## 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:**
  
  \[
  p(x|\theta_{x}, \theta_{y}) = \int dy f_x(x) e^{\phi(\theta_{x}, y)^T T_x(x)} f_y(y) e^{\theta_{y}^T T_y(y)} / Z_{x}(\theta_{x}) Z_{y}(\theta_{y})
  \]

  \[
  \ell(\theta_{x}, \theta_{y}) = \sum_n \log \int dy f_x(x) e^{\phi(\theta_{x}, y)^T T_x(x)} f_y(y) e^{\theta_{y}^T T_y(y)} / Z_{x}(\theta_{x}) Z_{y}(\theta_{y})
  \]

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

## Example: mixture of Gaussians

**Data:** \( \mathcal{X} = \{x_1, \ldots, x_n\} \)

**Latent process:**

\( s_i \overset{i.i.d.}{\sim} \text{Disc}[\pi] \)

**Component distributions:**

\( x_i | (s_i = m) \sim P_m[\theta_m] = \mathcal{N}(\mu_m, \Sigma_m) \)

**Marginal distribution:**

\( P(x) = \sum_{m=1}^{k} \pi_m P_m(x; \theta_m) \)

**Log-likelihood:**

\[
\ell(\{\mu_m\}, \{\Sigma_m\}, \pi) = \sum_{i=1}^{n} \log \sum_{m=1}^{k} \pi_m \frac{1}{\sqrt{2\pi \Sigma_m}} e^{-\frac{1}{2}(x_i - \mu_m)^{T} \Sigma_m^{-1} (x_i - \mu_m)}
\]

## The joint-data likelihood and EM

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

  \[
  \ell(\theta_{x}, \theta_{y}) = \sum_{n} \phi(\theta_{x}, y_n)^T T_{x}(x_n) + \theta_{y}^T \sum_{n} T_{y}(y_n) - \sum_{n} \log Z_{x}(\phi(\theta_{x}, y_n)) - N \log Z_{y}(\theta_{y})
  \]

- Conversely, if we knew \( \theta \), we might easily compute (the posterior over) the values of \( y \).

- **Idea:** update \( \theta \) and (the distribution on) \( y \) in alternation, to reach 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.

Start from arbitrary values of the parameters, and iterate 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.

- Decomposes difficult problems into series of tractable steps.
- An alternative to gradient-based iterative methods.
- No learning rate.
- In ML, the E step is called inference, and the M step learning. In stats, these are often called imputation and inference or estimation.
- Not essential for simple models (like MoGs/FA), though often more efficient than alternatives. Crucial for learning in complex settings.
- Provides a framework for principled approximations.

The lower bound for EM – “free energy”

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

Log-likelihood:

\[
\ell(\theta) = \log P(X|\theta) = \log \int dY P(Y, X|\theta)
\]

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

\[
\ell(\theta) = \log \int dY q(Y) \frac{P(Y, X|\theta)}{q(Y)} \geq \int dY q(Y) \log \frac{P(Y, X|\theta)}{q(Y)} \quad \text{def} \quad \mathcal{F}(q, \theta).
\]

Now,

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

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

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.

\[
\log (\alpha x_1 + (1-\alpha) x_2) \geq \alpha \log (x_1) + (1-\alpha) \log (x_2)
\]

In general:

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

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

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 free-energy lower bound on \( \ell(\theta) \) is a function of \( \theta \) and a distribution \( q \):

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

The EM steps can be re-written:

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

  \[
  q^{(k)}(Y) := \operatorname{argmax}_{q(Y)} \mathcal{F}(q(Y), \theta^{(k-1)})
  \]

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

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

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

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

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

The second term is the Kullback-Leibler divergence.

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

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

So, the E step sets

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

[inference / imputation]

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

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

EM Never Decreases the Likelihood

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

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

- 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).

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

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

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

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

Single data point \( x_1 = 0.3 \).

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

- An iterative algorithm that finds (local) maxima of the likelihood of a latent variable model.

\[ \ell(\theta) = \log P(X|\theta) = \log \int dY P(X|Y, \theta)P(Y|\theta) \]

- Increases a variational lower bound on the likelihood by coordinate ascent.

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

- **E step:**

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

- **M step:**

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

- After E-step \( \mathcal{F}(q, \theta) = \ell(\theta) \Rightarrow \) maximum of free-energy is maximum of likelihood.

EM for MoGs

- Evaluate responsibilities

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

- Update parameters

\[ \mu_m \leftarrow \frac{\sum_i r_{mi}x_i}{\sum_i r_{mi}} \]

\[ \Sigma_m \leftarrow \frac{\sum_i r_{mi}(x_i - \mu_m)(x_i - \mu_m)^T}{\sum_i r_{mi}} \]

\[ \pi_m \leftarrow \frac{\sum_i r_{mi}}{N} \]

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).

In fact, immediately after an E step

\[ \frac{\partial}{\partial \theta} \langle \log P(Y, X|\theta) \rangle_{q^{(k)}(Y)} \leq \text{KL}[q^{(k)}(\theta)||P(Y|X, \theta)] \]

[cf. mixture gradients from last lecture.] 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 \( F \) 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. 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:

\[ p(x|\theta) = \sum_{m=1}^{k} \pi_m \phi(x|m, \theta) = \sum_{m=1}^{k} \frac{\pi_m}{\sigma_m} \exp \left\{ -\frac{1}{2\sigma_m}(x - \mu_m)^2 \right\}, \]

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 \( s \) indicates which component generated observation \( x \).

The E-step computes the posterior for \( s \) given the current parameters:

\[ q(s|\theta) = p(s|x, \theta) \propto p(x|s, \theta)p(s|\theta) \]

\[ r_m \overset{\text{def}}{=} q(s = m) \propto \frac{\pi_m}{\sigma_m} \exp \left\{ -\frac{1}{2\sigma_m^2}(x - \mu_m)^2 \right\} \text{ (responsibilities)} \leftarrow \langle \delta_{s = m} \rangle_q \]

with the normalization such that \( \sum_m r_m = 1 \).
The Gaussian mixture model (M-step)

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

\[ E = \langle \log p(x, s|\theta) \rangle_{q(s)} = \sum_{i,m} q(s_i) \log p(s_i|\theta) \times p(x_i|s_i, \theta) \]

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

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

\[ \frac{\partial}{\partial \mu_m} E = \sum_i \tau_m (x_i - \mu_m) = 0 \Rightarrow \mu_m = \frac{\sum_i \tau_m x_i}{\sum_i \tau_m}, \]

\[ \frac{\partial}{\partial \sigma_m} E = \sum_i \tau_m \left[ -\frac{1}{\sigma_m} \sigma_m^{-1} + \frac{(x_i - \mu_m)^2}{\sigma_m^2} \right] = 0 \Rightarrow \sigma_m^2 = \frac{\sum_i \tau_m (x_i - \mu_m)^2}{\sum_i \tau_m}, \]

\[ \frac{\partial}{\partial \pi_m} E = \sum_i \tau_m \frac{1}{\pi_m}, \quad \frac{\partial E}{\partial \pi_m} + \lambda = 0 \Rightarrow \pi_m = \frac{1}{\lambda} \sum_i \tau_m, \]

where \( \lambda \) is a Lagrange multiplier ensuring that the mixing proportions sum to unity.

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: \( q_n(y_n) = \log p(y_n|x_n, \theta) \times p(y_n|x_n|\theta)/p(x_n|\theta) \)

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

\[ p(y_n|x_n, \theta) = p(y_n|x_n, \theta) = p(y_n|x_n, \theta) \]

\[ = (2\pi)^{-\frac{d}{2}} \exp\left\{-\frac{1}{2} y_n^T \Sigma^{-1} y_n - \frac{1}{2} (x_n - \Lambda y_n)^T \Sigma^{-1} (x_n - \Lambda y_n) \right\} \]

\[ = c \times \exp\left\{-\frac{1}{2} y_n^T \Sigma^{-1} y_n - \frac{1}{2} (x_n - \Lambda y_n)^T \Sigma^{-1} (x_n - \Lambda y_n) \right\} \]

\[ = c' \times \exp\left\{-\frac{1}{2} y_n^T \Sigma^{-1} y_n - 2y_n^T \Sigma^{-1} \Lambda \sigma_n + \sigma_n^T \Sigma^{-1} \sigma_n \right\} \]

\[ \text{So } \Sigma = (I + \Sigma^{-1} \Lambda)^{-1} = I - \beta \Lambda \text{ and } \mu_n = \Sigma^{-1} \Lambda \sigma_n = \beta \sigma_n. \]

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 \( \theta_{t+1} \) by maximising \( \mathcal{F} = \sum_n \langle \log p(y_n|\theta) + \log p(x_n|y_n, \theta) \rangle_{q_n(y_n)} + c \)

\[ \log p(y_n|\theta) + \log p(x_n|y_n, \theta) \]

\[ = c - \frac{1}{2} y_n^T y_n - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x_n - \Lambda y_n)^T \Sigma^{-1} (x_n - \Lambda y_n) \]

\[ = c' - \frac{1}{2} \log |\Sigma| - 2y_n^T \Sigma^{-1} \sigma_n + y_n^T \Sigma^{-1} \Lambda \sigma_n \]

\[ = \frac{1}{2} \log |\Sigma| - 2y_n^T \Sigma^{-1} \sigma_n + y_n^T \Sigma^{-1} \Lambda \sigma_n + \text{Tr} \left[ \Sigma^{-1} \Lambda (\sigma_n \sigma_n^T + \Sigma) \right] \]

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

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

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

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

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

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

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

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

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

(Squared residuals)

Note: We should actually only take derivatives wrt \(\Psi_m\) since \(\Psi\) is diagonal.

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

**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:

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

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

\[
= \theta^T \langle T(z) \rangle_{q|y} - \log Z(\theta) + \text{const wrt} \theta
\]

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

We also have:

\[
\frac{\partial}{\partial \theta} \log Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} Z(\theta)
\]

\[
= \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} \int f(z) \exp(\theta^T T(z)) dz
\]

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

Thus, the **M step** solves:

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

**Mixtures of Factor Analysers**

Simultaneous clustering and dimensionality reduction.

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

where \(\pi_k\) is the mixing proportion for FA \(k\), \(\mu_k\) is its centre, \(\Lambda_k\) is its “factor loading matrix”, and \(\Sigma\) 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 \(y_{s_n, k} \in \mathbb{R}^Q\)

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

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 \(\langle \delta_{s_n=m} y_n y_n^T \rangle\).


**EM for exponential family mixtures**

To derive EM formally for models with discrete latents (including mixtures) it is useful to introduce an indicator vector \(s\) in place of the discrete \(s\).

\[
s_i = m \iff s_i = [0, \ldots, 1, \ldots, 0] \text{ mth position}
\]

Collecting the \(M\) component distributions’ natural params into a matrix \(\Theta = [\theta_m]\):

\[
\log P(x, S) = \sum_m \left[ \log \pi^T s_m \Theta^T T(x) - s_m^T \log Z(\Theta) \right] + \text{const}
\]

where \(\log Z(\Theta)\) collects the log-normalisers for all components into an \(M\)-element vector.

Then, the expected sufficient statistics (E-step) are:

\[
\sum_m \langle s_m \rangle_q \quad \text{(responsibilities \(r_m\))}
\]

\[
\sum_m T(x) \langle s_m \rangle_q \quad \text{(responsibility-weighted sufficient stats)}
\]

And maximisation of the expected log-joint (M-step) gives:

\[
\pi^{(k+1)} \propto \sum_m \langle s_m \rangle_q \left( \sum_i T(x) \langle [s_m]_i \rangle_q \right) / \left( \sum_i \langle [s_m]_i \rangle_q \right)
\]
EM for MAP

What if we have a prior?

\[ p(z|\theta) = f(z) \exp(\theta^T T(z))/Z(\theta) \quad p(\theta) = F(\nu, \tau) \exp(\theta^T \tau)/Z(\theta)^\nu \]

Augment the free energy by adding the log prior:

\[ F_{\text{MAP}}(q, \theta) = \int q(Y) \log p(Y, \chi; \theta) d\chi + \log p(\chi|\theta) + \log P(\theta) \]

\[ = \int q(Y) [\alpha (T(z) + \tau) - (N + \nu) \log Z(\theta)] d\chi + \text{const wrt } \theta \]

\[ = \theta^T (T(z) q(\chi) + \tau) - (N + \nu) \log Z(\theta) + \text{const wrt } \theta \]

So, the expected sufficient statistics in the E step are unchanged.

Thus, after an E-step the augmented free-energy equals the log-joint, and so free-energy maxima are log-joint maxima (i.e. MAP values).

Can we find posteriors? Only approximately – we’ll return to this later as “Variational Bayes”.

Proof of the Matrix Inversion Lemma

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

Need to prove:

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

Expand:

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

Regroup:

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

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

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

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

References

  http://www.jstor.org/stable/2984875

  A view of the EM algorithm that justifies incremental, sparse, and other variants.

  Optimization with EM and expectation-conjugate-gradient.
  In ICML (pp. 672-679).

  The EM Algorithm for Mixtures of Factor Analyzers.

KL(q(x)||p(x)) ≥ 0, with equality iff ∀x : p(x) = q(x)

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

\[ \text{KL}[q||p] = \sum_j q_j \log \frac{q_j}{p_j}. \]

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

\[ E \overset{\text{def}}{=} \text{KL}[q||p] + \lambda (1 - \sum_j q_j) = \sum_j q_j \log \frac{q_j}{p_j} + \lambda (1 - \sum_j q_j) \]

Find conditions for stationarity

\[ \frac{\partial E}{\partial q_j} = - \log q_j - \log p_j + 1 - \lambda \Rightarrow q_j = \rho \exp(\lambda - 1) \]

\[ \frac{\partial E}{\partial \lambda} = 1 - \sum_j q_j = 0 \Rightarrow \sum_j q_j = 1 \]

Check sign of curvature (Hessian):

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

so unique stationary point \( q_j = \rho_j \) is indeed a minimum. Easily verified that at that minimum, \( \text{KL}[q||p] = \text{KL}[p||p] = 0 \).

A similar proof holds for continuous densities, using functional derivatives.
Fixed Points of EM are Stationary Points in $\ell$

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

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

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

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

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

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

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

$$\frac{d}{d\theta} \ell(\theta) \bigg|_{\theta^*} = 0$$

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

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

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

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

$\theta^*$ is a maximum of $\ell$.

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