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: \( p(x|\theta) = f(x)e^{\theta^T T(x)/Z(\theta)} \)
  \[
  \ell(\theta) = \theta^T \sum_n T(x_n) - N \log Z(\theta) \quad (+ \text{constants})
  \]
  - Concave function.
  - Maximum may be closed-form.
  - If not, numerical optimisation is still generally straightforward.

- Latent variable models: \( p(x|\theta, y) = \int dy f_y(y)e^{\phi(\theta_x, y) T_x(x)} Z_x(\phi(\theta_x, y)) f_x(x) \)
  \[
  \ell(\theta_x, \theta_y) = \sum_n \log \int dy f_y(y)e^{\phi(\theta_x, y) T_x(x)} Z_x(\phi(\theta_x, y)) f_x(x)
  \]
  - Usually no closed form optimum.
  - Often multiple local maxima.
  - Direct numerical optimisation may be possible but infrequently easy.

Example: mixture of Gaussians

Data: \( \mathcal{X} = \{x_1 \ldots x_N\} \)

Latent process:
\( s_i \sim \text{Disc}[\pi] \)

Component distributions:
\( x_i \mid (s_i = m) \sim p_m[x] = N(\mu_m, \Sigma_m) \)

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

Log-likelihood:
\[
\ell(\{\mu_m\}, \{\Sigma_m\}, \pi) = \sum_{i=1}^N \sum_{m=1}^k \pi_m e^{-\frac{1}{2}(x_i - \mu_m)^T \Sigma_m^{-1}(x_i - \mu_m)}
\]

The joint-data likelihood

- For many models, maximisation might be straightforward if \( y \) were not latent, and we could just maximise the joint-data likelihood:
  \[
  \ell(\theta_x, \theta_y) = \sum_n \phi(\theta_x, y_n) T_x(x_n) + \theta_y^T \sum_n T_y(y_n) - \sum_n \log Z_x(\phi(\theta_x, y_n)) - N \log Z_y(\theta_y)
  \]
  - Conversely, if we knew \( \theta \), we could compute (the posterior over) the values of \( y \).
  - Idea: update \( \theta \) and (the distribution on) \( y \) 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?

### The lower bound for EM – “free energy”

Observed data \(X = \{x_i\}\); Latent variables \(Y = \{y_i\}\); Parameters \(\theta = \{\theta_x, \theta_y\}\).

The lower bound on the log likelihood is given by:

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

where \(H[q]\) is the entropy of \(q(Y)\).

So:

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

**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 \(\{\alpha_i > 0\}\)):

\[
\log \left( \sum_i \alpha_i x_i \right) \geq \sum_i \alpha_i \log(x_i)
\]

For probability measure \(\alpha\) and concave \(f\):

\[
\log \left( \sum_i \alpha_i x_i \right) \geq \sum_i \alpha_i \log(x_i)
\]

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

### The E and M steps of EM

The E and M steps of EM:

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

  \[
  q^{(k)}(Y) := \arg\max_{q(Y)} \mathcal{F}(q(Y), q^{(k-1)}).
  \]

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

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

The second equality comes from the fact \(H[q^{(k)}(Y)]\) does not depend directly on \(\theta\).
The E Step

The free energy can be re-written

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

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

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

\[ = \ell(\theta) - KL[q(Y)||P(Y|X, \theta)] \]

The second term is the Kullback-Leibler divergence.

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

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

So, the E step sets

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

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

Coordinate Ascent in \( F \) (Demo)

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

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

\[ x|s=0 \sim N[-1, 1] \quad x|s=1 \sim N[1, 1]. \]

Single data point \( x_1 = .3 \).

\( q(s) \) is a distribution on a single binary latent, and so is represented by \( \pi \in [0, 1] \).

![Coordinate Ascent in \( F \) (Demo)](image)

EM Never Decreases the Likelihood

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

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

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt \( \theta \).
- \( \mathcal{F} \leq \ell \) 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 \( \theta \) 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 $\theta$ rather than maximize. (Dempster, Laird and Rubin (1977) call this the generalized EM, or GEM, algorithm).

In fact, immediately after an E step
\[
\frac{\partial}{\partial \theta} \log P(X, Y|\theta)\Big|_{\theta^{(k-1)}} = \frac{\partial}{\partial \theta} \log P(X|\theta)
\]
So E-step (inference) can be used to construct other gradient-based optimisation schemes (e.g. “Expectation Conjugate Gradient”, Salakhutdinov et al. ICML 2003).

Partial E steps: We can also just increase $\mathcal{F}$ wrt 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. One might also update the posterior over a subset of the hidden variables, while holding others fixed...

The Gaussian mixture model (E-step)

In a univariate Gaussian mixture model, the density of a data point $x$ is:
\[
p(x|\theta) = \sum_{m=1}^{k} p(s = m|\theta)p(x|s = m, \theta) \propto \sum_{m=1}^{k} \pi_m \exp \left\{ -\frac{1}{2\sigma_m^2} (x - \mu_m)^2 \right\},
\]
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 generated observation $x_i$.

The E-step computes the posterior for $s_i$ given the current parameters:
\[
q(s_i) = p(s_i = m|x_i, \theta) \propto p(x_i|s_i, \theta)p(s_i|\theta)
\]
\[
r_m \overset{def}{=} q(s_i = m) \propto \frac{\pi_m}{\sigma_m^2} \exp \left\{ -\frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right\} \quad \text{(responsibilities)} \leftarrow \langle \delta_{s_i = m} \rangle_q
\]
with the normalization such that $\sum_m r_m = 1$.

EM for MoGs

Evaluate responsibilities
\[
r_m = \frac{P_m(x)\pi_m}{\sum_{m'} P_{m'}(x)\pi_{m'}}
\]

Update parameters
\[
\mu_m \leftarrow \frac{\sum_i r_m x_i}{\sum_i r_m},
\]
\[
\sigma_m = \frac{\sum_i r_m (x_i - \mu_m)^2}{\sum_i r_m}
\]
\[
\pi_m \leftarrow \frac{\sum_i r_m}{N}
\]

The Gaussian mixture model (M-step)

In the M-step we optimize the sum (since $s$ is discrete):
\[
E = \langle \log p(x, s|\theta) \rangle_{q(s)} = \sum_{i,m} q(s_i) \log[p(x_i|s_i, \theta) p(s_i|s, \theta)]
\]
\[
= \sum_i \sum_m r_m [\log \pi_m - \log \sigma_m - \frac{1}{2\sigma_m^2} (x_i - \mu_m)^2].
\]

Optimum is found by setting the partial derivatives of $E$ to zero:
\[
\frac{\partial}{\partial \mu_m} E = \sum_i r_m \frac{x_i - \mu_m}{2\sigma_m^2} = 0 \Rightarrow \mu_m = \frac{\sum_i r_m x_i}{\sum_i r_m},
\]
\[
\frac{\partial}{\partial \sigma_m} E = \sum_i r_m \left[ -\frac{1}{\sigma_m} + \frac{(x_i - \mu_m)^2}{\sigma_m^3} \right] = 0 \Rightarrow \sigma_m^2 = \frac{\sum_i r_m (x_i - \mu_m)^2}{\sum_i r_m},
\]
\[
\frac{\partial}{\partial \pi_m} E = \sum_i r_m \frac{1}{\pi_m}, \quad \frac{\partial E}{\partial \pi_m} + \lambda = 0 \Rightarrow \pi_m = \frac{1}{N} \sum_i r_m,
\]
where $\lambda$ is a Lagrange multiplier ensuring that the mixing proportions sum to unity.
The model for \(x\):

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

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

**The E step for Factor Analysis**

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

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

\[
p(y, x) = p(y|x) = (2\pi)^{-\frac{n}{2}} \exp\{-\frac{1}{2}(y - \mu_n)^T \Sigma^{-1} (y - \mu_n)\} + \frac{1}{2} \sum_{k} (\mu_k - \mu_n)^T \Sigma (\mu_k - \mu_n)
\]

So \(\Sigma = (I + \Lambda^T \Psi^{-1} \Lambda)^{-1} = I - \beta \Lambda\) and \(\mu_n = \Sigma \Lambda^T \Psi^{-1} x_n = \beta x_n\). Where \(\beta = \Sigma \Lambda^T \Psi^{-1}\). Note that \(\mu_n\) is a linear function of \(x_n\) and \(\Sigma\) does not depend on \(x_n\).

**The M step for Factor Analysis**

**M step:** Find \(\theta_{n+1}\) by maximising \(\mathcal{F} = \sum_n (\log p(y_n|\theta) + \log p(x_n|y_n, \theta))q_n(y_n) + c\)

\[
\log p(y_n|\theta) + \log p(x_n|y_n, \theta)
= c - \frac{1}{2} y_n^T y_n - \frac{1}{2} \log |\Psi| - \frac{1}{2} (x_n - \Lambda y_n)^T \Psi^{-1} (x_n - \Lambda y_n)
= c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda y_n + y_n^T \Lambda^T \Psi^{-1} \Lambda y_n\right]
= c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda y_n + \text{Tr} \left[\Lambda^T \Psi^{-1} \Lambda y_n y_n^T\right]\right]
\]

Taking expectations wrt \(q_n(y_n)\):

\[
c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[\Lambda^T \Psi^{-1} \Lambda (\mu_n + \Sigma)\right]\right]
\]

Note that we don’t need to know everything about \(q(y_n)\), just the moments \(\langle y_n \rangle\) and \(\langle y_n y_n^T \rangle\). These are the expected sufficient statistics.

**The M step for Factor Analysis (cont.)**

\[
\mathcal{F} = c' - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \left[x_n^T \Psi^{-1} x_n - 2x_n^T \Psi^{-1} \Lambda \mu_n + \text{Tr} \left[\Lambda^T \Psi^{-1} \Lambda (\mu_n + \Sigma)\right]\right]
\]

Taking derivatives wrt \(\Lambda\) and \(\Psi^{-1}\), using \(\frac{\partial \log |A|}{\partial A} = A^{-T}\) and \(\frac{\partial \log |A|}{\partial \Lambda} = A^{-T}\):

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

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

\[
\frac{\partial \mathcal{F}}{\partial \Psi^{-1}} = \frac{N}{2} \Psi - \frac{1}{2} \sum_n \left[x_n x_n^T - \Lambda \mu_n \mu_n^T - x_n \mu_n \mu_n^T + \Lambda (\mu_n + \Sigma)\right] \Lambda^T = 0
\]

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

\[
\tilde{\Psi} = \Lambda \Sigma \Lambda^T + \frac{1}{N} \sum_n (x_n - \Lambda \mu_n) (x_n - \Lambda \mu_n)^T \quad (\text{quared residuals})
\]

Note: we should actually only take derivatives w.r.t. \(\Psi_{ii}\) since \(\Psi\) is diagonal.

As \(\Sigma \to 0\) these become the equations for ML linear regression.
Mixtures of Factor Analysers

Simultaneous clustering and dimensionality reduction.

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

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 \( y_{nk} \in \mathcal{R} \)

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

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

**E step:** We need moments of \( p(y_n, s_n, x_n, \theta) \), specifically: \( \langle \delta_{s_n=m} \rangle, \langle \delta_{s_n=m} y_n \rangle \) and \( \langle \delta_{s_n=m} y_n^T \rangle \).

**M step:** Similar to M-step for FA with responsibility-weighted moments.


EM for exponential families

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

\[ p(z|\theta) = f(z) \exp(\theta^T T(z)) / Z(\theta) \]

(with \( Z(\theta) = \int f(z) \exp(\theta^T T(z)) \, dz \)) but whose marginal \( p(x) \not\in \text{ExpFam} \).

The free energy dependence on \( \theta \) is given by:

\[ \mathcal{F}(q, \theta) = \int q(y) \log p(y, x|\theta) \, dy - H[q] \]

\[ = \int q(y) [\theta^T T(z) - \log Z(\theta)] \, dy + \text{const wrt} \theta \]

So, in the **E step** all we need to compute are the expected sufficient statistics under \( q \).

We also have:

\[ \frac{\partial}{\partial \theta} \log Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} Z(\theta) = \int f(z) \exp(\theta^T T(z)) \cdot T(z) = \langle T(z) \rangle_{q(y)} \]

Thus, the **M step** solves:

\[ \frac{\partial \mathcal{F}}{\partial \theta} = \langle T(z) \rangle_{q(y)} - \langle T(z) \rangle_{\theta} = 0 \]

Proof of the Matrix Inversion Lemma

\[ (A + XB^T)^{-1} = A^{-1} - A^{-1}X(B^{-1} + X^T A^{-1}X)^{-1}X^T A^{-1} \]

Need to prove:

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

Expand:

\[ I + A^{-1}XB^T - A^{-1}X(B^{-1} + X^T A^{-1}X)^{-1}X^T - A^{-1}X(B^{-1} + X^T A^{-1}X)^{-1}X^T A^{-1}XB^T \]

Regroup:

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

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

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

\[ = I + A^{-1}X(B^T - B^T) = I \]
Thus the curvature of the likelihood is negative and

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

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

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

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

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

Find conditions for stationarity

\[ \begin{align*}
\frac{\partial E}{\partial q_i} &= \log q_i - \log p_i + 1 - \lambda \Rightarrow q_i = p_i \exp(\lambda - 1) \\
\frac{\partial E}{\partial \lambda} &= 1 - \sum_i q_i = 0 \Rightarrow \sum_i q_i = 1
\end{align*} \]

\[ \Rightarrow q_i = p_i. \]

Check sign of curvature (Hessian):

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

so unique stationary point \( q_i = p_i \) is indeed a minimum. Easily verified that at that minimum, \( KL[q||p] = KL[p||p] = 0 \).

A similar proof holds for continuous densities, using functional derivatives.

**Maxima in \( \mathcal{F} \) correspond to maxima in \( \ell \)**

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

Differentiating the previous expression wrt \( \theta \) again we find

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

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

\[ \theta^* \text{ is a maximum of } \ell. \]

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