Probabilistic & Unsupervised Learning

The EM algorithm

Maneesh Sahani
maneesh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit, and
MSc ML/CSML, Dept Computer Science
University College London

Term 1, Autumn 2013

Log-likelihoods

- Exponential family models: $p(\mathbf{x}|\boldsymbol{\theta}) = f(\mathbf{x})e^{\boldsymbol{\theta}^{\mathsf{T}}\mathbf{T}(\mathbf{x})}/Z(\boldsymbol{\theta})$

$$\ell(\theta) = \theta^{\mathsf{T}} \sum_n T(\mathbf{x}_n) - N \log Z(\theta) \quad (+\text{ constants})$$

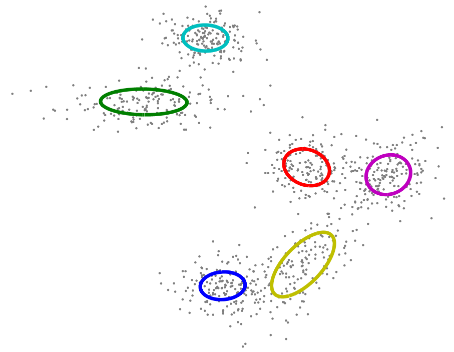
  - Concave function.
  - Maximum may be closed-form.
  - If not, numerical optimisation is still generally straightforward.

- Latent variable models: $p(\mathbf{x}|\boldsymbol{\theta}_x, \boldsymbol{\theta}_y) = \int d\mathbf{y}\, f_x(\mathbf{x}) \frac{e^{\phi(\boldsymbol{\theta}_x,\mathbf{y})^{\mathsf{T}}\mathbf{T}_x(\mathbf{x})}}{Z_x(\phi(\boldsymbol{\theta}_x,\mathbf{y}))} f_y(\mathbf{y}) \frac{e^{\boldsymbol{\theta}_y^{\mathsf{T}}\mathbf{T}_y(\mathbf{y})}}{Z_y(\boldsymbol{\theta}_y)}$

$$\ell(\theta_x, \theta_y) = \sum_n \log \int d\mathbf{y}\, f_x(\mathbf{x}) \frac{e^{\phi(\boldsymbol{\theta}_x,\mathbf{y})^{\mathsf{T}}\mathbf{T}_x(\mathbf{x})}}{Z_x(\phi(\boldsymbol{\theta}_x,\mathbf{y}))} f_y(\mathbf{y}) \frac{e^{\boldsymbol{\theta}_y^{\mathsf{T}}\mathbf{T}_y(\mathbf{y})}}{Z_y(\boldsymbol{\theta}_y)}$$

  - Usually no closed form optimum.
  - Often multiple local maxima.
  - Direct numerical optimisation may be possible but infrequently easy.

Example: mixture of Gaussians



Data: $\mathcal{X} = \{\mathbf{x}_1 \ldots \mathbf{x}_N\}$

Latent process:
$$s_i \overset{\text{iid}}{\sim} \text{Disc}[\boldsymbol{\pi}]$$

Component distributions:
$$\mathbf{x}_i \mid (s_i = m) \sim \mathcal{P}_m[\theta_m] = \mathcal{N}(\boldsymbol{\mu}_m, \Sigma_m)$$

Marginal distribution:
$$P(\mathbf{x}_i) = \sum_{m=1}^{k} \pi_m P_m(\mathbf{x}; \theta_m)$$

Log-likelihood:
$$\ell(\{\boldsymbol{\mu}_m\}, \{\Sigma_m\}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \log \sum_{m=1}^{k} \frac{\pi_m}{\sqrt{|2\pi\Sigma_m|}} e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_m)^{\mathsf{T}}\Sigma_m^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_m)}$$

The joint-data likelihood

- For many models, maximisation might be straightforward if $\mathbf{y}$ were not latent, and we could just maximise the joint-data likelihood:

$$\ell(\theta_x, \theta_y) = \sum_n \phi(\boldsymbol{\theta}_x, \mathbf{y}_n)^{\mathsf{T}}\mathbf{T}_x(\mathbf{x}_n) + \boldsymbol{\theta}_y^{\mathsf{T}} \sum_n \mathbf{T}_y(\mathbf{y}_n) - \sum_n \log Z_x(\phi(\boldsymbol{\theta}_x, \mathbf{y}_n)) - N \log Z_y(\boldsymbol{\theta}_y)$$

- Conversely, if we knew θ, we could compute (the posterior over) the values of $\mathbf{y}$.
- Idea: update $\boldsymbol{\theta}$ and (the distribution on) $\mathbf{y}$ in alternation, converging to a self-consistent answer.
- Will this yield the right answer?
- Typically, it will (as we shall see). This is the Expectation Maximisation (EM) algorithm.

The Expectation Maximisation (EM) algorithm

The EM algorithm (Dempster, Laird & Rubin, 1977; but significant earlier precedents) finds a (local) maximum of a latent variable model likelihood. It starts from arbitrary values of the parameters, and iterates two steps:

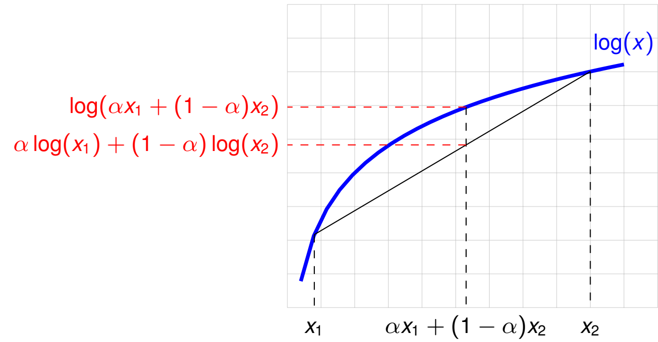
E step: Fill in values of latent variables according to posterior given data.

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.
- How does it work?

Jensen's inequality

One view: EM iteratively refines a lower bound on the log-likelihood.



In general:

For $\alpha_i \geq 0$, $\sum \alpha_i = 1$ (and $\{x_i > 0\}$):

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

For probability measure α and concave f

$$f\left(\mathrm{E}_\alpha[x]\right) \geq \mathrm{E}_\alpha[f(x)]$$

Equality (if and) only if $f(x)$ is almost surely constant or linear on (convex) support of α.

The lower bound for EM – "free energy"

Observed data $\mathcal{X} = \{\mathbf{x}_i\}$; Latent variables $\mathcal{Y} = \{\mathbf{y}_i\}$; Parameters $\theta = \{\theta_x, \theta_y\}$.

Log-likelihood:

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

By Jensen, any distribution, $q(\mathcal{Y})$, over the latent variables generates a lower bound:

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

Now,

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

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

where $\mathbf{H}[q]$ is the entropy of $q(\mathcal{Y})$.

So:

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

The E and M steps of EM

The lower bound on the log likelihood is given by:

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

EM alternates between:

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

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})}{\operatorname{argmax}}\ \mathcal{F}\left(q(\mathcal{Y}), \theta^{(k-1)}\right).$$

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

$$\theta^{(k)} := \underset{\theta}{\operatorname{argmax}}\ \mathcal{F}\left(q^{(k)}(\mathcal{Y}), \theta\right) = \underset{\theta}{\operatorname{argmax}}\ \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q^{(k)}(\mathcal{Y})}$$

The second equality comes from the fact $\mathbf{H}\left[q^{(k)}(\mathcal{Y})\right]$ does not depend directly on θ.

The E Step

The free energy can be re-written

$$\begin{aligned}
\mathcal{F}(q,\theta) &= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y},\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)P(\mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \int q(\mathcal{Y}) \log P(\mathcal{X}|\theta)\, d\mathcal{Y} + \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}|\mathcal{X},\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \\
&= \ell(\theta) - \mathbf{KL}[q(\mathcal{Y})\|P(\mathcal{Y}|\mathcal{X},\theta)]
\end{aligned}$$

The second term is the Kullback-Leibler divergence.

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

But $\mathbf{KL}[q\|p]$ is zero if and only if $q = p$ (see appendix.)

So, the E step sets

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

and, after an E step, the free energy equals the likelihood.

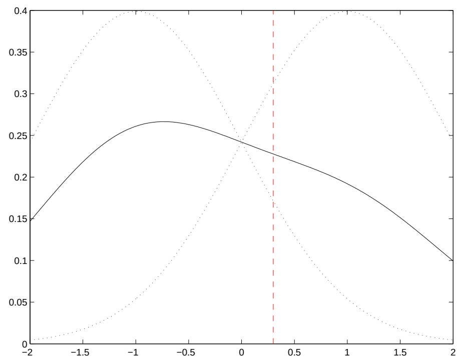
Coordinate Ascent in $\mathcal{F}$ (Demo)

To visualise, we consider a one parameter / one latent mixture:

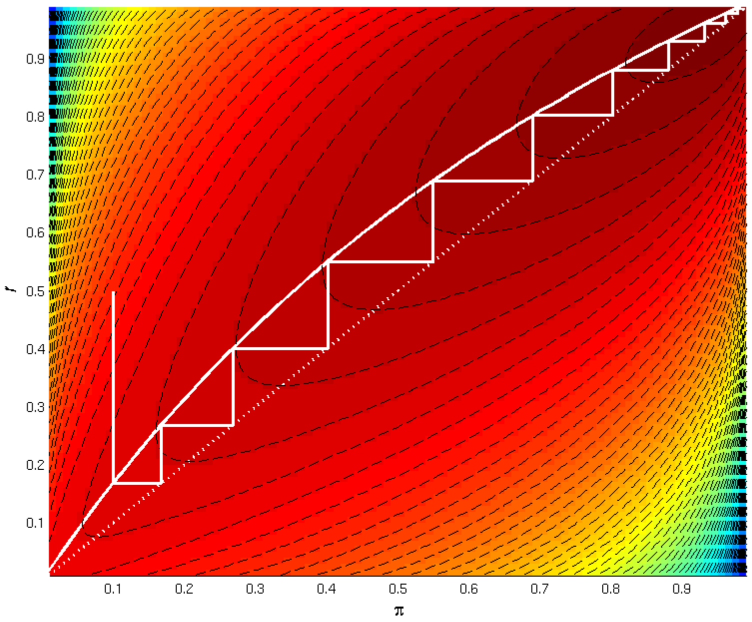
$$s \sim \text{Bernoulli}[\pi]$$

$$x|s=0 \sim \mathcal{N}[-1,1] \qquad x|s=1 \sim \mathcal{N}[1,1]\,.$$

Single data point $x_1 = .3$.
$q(s)$ is a distribution on a single binary latent, and so is represented by $r_1 \in [0,1]$.



Coordinate Ascent in $\mathcal{F}$ (Demo)



EM Never Decreases the Likelihood

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

$$\ell(\theta^{(k-1)}) \underset{\text{E step}}{=} \mathcal{F}(q^{(k)},\theta^{(k-1)}) \underset{\text{M step}}{\leq} \mathcal{F}(q^{(k)},\theta^{(k)}) \underset{\text{Jensen}}{\leq} \ell(\theta^{(k)}),$$

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ.
- $\mathcal{F} \leq \ell$ 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 θ iff the likelihood increases.

Can also show that fixed points of EM (generally) correspond to maxima of the likelihood (see appendices).

Partial M steps and Partial E steps

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

In fact, immediately after an E step

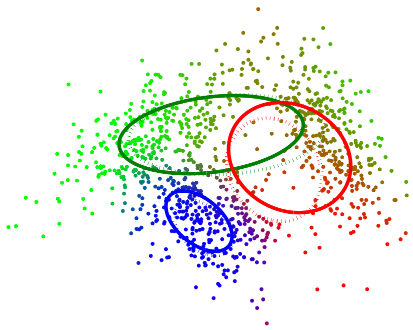
$$\left.\frac{\partial}{\partial\theta}\right|_{\theta^{(k-1)}} \langle \log P(\mathcal{X},\mathcal{Y}|\theta)\rangle_{q^{(k)}(\mathcal{Y})[=P(\mathcal{Y}|\mathcal{X},\theta^{(k-1)})]} = \left.\frac{\partial}{\partial\theta}\right|_{\theta^{(k-1)}} \log P(\mathcal{X}|\theta)$$

So E-step (inference) can be used to construct other gradient-based optimisation schemes (e.g. "Expectaton Conjugate Gradient", Salakhutdinov et al. *ICML* 2003).

Partial E steps: We can also just *increase* $\mathcal{F}$ wrt to some of the qs.

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. One might also update the posterior over a subset of the hidden variables, while holding others fixed...

EM for MoGs



- Evaluate responsibilities

$$r_{im} = \frac{P_m(\mathbf{x})\pi_m}{\sum_{m'} P_{m'}(\mathbf{x})\pi_{m'}}$$

- Update parameters

$$\boldsymbol{\mu}_m \leftarrow \frac{\sum_i r_{im}\mathbf{x}_i}{\sum_i r_{im}}$$

$$\Sigma_m \leftarrow \frac{\sum_i r_{im}(\mathbf{x}_i - \boldsymbol{\mu}_m)(\mathbf{x}_i - \boldsymbol{\mu}_m)^{\mathsf{T}}}{\sum_i r_{im}}$$

$$\pi_m \leftarrow \frac{\sum_i r_{im}}{N}$$

The Gaussian mixture model (E-step)

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

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

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

The hidden variable s_i indicates which component generated observation x_i.

The E-step computes the posterior for s_i given the current parameters:

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

$$r_{im} \stackrel{\text{def}}{=} q(s_i=m) \propto \frac{\pi_m}{\sigma_m} \exp\Big\{-\frac{1}{2\sigma_m^2}(x_i-\mu_m)^2\Big\} \quad \text{(responsibilities)} \quad \leftarrow \langle\delta_{s_i=m}\rangle_q$$

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):

$$\begin{aligned} E = \langle \log p(x,s|\theta)\rangle_{q(s)} &= \sum q(s)\log[p(s|\theta)\ p(x|s,\theta)] \\ &= \sum_{i,m} r_{im}\Big[\log\pi_m - \log\sigma_m - \frac{1}{2\sigma_m^2}(x_i-\mu_m)^2\Big]. \end{aligned}$$

Optimum is found by setting the partial derivatives of E to zero:

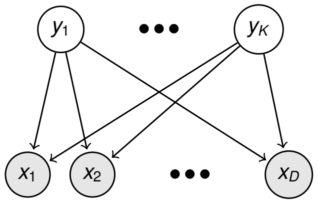
$$\frac{\partial}{\partial\mu_m}E = \sum_i r_{im}\frac{(x_i-\mu_m)}{2\sigma_m^2} = 0 \quad\Rightarrow\quad \mu_m = \frac{\sum_i r_{im}x_i}{\sum_i r_{im}},$$

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

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

where λ is a Lagrange multiplier ensuring that the mixing proportions sum to unity.

EM for Factor Analysis



The model for $\mathbf{x}$:

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

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

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

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

$$\begin{aligned}\mathcal{F}(q,\theta) &= \sum_n \int q_n(\mathbf{y}_n)\left[\log p(\mathbf{y}_n|\theta) + \log p(\mathbf{x}_n|\mathbf{y}_n,\theta) - \log q_n(\mathbf{y}_n)\right] d\mathbf{y}_n \\ &= \sum_n \int q_n(\mathbf{y}_n)\left[\log p(\mathbf{y}_n|\theta) + \log p(\mathbf{x}_n|\mathbf{y}_n,\theta)\right] d\mathbf{y}_n + \text{c}.\end{aligned}$$

The E step for Factor Analysis

E step: For each data point $\mathbf{x}_n$, compute the posterior distribution of hidden factors given the observed data: $q_n(\mathbf{y}_n) = p(\mathbf{y}_n|\mathbf{x}_n,\theta) = p(\mathbf{y}_n,\mathbf{x}_n|\theta)/p(\mathbf{x}_n|\theta)$

Tactic: write $p(\mathbf{y}_n,\mathbf{x}_n|\theta)$, consider $\mathbf{x}_n$ to be fixed. What is this as a function of $\mathbf{y}_n$?

$$\begin{aligned}p(\mathbf{y}_n,\mathbf{x}_n) &= p(\mathbf{y}_n)p(\mathbf{x}_n|\mathbf{y}_n) \\ &= (2\pi)^{-\frac{K}{2}}\exp\{-\frac{1}{2}\mathbf{y}_n^\top\mathbf{y}_n\}\;|2\pi\Psi|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\mathbf{x}_n-\Lambda\mathbf{y}_n)^\top\Psi^{-1}(\mathbf{x}_n-\Lambda\mathbf{y}_n)\} \\ &= \text{c}\times\exp\{-\frac{1}{2}[\mathbf{y}_n^\top\mathbf{y}_n + (\mathbf{x}_n-\Lambda\mathbf{y}_n)^\top\Psi^{-1}(\mathbf{x}_n-\Lambda\mathbf{y}_n)]\} \\ &= \text{c'}\times\exp\{-\frac{1}{2}[\mathbf{y}_n^\top(I+\Lambda^\top\Psi^{-1}\Lambda)\mathbf{y}_n - 2\mathbf{y}_n^\top\Lambda^\top\Psi^{-1}\mathbf{x}_n]\} \\ &= \text{c''}\times\exp\{-\frac{1}{2}[\mathbf{y}_n^\top\Sigma^{-1}\mathbf{y}_n - 2\mathbf{y}_n^\top\Sigma^{-1}\boldsymbol{\mu}_n + \boldsymbol{\mu}_n^\top\Sigma^{-1}\boldsymbol{\mu}_n]\}\end{aligned}$$

So $\Sigma = (I+\Lambda^\top\Psi^{-1}\Lambda)^{-1} = I - \beta\Lambda$ and $\boldsymbol{\mu}_n = \Sigma\Lambda^\top\Psi^{-1}\mathbf{x}_n = \beta\mathbf{x}_n$. Where $\beta = \Sigma\Lambda^\top\Psi^{-1}$.
Note that $\boldsymbol{\mu}_n$ is a linear function of $\mathbf{x}_n$ and Σ does not depend on $\mathbf{x}_n$.

The M step for Factor Analysis

M step: Find θ_{t+1} by maximising $\mathcal{F} = \sum_n \langle \log p(\mathbf{y}_n|\theta) + \log p(\mathbf{x}_n|\mathbf{y}_n,\theta)\rangle_{q_n(\mathbf{y}_n)} + \text{c}$

$$\begin{aligned}&\log p(\mathbf{y}_n|\theta) + \log p(\mathbf{x}_n|\mathbf{y}_n,\theta) \\ &= \text{c} - \frac{1}{2}\mathbf{y}_n^\top\mathbf{y}_n - \frac{1}{2}\log|\Psi| - \frac{1}{2}(\mathbf{x}_n-\Lambda\mathbf{y}_n)^\top\Psi^{-1}(\mathbf{x}_n-\Lambda\mathbf{y}_n) \\ &= \text{c'} - \frac{1}{2}\log|\Psi| - \frac{1}{2}\left[\mathbf{x}_n^\top\Psi^{-1}\mathbf{x}_n - 2\mathbf{x}_n^\top\Psi^{-1}\Lambda\mathbf{y}_n + \mathbf{y}_n^\top\Lambda^\top\Psi^{-1}\Lambda\mathbf{y}_n\right] \\ &= \text{c'} - \frac{1}{2}\log|\Psi| - \frac{1}{2}\left[\mathbf{x}_n^\top\Psi^{-1}\mathbf{x}_n - 2\mathbf{x}_n^\top\Psi^{-1}\Lambda\mathbf{y}_n + \text{Tr}\left[\Lambda^\top\Psi^{-1}\Lambda\mathbf{y}_n\mathbf{y}_n^\top\right]\right]\end{aligned}$$

Taking expectations wrt $q_n(\mathbf{y}_n)$:

$$= \text{c'} - \frac{1}{2}\log|\Psi| - \frac{1}{2}\left[\mathbf{x}_n^\top\Psi^{-1}\mathbf{x}_n - 2\mathbf{x}_n^\top\Psi^{-1}\Lambda\boldsymbol{\mu}_n + \text{Tr}\left[\Lambda^\top\Psi^{-1}\Lambda(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top + \Sigma)\right]\right]$$

Note that we don't need to know everything about $q(\mathbf{y}_n)$, just the moments $\langle\mathbf{y}_n\rangle$ and $\langle\mathbf{y}_n\mathbf{y}_n^\top\rangle$.
These are the expected sufficient statistics.

The M step for Factor Analysis (cont.)

$$\mathcal{F} = c' - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n\left[\mathbf{x}_n^\top\Psi^{-1}\mathbf{x}_n - 2\mathbf{x}_n^\top\Psi^{-1}\Lambda\boldsymbol{\mu}_n + \text{Tr}\left[\Lambda^\top\Psi^{-1}\Lambda(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top + \Sigma)\right]\right]$$

Taking derivatives wrt Λ and Ψ^{-1}, using $\frac{\partial \text{Tr}[AB]}{\partial B} = A^\top$ and $\frac{\partial \log|A|}{\partial A} = A^{-\top}$:

$$\begin{aligned}\frac{\partial\mathcal{F}}{\partial\Lambda} &= \Psi^{-1}\sum_n\mathbf{x}_n\boldsymbol{\mu}_n^\top - \Psi^{-1}\Lambda\left(N\Sigma + \sum_n\boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top\right) = 0 \\ \Rightarrow \widehat{\Lambda} &= \left(\sum_n\mathbf{x}_n\boldsymbol{\mu}_n^\top\right)\left(N\Sigma + \sum_n\boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top\right)^{-1} \\ \frac{\partial\mathcal{F}}{\partial\Psi^{-1}} &= \frac{N}{2}\Psi - \frac{1}{2}\sum_n\left[\mathbf{x}_n\mathbf{x}_n^\top - \Lambda\boldsymbol{\mu}_n\mathbf{x}_n^\top - \mathbf{x}_n\boldsymbol{\mu}_n^\top\Lambda^\top + \Lambda(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top + \Sigma)\Lambda^\top\right] \\ \Rightarrow \widehat{\Psi} &= \frac{1}{N}\sum_n\left[\mathbf{x}_n\mathbf{x}_n^\top - \Lambda\boldsymbol{\mu}_n\mathbf{x}_n^\top - \mathbf{x}_n\boldsymbol{\mu}_n^\top\Lambda^\top + \Lambda(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^\top + \Sigma)\Lambda^\top\right] \\ \widehat{\Psi} &= \Lambda\Sigma\Lambda^\top + \frac{1}{N}\sum_n(\mathbf{x}_n - \Lambda\boldsymbol{\mu}_n)(\mathbf{x}_n - \Lambda\boldsymbol{\mu}_n)^\top \qquad \text{(squared residuals)}\end{aligned}$$

Note: we should actually only take derivatives w.r.t. Ψ_{dd} since Ψ is diagonal.
As $\Sigma \to 0$ these become the equations for ML linear regression

Mixtures of Factor Analysers

Simultaneous clustering and dimensionality reduction.

$$p(\mathbf{x}|\theta) = \sum_k \pi_k \, \mathcal{N}(\boldsymbol{\mu}_k, \Lambda_k \Lambda_k^{\mathsf{T}} + \Psi)$$

where π_k is the mixing proportion for FA k, $\boldsymbol{\mu}_k$ is its centre, Λ_k is its "factor loading matrix", and Ψ is a common sensor noise model. $\theta = \{\{\pi_k, \boldsymbol{\mu}_k, \Lambda_k\}_{k=1\ldots K}, \Psi\}$
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 continous factor vector $\mathbf{y}_{n,k} \in \mathcal{R}^{D_k}$

$$p(\mathbf{x}|\theta) = \sum_{s_n=1}^{K} p(s_n|\theta) \int p(\mathbf{y}|s_n, \theta) p(\mathbf{x}_n|\mathbf{y}, s_n, \theta) \; d\mathbf{y}$$

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

E step: We need moments of $p(\mathbf{y}_n, s_n|\mathbf{x}_n, \theta)$, specifically: $\langle \delta_{s_n=m} \rangle$, $\langle \delta_{s_n=m} \mathbf{y}_n \rangle$ and $\langle \delta_{s_n=m} \mathbf{y}_n \mathbf{y}_n^{\mathsf{T}} \rangle$.

M step: Similar to M-step for FA with responsibility-weighted moments.
See `http://www.learning.eng.cam.ac.uk/zoubin/papers/tr-96-1.pdf`

EM for exponential families

EM is often applied to models whose **joint** over $\mathbf{z} = (\mathbf{y}, \mathbf{x})$ has exponential-family form:

$$p(\mathbf{z}|\theta) = f(\mathbf{z}) \exp\{\theta^{\mathsf{T}} \mathsf{T}(\mathbf{z})\}/Z(\theta)$$

(with $Z(\theta) = \int f(\mathbf{z}) \exp\{\theta^{\mathsf{T}} \mathsf{T}(\mathbf{z})\} d\mathbf{z}$) but whose marginal $p(\mathbf{x}) \notin$ *ExpFam*.
The free energy dependence on θ is given by:

$$\begin{aligned}
\mathcal{F}(q, \theta) &= \int q(\mathbf{y}) \log p(\mathbf{y}, \mathbf{x}|\theta) d\mathbf{y} - \mathbf{H}[q] \\
&= \int q(\mathbf{y}) \left[\theta^{\mathsf{T}} \mathsf{T}(\mathbf{z}) - \log Z(\theta)\right] d\mathbf{y} + \text{const wrt } \theta \\
&= \theta^{\mathsf{T}} \langle \mathsf{T}(\mathbf{z}) \rangle_{q(\mathbf{y})} - \log Z(\theta) + \text{const wrt } \theta
\end{aligned}$$

So, in the **E step** all we need to compute are the expected sufficient statistics under q.
We also have:

$$\begin{aligned}
\frac{\partial}{\partial \theta} \log Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} Z(\theta) &= \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} \int f(\mathbf{z}) \exp\{\theta^{\mathsf{T}} \mathsf{T}(\mathbf{z})\} \\
&= \int \tfrac{1}{Z(\theta)} f(\mathbf{z}) \exp\{\theta^{\mathsf{T}} \mathsf{T}(\mathbf{z})\} \cdot \mathsf{T}(\mathbf{z}) = \langle \mathsf{T}(\mathbf{z}) | \theta \rangle
\end{aligned}$$

Thus, the **M step** solves: $\displaystyle \frac{\partial \mathcal{F}}{\partial \theta} = \langle \mathsf{T}(\mathbf{z}) \rangle_{q(\mathbf{y})} - \langle \mathsf{T}(\mathbf{z}) | \theta \rangle = 0$

References

- A. P. Dempster, N. M. Laird and D. B. Rubin (1977).
  Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1 (1977), pp. 1-38.
  `http://www.jstor.org/stable/2984875`
- R. M. Neal and G. E. Hinton (1998).
  A view of the EM algorithm that justifies incremental, sparse, and other variants.
  In M. I. Jordan (editor) Learning in Graphical Models, pp. 355-368, Dordrecht: Kluwer Academic Publishers.
  `http://www.cs.utoronto.ca/∼radford/ftp/emk.pdf`
- R. Salakhutdinov, S. Roweis and Z. Ghahramani, (2003).
  Optimization with EM and expectation-conjugate-gradient.
  In ICML (pp. 672-679).
  `http://www.cs.utoronto.ca/∼rsalakhu/papers/emecg.pdf`
- Z. Ghahramani and G. E. Hinton (1996).
  The EM Algorithm for Mixtures of Factor Analyzers.
  University of Toronto Technical Report CRG-TR-96-1.
  `http://learning.eng.cam.ac.uk/zoubin/papers/tr-96-1.pdf`

Proof of the Matrix Inversion Lemma

$$(A + XBX^{\mathsf{T}})^{-1} = A^{-1} - A^{-1}X(B^{-1} + X^{\mathsf{T}}A^{-1}X)^{-1}X^{\mathsf{T}}A^{-1}$$

Need to prove:

$$\left(A^{-1} - A^{-1}X(B^{-1} + X^{\mathsf{T}}A^{-1}X)^{-1}X^{\mathsf{T}}A^{-1}\right)(A + XBX^{\mathsf{T}}) = I$$

Expand:

$$I + A^{-1}XBX^{\mathsf{T}} - A^{-1}X(B^{-1} + X^{\mathsf{T}}A^{-1}X)^{-1}X^{\mathsf{T}} - A^{-1}X(B^{-1} + X^{\mathsf{T}}A^{-1}X)^{-1}X^{\mathsf{T}}A^{-1}XBX^{\mathsf{T}}$$

Regroup:

$$\begin{aligned}
&= I + A^{-1}X\left(BX^{\mathsf{T}} - (B^{-1} + X^{\mathsf{T}}A^{-1}X)^{-1}X^{\mathsf{T}} - (B^{-1} + X^{\mathsf{T}}A^{-1}X)^{-1}X^{\mathsf{T}}A^{-1}XBX^{\mathsf{T}}\right) \\
&= I + A^{-1}X\left(BX^{\mathsf{T}} - (B^{-1} + X^{\mathsf{T}}A^{-1}X)^{-1}B^{-1}BX^{\mathsf{T}} - (B^{-1} + X^{\mathsf{T}}A^{-1}X)^{-1}X^{\mathsf{T}}A^{-1}XBX^{\mathsf{T}}\right) \\
&= I + A^{-1}X\left(BX^{\mathsf{T}} - (B^{-1} + X^{\mathsf{T}}A^{-1}X)^{-1}(B^{-1} + X^{\mathsf{T}}A^{-1}X)BX^{\mathsf{T}}\right) \\
&= I + A^{-1}X(BX^{\mathsf{T}} - BX^{\mathsf{T}}) = I
\end{aligned}$$

$\mathbf{KL}[q(x)\|p(x)] \geq 0$, with equality iff $\forall x: \; p(x) = q(x)$

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

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

To minimize wrt distribution q we need a Lagrange multiplier to enforce normalisation:

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

Find conditions for stationarity

$$\left.\begin{aligned} \frac{\partial E}{\partial q_i} &= \log q_i - \log p_i + 1 - \lambda = 0 \Rightarrow q_i = p_i \exp(\lambda - 1) \\ \frac{\partial E}{\partial \lambda} &= 1 - \sum_i q_i = 0 \Rightarrow \sum_i q_i = 1 \end{aligned}\right\} \Rightarrow q_i = p_i.$$

Check sign of curvature (Hessian):

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

so unique stationary point $q_i = p_i$ is indeed a minimum. Easily verified that at that minimum, $\mathbf{KL}[q\|p] = \mathbf{KL}[p\|p] = 0$.
A similar proof holds for continuous densities, using functional derivatives.

Maxima in $\mathcal{F}$ correspond to maxima in ℓ

Let θ^* now be the parameter value at a local maximum of $\mathcal{F}$ (and thus at a fixed point)

Differentiating the previous expression wrt θ again we find

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

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

θ^* is a maximum of ℓ.

[...as long as the derivatives exist. They sometimes don't (zero-noise ICA)].

Fixed Points of EM are Stationary Points in ℓ

Let a fixed point of EM occur with parameter θ^*. Then:

$$\frac{\partial}{\partial \theta}\langle \log P(\mathcal{Y}, \mathcal{X} \mid \theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)}\Big|_{\theta^*} = 0$$

Now,

$$\begin{aligned} \ell(\theta) = \log P(\mathcal{X}|\theta) &= \langle \log P(\mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \\ &= \left\langle \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{P(\mathcal{Y}|\mathcal{X}, \theta)} \right\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \\ &= \langle \log P(\mathcal{Y}, \mathcal{X}|\theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} - \langle \log P(\mathcal{Y}|\mathcal{X}, \theta)\rangle_{P(\mathcal{Y}|\mathcal{X},\theta^*)} \end{aligned}$$

so,

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

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

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

So, EM converges to a stationary point of $\ell(\theta)$.