood.
The KL$[q(x) \| p(x)]$ is non-negative and zero iff $\forall x : p(x) = q(x)$

First let’s consider discrete distributions; the Kullback-Liebler divergence is:

$$\text{KL}[q \| p] = \sum_i q_i \log \frac{q_i}{p_i}.$$ 

To find the distribution $q$ which minimizes $\text{KL}[q \| p]$ we add a Lagrange multiplier to enforce the normalization constraint:

$$E \overset{\text{def}}{=} \text{KL}[q \| p] + \lambda (1 - \sum_i q_i) = \sum_i q_i \log \frac{q_i}{p_i} + \lambda (1 - \sum_i q_i)$$

We then take partial derivatives and set to zero:

$$\frac{\partial E}{\partial q_i} = \log q_i - \log p_i + 1 - \lambda = 0 \Rightarrow q_i = p_i \exp(\lambda - 1)$$
$$\frac{\partial E}{\partial \lambda} = 1 - \sum_i q_i = 0 \Rightarrow \sum_i q_i = 1$$

$$\Rightarrow q_i = p_i.$$
Why $\text{KL}[q\|p]$ is non-negative and zero iff $p(x) = q(x)$ 

Check that the curvature (Hessian) is positive (definite), corresponding to a minimum:

$$\frac{\partial^2 E}{\partial q_i \partial q_i} = \frac{1}{q_i} > 0, \quad \frac{\partial^2 E}{\partial q_i \partial q_j} = 0,$$

showing that $q_i = p_i$ is a genuine minimum.

At the minimum is it easily verified that $\text{KL}[p\|p] = 0$.

A similar proof holds for $\text{KL}[\cdot\|\cdot]$ between continuous densities, the derivatives being substituted by functional derivatives.
EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

\[
\mathcal{L}(\theta^{(k-1)}) = \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \leq \mathcal{F}(q^{(k)}, \theta^{(k)}) \leq \mathcal{L}(\theta^{(k)}),
\]

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt \( \theta \).
- \( \mathcal{F} \leq \mathcal{L} \) by Jensen – or, equivalently, from the non-negativity of KL

If the M-step is executed so that \( \theta^{(k)} \neq \theta^{(k-1)} \) iff \( \mathcal{F} \) increases, then the overall EM iteration will step to a new value of \( \theta \) iff the likelihood increases.
Fixed Points of EM are Stationary Points in $\mathcal{L}$

Let a fixed point of EM occur with parameter $\theta^*$. Then:

$$\frac{\partial}{\partial \theta} \left[ \langle \log P(\mathcal{X}, \mathcal{Y} | \theta) \rangle_{P(\mathcal{X}|\mathcal{Y},\theta^*)} \right]_{\theta^*} = 0$$

Now,

$$\mathcal{L}(\theta) = \log P(\mathcal{Y}|\theta) = \langle \log P(\mathcal{Y}|\theta) \rangle_{P(\mathcal{X}|\mathcal{Y},\theta^*)}$$

$$= \left\langle \log \frac{P(\mathcal{X}, \mathcal{Y}|\theta)}{P(\mathcal{X}|\mathcal{Y}, \theta)} \right\rangle_{P(\mathcal{X}|\mathcal{Y},\theta^*)}$$

$$= \langle \log P(\mathcal{X}, \mathcal{Y}|\theta) \rangle_{P(\mathcal{X}|\mathcal{Y},\theta^*)} - \langle \log P(\mathcal{X}|\mathcal{Y}, \theta) \rangle_{P(\mathcal{X}|\mathcal{Y},\theta^*)}$$

so,

$$\frac{d}{d\theta} \mathcal{L}(\theta) = \frac{d}{d\theta} \langle \log P(\mathcal{X}, \mathcal{Y}|\theta) \rangle_{P(\mathcal{X}|\mathcal{Y},\theta^*)} - \frac{d}{d\theta} \langle \log P(\mathcal{X}|\mathcal{Y}, \theta) \rangle_{P(\mathcal{X}|\mathcal{Y},\theta^*)}$$

The second term is 0 at $\theta^*$ if the derivative exists (minimum of $\text{KL}[\cdot || \cdot]$), and thus:

$$\frac{d}{d\theta} \mathcal{L}(\theta) \bigg|_{\theta^*} = \frac{d}{d\theta} \langle \log P(\mathcal{X}, \mathcal{Y}|\theta) \rangle_{P(\mathcal{X}|\mathcal{Y},\theta^*)} \bigg|_{\theta^*} = 0$$

So, EM converges to a stationary point of $\mathcal{L}(\theta)$. 
Maxima in $F$ correspond to maxima in $L$

Let $\theta^*$ now be the parameter value at a local maximum of $F$ (and thus at a fixed point).

Differentiating the previous expression wrt $\theta$ again we find

$$
\frac{d^2}{d\theta^2} L(\theta) = \frac{d^2}{d\theta^2} \langle \log P(X, Y|\theta) \rangle_{P(X|Y, \theta^*)} - \frac{d^2}{d\theta^2} \langle \log P(X|Y, \theta) \rangle_{P(X|Y, \theta^*)}
$$

The first term on the right is negative (a maximum) and the second term is positive (a minimum). Thus the curvature of the likelihood is negative and

$$
\theta^* \text{ is a maximum of } L.
$$
The Gaussian mixture model (E-step)

In a univariate Gaussian mixture model, the density of a data point $y$ is:

$$p(y|\theta) = \sum_{m=1}^{k} p(s = m|\theta)p(y|s = m, \theta) \propto \sum_{m=1}^{k} \frac{\pi_m}{\sigma_m} \exp \left\{ - \frac{1}{2\sigma_m^2} (y - \mu_m)^2 \right\},$$

where $\theta$ is the collection of parameters: means $\mu_m$, variances $\sigma_m^2$ and mixing proportions $\pi_m = p(s = m|\theta)$.

The hidden variable $s_i$ indicates which component observation $y_i$ belongs to. The E-step computes the posterior for $s_i$ given the current parameters:

$$q(s_i) = p(s_i|y_i, \theta) \propto p(y_i|s_i, \theta)p(s_i|\theta)$$

$$r_{im} \overset{\text{def}}{=} q(s_i = m) \propto \frac{\pi_m}{\sigma_m} \exp \left\{ - \frac{1}{2\sigma_m^2} (y_i - \mu_m)^2 \right\} \quad \text{(responsibilities)}$$

with the normalization such that $\sum_m r_{im} = 1$. 

The Gaussian mixture model (M-step)

In the M-step we optimize the sum (since $s$ is discrete):

$$E = \langle \log p(y, s|\theta) \rangle_{q(s)} = \sum q(s) \log[p(s|\theta) p(y|s, \theta)]$$

$$= \sum_{i,m} r_{im} \left[ \log \pi_m - \log \sigma_m - \frac{1}{2\sigma_m^2}(y_i - \mu_m)^2 \right].$$

Optimization is done by setting the partial derivatives of $E$ to zero:

$$\frac{\partial E}{\partial \mu_m} = \sum_i r_{im} \frac{(y_i - \mu_m)}{2\sigma_m^2} = 0 \Rightarrow \mu_m = \frac{\sum_i r_{im} y_i}{\sum_i r_{im}},$$

$$\frac{\partial E}{\partial \sigma_m} = \sum_i r_{im} \left[ -\frac{1}{\sigma_m} + \frac{(y_i - \mu_m)^2}{\sigma_m^3} \right] = 0 \Rightarrow \sigma_m^2 = \frac{\sum_i r_{im}(y_i - \mu_m)^2}{\sum_i r_{im}},$$

$$\frac{\partial E}{\partial \pi_m} = \sum_i r_{im} \frac{1}{\pi_m}, \quad \frac{\partial E}{\partial \pi_m} + \lambda = 0 \Rightarrow \pi_m = \frac{1}{n} \sum_i r_{im},$$

where $\lambda$ is a Lagrange multiplier ensuring that the mixing proportions sum to unity.
Factor Analysis

\[ \begin{align*}
Y_1 & \quad \cdots \quad Y_D \\
X_1 & \quad \cdots \quad X_K \\
\Lambda
\end{align*} \]

Linear generative model: 
\[ y_d = \sum_{k=1}^{K} \Lambda_{dk} x_k + \epsilon_d \]

- \( x_k \) are independent \( \mathcal{N}(0, 1) \) Gaussian factors
- \( \epsilon_d \) are independent \( \mathcal{N}(0, \Psi_{dd}) \) Gaussian noise
- \( K < D \)

So, \( y \) is Gaussian with:
\[ p(y) = \int p(x)p(y|x)dx = \mathcal{N}(0, \Lambda\Lambda^\top + \Psi) \]

where \( \Lambda \) is a \( D \times K \) matrix, and \( \Psi \) is diagonal.

**Dimensionality Reduction:** Finds a low-dimensional projection of high dimensional data that captures the correlation structure of the data.
EM for Factor Analysis

The model for $y$:

$$p(y|\theta) = \int p(x|\theta)p(y|x, \theta)dx = \mathcal{N}(0, \Lambda\Lambda^T + \Psi)$$

Model parameters: $\theta = \{\Lambda, \Psi\}$.

**E step:** For each data point $y_n$, compute the posterior distribution of hidden factors given the observed data: $q_n(x) = p(x|y_n, \theta_t)$.

**M step:** Find the $\theta_{t+1}$ that maximises $F(q, \theta)$:

$$F(q, \theta) = \sum_n \int q_n(x) \left[ \log p(x|\theta) + \log p(y_n|x, \theta) - \log q_n(x) \right] dx$$

$$= \sum_n \int q_n(x) \left[ \log p(x|\theta) + \log p(y_n|x, \theta) \right] dx + c.$
The E step for Factor Analysis

**E step:** For each data point \(y_n\), compute the posterior distribution of hidden factors given the observed data: 
\[
q_n(x) = p(x|y_n, \theta) = p(x, y_n|\theta)/p(y_n|\theta)
\]

**Tactic:** write \(p(x, y_n|\theta)\), consider \(y_n\) to be fixed. What is this as a function of \(x\)?

\[
p(x, y_n) = p(x)p(y_n|x) = (2\pi)^{-\frac{K}{2}} \exp\left\{ -\frac{1}{2} x^\top x \right\} |2\pi\Psi|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (y_n - \Lambda x)^\top \Psi^{-1} (y_n - \Lambda x) \right\}
\]
\[
= c \times \exp\left\{ -\frac{1}{2} \left[ x^\top x + (y_n - \Lambda x)^\top \Psi^{-1} (y_n - \Lambda x) \right] \right\}
\]
\[
= c' \times \exp\left\{ -\frac{1}{2} \left[ x^\top (I + \Lambda^\top \Psi^{-1} \Lambda)x - 2x^\top \Lambda^\top \Psi^{-1} y_n \right] \right\}
\]
\[
= c'' \times \exp\left\{ -\frac{1}{2} \left[ x^\top \Sigma^{-1} x - 2x^\top \Sigma^{-1} \mu + \mu^\top \Sigma^{-1} \mu \right] \right\}
\]
So \(\Sigma = (I + \Lambda^\top \Psi^{-1} \Lambda)^{-1} = I - \beta \Lambda\) and \(\mu = \Sigma \Lambda^\top \Psi^{-1} y_n = \beta y_n\). Where \(\beta = \Sigma \Lambda^\top \Psi^{-1}\). Note that \(\mu\) is a linear function of \(y_n\) and \(\Sigma\) does not depend on \(y_n\).
The M step for Factor Analysis

**M step:** Find $\theta_{t+1}$ maximising $\mathcal{F} = \sum_n \int q_n(x) \left[ \log p(x|\theta) + \log p(y_n|x, \theta) \right] dx + c$

$$
\log p(x|\theta) + \log p(y_n|x, \theta) = c - \frac{1}{2} x^\top x - \frac{1}{2} \log |\Psi| - \frac{1}{2} (y_n - \Lambda x)^\top \Psi^{-1} (y_n - \Lambda x)
$$

$$
= c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[ y_n^\top \Psi^{-1} y_n - 2 y_n^\top \Psi^{-1} \Lambda x + x^\top \Lambda^\top \Psi^{-1} \Lambda x \right]
$$

$$
= c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[ y_n^\top \Psi^{-1} y_n - 2 y_n^\top \Psi^{-1} \Lambda x + \text{Tr} \left[ \Lambda^\top \Psi^{-1} \Lambda x x^\top \right] \right]
$$

Taking expectations over $q_n(x)\ldots$

$$
= c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[ y_n^\top \Psi^{-1} y_n - 2 y_n^\top \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^\top \Psi^{-1} \Lambda (\mu_n \mu_n^\top + \Sigma) \right] \right]
$$

Note that we don’t need to know everything about $q$, just the expectations of $x$ and $xx^\top$ under $q$ (i.e. the expected sufficient statistics).
The M step for Factor Analysis (cont.)

\[ F = e - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \left[ y_n^T \Psi^{-1} y_n - 2y_n^T \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda (\mu_n \mu_n^T + \Sigma) \right] \right] \]

Taking derivatives w.r.t. \( \Lambda \) and \( \Psi^{-1} \), using \( \frac{\partial \text{Tr}[AB]}{\partial B} = A^T \) and \( \frac{\partial \log |A|}{\partial A} = A^{-T} \):

\[
\frac{\partial F}{\partial \Lambda} = \Psi^{-1} \sum_n y_n \mu_n^T - \Psi^{-1} \Lambda \left( N\Sigma + \sum_n \mu_n \mu_n^T \right) = 0
\]

\[
\hat{\Lambda} = \left( \sum_n y_n \mu_n^T \right) \left( N\Sigma + \sum_n \mu_n \mu_n^T \right)^{-1}
\]

\[
\frac{\partial F}{\partial \Psi^{-1}} = \frac{N}{2} \Psi - \frac{1}{2} \sum_n \left[ y_n y_n^T - \Lambda \mu_n y_n^T - y_n \mu_n^T \Lambda^T + \Lambda (\mu_n \mu_n^T + \Sigma) \Lambda^T \right]
\]

\[
\hat{\Psi} = \frac{1}{N} \sum_n \left[ y_n y_n^T - \Lambda \mu_n y_n^T - y_n \mu_n^T \Lambda^T + \Lambda (\mu_n \mu_n^T + \Sigma) \Lambda^T \right]
\]

\[
\hat{\Psi} = \Lambda \Sigma \Lambda^T + \frac{1}{N} \sum_n (y_n - \Lambda \mu_n) (y_n - \Lambda \mu_n)^T \quad \text{(squared residuals)}
\]

Note: we should actually only take derivatives w.r.t. \( \Psi_{dd} \) since \( \Psi \) is diagonal. When \( \Sigma \to 0 \) these become the equations for linear regression!
Partial M steps and Partial E steps

**Partial M steps:** The proof holds even if we just increase $F$ wrt $\theta$ rather than maximize. (Dempster, Laird and Rubin (1977) call this the generalized EM, or GEM, algorithm).

**Partial E steps:** We can also just increase $F$ wrt to some of the $q$s.

For example, sparse or online versions of the EM algorithm would compute the posterior for a subset of the data points or as the data arrives, respectively. You can also update the posterior over a subset of the hidden variables, while holding others fixed...
EM for exponential families

**Defn:** $p$ is in the exponential family for $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ if it can be written:

$$
p(\mathbf{z}|\theta) = b(\mathbf{z}) \exp\{\theta^\top s(\mathbf{z})\}/\alpha(\theta)
$$

where $\alpha(\theta) = \int b(\mathbf{z}) \exp\{\theta^\top s(\mathbf{z})\} d\mathbf{z}$

**E step:** $q(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \theta)$

**M step:** $\theta^{(k)} := \operatorname{arg\ max}_\theta \mathcal{F}(q, \theta)$

$$
\mathcal{F}(q, \theta) = \int q(\mathbf{x}) \log p(\mathbf{x}, \mathbf{y}|\theta) d\mathbf{x} - \mathcal{H}(q)
$$

$$
= \int q(\mathbf{x})[\theta^\top s(\mathbf{z}) - \log \alpha(\theta)] d\mathbf{x} + \text{const}
$$

It is easy to verify that:

$$
\frac{\partial \log \alpha(\theta)}{\partial \theta} = \mathbb{E}[s(\mathbf{z})|\theta]
$$

Therefore, M step solves:

$$
\frac{\partial \mathcal{F}}{\partial \theta} = \mathbb{E}_q(\mathbf{x})[s(\mathbf{z})] - \mathbb{E}[s(\mathbf{z})|\theta] = 0
$$
Mixtures of Factor Analysers

Simultaneous clustering and dimensionality reduction.

\[
p(y|\theta) = \sum_k \pi_k \mathcal{N}(\mu_k, \Lambda_k \Lambda_k^\top + \Psi)
\]

where \(\pi_k\) is the mixing proportion for FA \(k\), \(\mu_k\) is its centre, \(\Lambda_k\) is its “factor loading matrix”, and \(\Psi\) is a common sensor noise model. \(\theta = \{\{\pi_k, \mu_k, \Lambda_k\}_{k=1}^{K}, \Psi\}\)

We can think of this model as having two sets of hidden latent variables:

- A discrete indicator variable \(s_n \in \{1, \ldots K\}\)
- For each factor analyzer, a continuous factor vector \(x_{n,k} \in \mathcal{R}^{D_k}\)

\[
p(y|\theta) = \sum_{s_n=1}^{K} p(s_n|\theta) \int p(x|s_n, \theta)p(y_n|x, s_n, \theta) \, dx
\]

As before, an EM algorithm can be derived for this model:

**E step**: Infer joint distribution of latent variables, \(p(x_n, s_n|y_n, \theta)\)

**M step**: Maximize \(\mathcal{F}\) with respect to \(\theta\).
Proof of the Matrix Inversion Lemma

\[(A + XBX^\top)^{-1} = A^{-1} - A^{-1}X(B^{-1} + X^\top A^{-1}X)^{-1}X^\top A^{-1}\]

Need to prove:

\[(A^{-1} - A^{-1}X(B^{-1} + X^\top A^{-1}X)^{-1}X^\top A^{-1}) (A + XBX^\top) = I\]

Expand:

\[I + A^{-1}XBX^\top - A^{-1}X(B^{-1} + X^\top A^{-1}X)^{-1}X^\top - A^{-1}X(B^{-1} + X^\top A^{-1}X)^{-1}X^\top A^{-1}XBX^\top\]

Regroup:

\[= I + A^{-1}X \left( BX^\top - (B^{-1} + X^\top A^{-1}X)^{-1}X^\top - (B^{-1} + X^\top A^{-1}X)^{-1}X^\top A^{-1}XBX^\top \right)\]

\[= I + A^{-1}X \left( BX^\top - (B^{-1} + X^\top A^{-1}X)^{-1}B^{-1}BX^\top - (B^{-1} + X^\top A^{-1}X)^{-1}X^\top A^{-1}XBX^\top \right)\]

\[= I + A^{-1}X \left( BX^\top - (B^{-1} + X^\top A^{-1}X)^{-1}(B^{-1} + X^\top A^{-1}X)BX^\top \right)\]

\[= I + A^{-1}X(BX^\top - BX^\top) = I\]