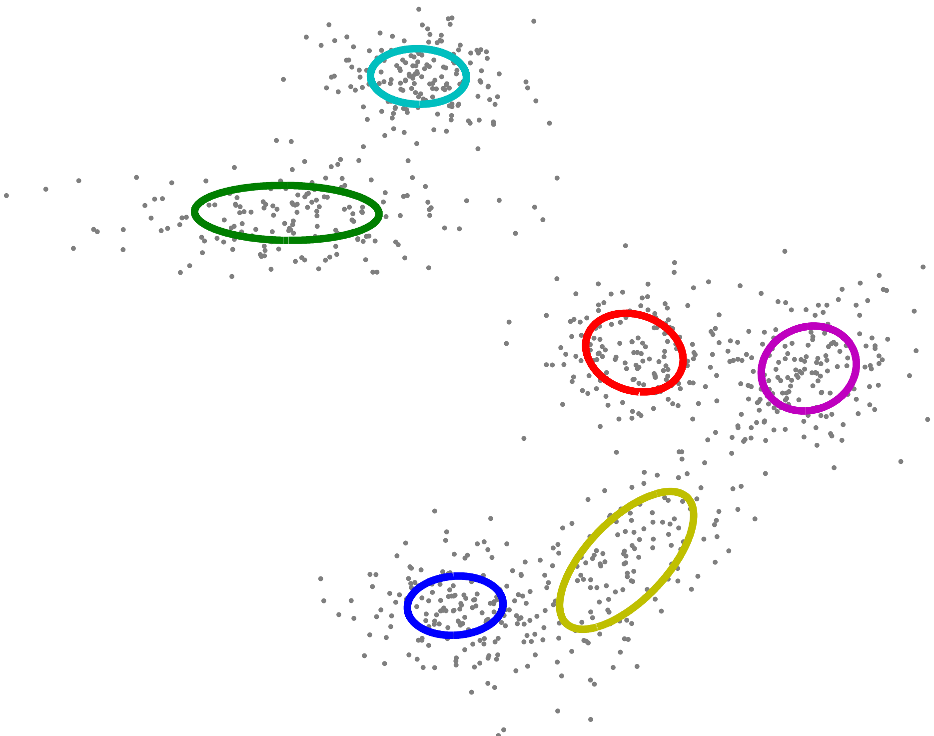
Week 3: The EM algorithm

Maneesh Sahani
`maneesh@gatsby.ucl.ac.uk`

Gatsby Computational Neuroscience Unit
University College London

Term 1, Autumn 2005

Mixtures of Gaussians



Data: $\mathcal{Y} = \{\mathbf{y}_1 \ldots \mathbf{y}_N\}$

Latent process:

$$s_i \overset{\text{iid}}{\sim} \ \mathit{Discrete}[\boldsymbol{\pi}]$$

Component distributions:

$$\mathbf{y}_i \mid (s_i = m) \sim \mathcal{P}_m[\theta_m] = \mathcal{N}\left[\boldsymbol{\mu}_m; \Sigma_m\right]$$

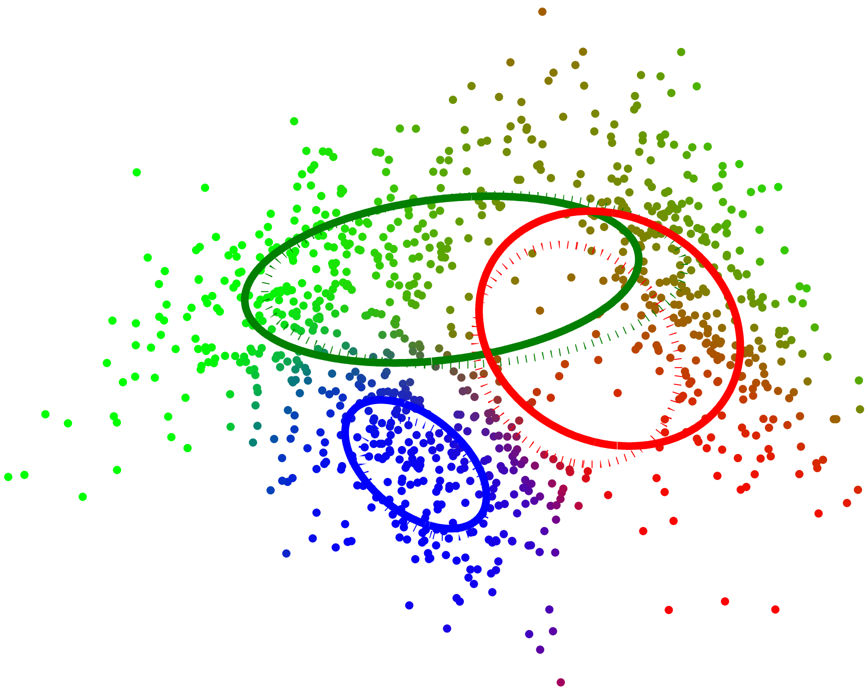
Marginal distribution:

$$P(\mathbf{y}_i) = \sum_{m=1}^{k} \pi_m P_m(\mathbf{y}; \theta_m)$$

Log-likelihood:

$$\log p(\mathcal{Y} \mid \{\boldsymbol{\mu}_m\}, \{\Sigma_m\}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \log \sum_{m=1}^{k} \pi_m \left|2\pi\Sigma_m\right|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_m)^{\mathsf{T}} \Sigma_m^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_m)\right]$$

EM for MoGs



- Evaluate responsibilities

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

- Update parameters

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

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

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

The Expectation Maximisation (EM) algorithm

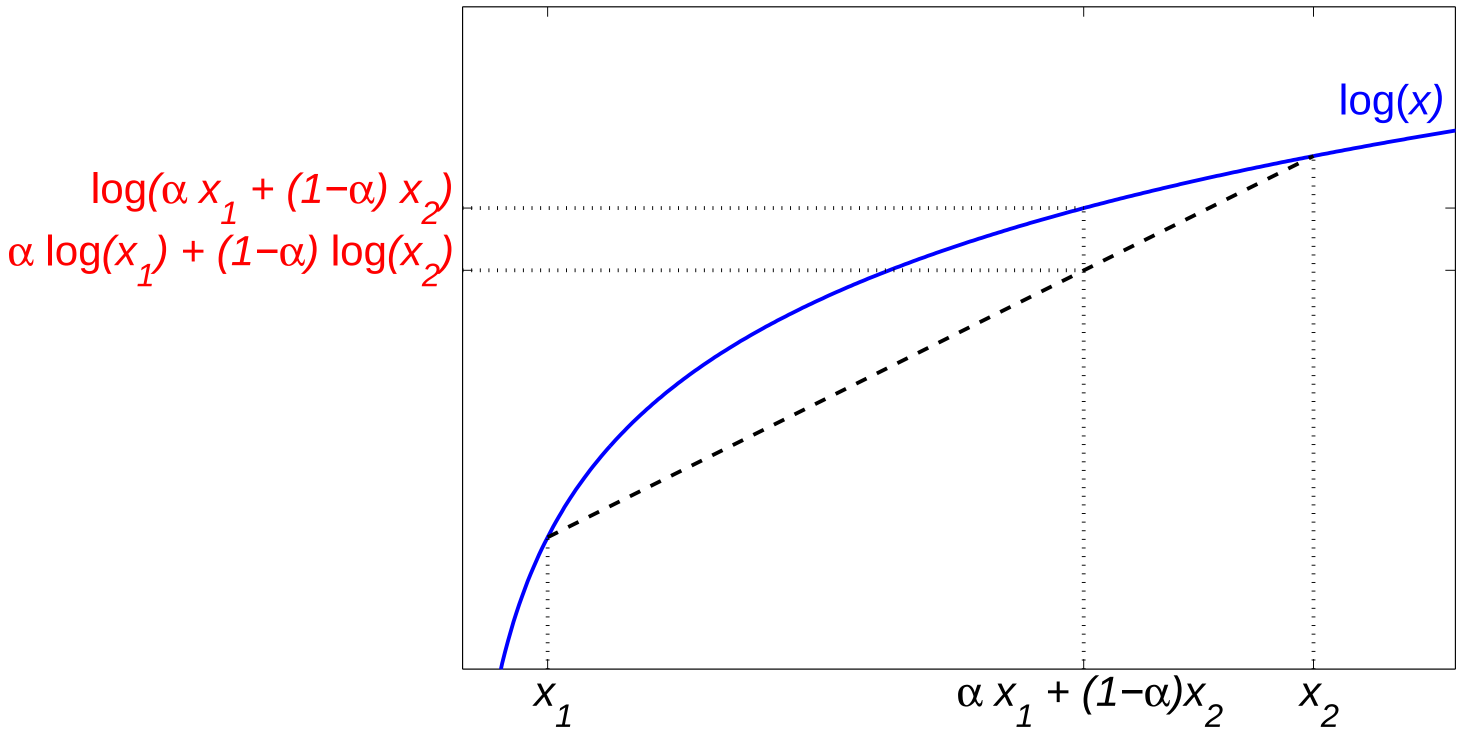
The EM algorithm finds a (local) maximum of a latent variable model likelihood. It starts from arbitrary values of the parameters, and iterates two steps:

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 $\mathcal{Y} = \{\mathbf{y}_i\}$; Latent variables $\mathcal{X} = \{\mathbf{x}_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

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

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:

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

Now,

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

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

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

The E and M steps of EM

The lower bound on the log likelihood is given by:

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

EM alternates between:

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

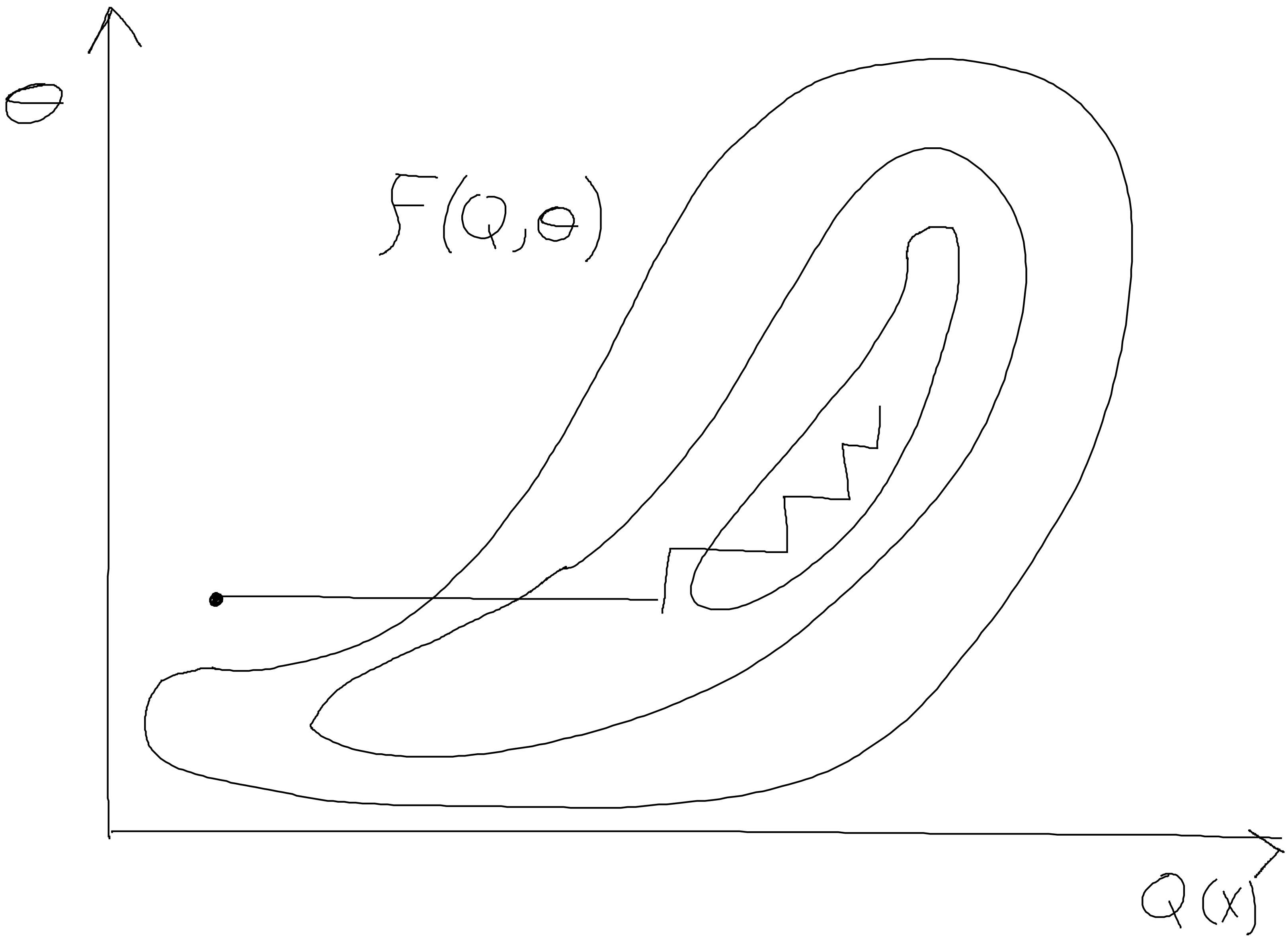
$$q^{(k)}(\mathcal{X}) := \underset{q(\mathcal{X})}{\text{argmax}}\ \mathcal{F}\big(q(\mathcal{X}), \theta^{(k-1)}\big).$$

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

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

The second equality comes from the fact that the entropy of $q(\mathcal{X})$ does not depend directly on θ.

EM as Coordinate Ascent in $\mathcal{F}$



The E Step

The free energy can be re-written

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

The second term is the Kullback-Leibler divergence.

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

But **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:

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

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

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

We then take partial derivatives and set to zero:

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

Why **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:

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

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

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

A similar proof holds for **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:

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

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

Fixed Points of EM are Stationary Points in $\mathcal{L}$

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

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

Now,

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

so,

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

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

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

So, EM converges to a stationary point of $\mathcal{L}(\theta)$.

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

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

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

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

θ^* is a maximum of $\mathcal{L}$.

The Gaussian mixture model (E-step)

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

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

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 observation y_i belongs to.
The E-step computes the posterior for s_i given the current parameters:

$$\begin{aligned} q(s_i) \ &= \ p(s_i|y_i,\theta) \propto p(y_i|s_i,\theta)p(s_i|\theta) \\ r_{im} \stackrel{\text{def}}{=} q(s_i=m) \ &\propto \ \frac{\pi_m}{\sigma_m} \exp\Big\{-\frac{1}{2\sigma_m^2}(y_i-\mu_m)^2\Big\} \qquad \text{(responsibilities)} \end{aligned}$$

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):

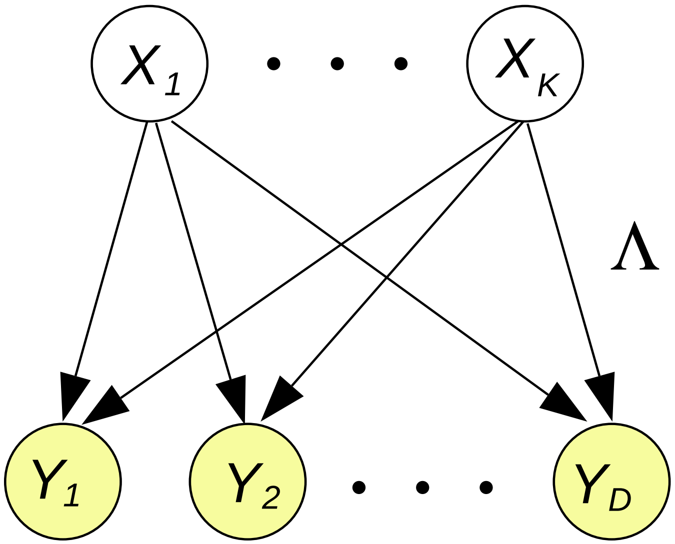
$$
\begin{aligned}
E = \langle \log p(y, s|\theta) \rangle_{q(s)} &= \sum q(s) \log[p(s|\theta)\; p(y|s,\theta)] \\
&= \sum_{i,m} r_{im} \Big[\log \pi_m - \log \sigma_m - \frac{1}{2\sigma_m^2}(y_i - \mu_m)^2 \Big].
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{\partial E}{\partial \mu_m} &= \sum_i r_{im} \frac{(y_i - \mu_m)}{2\sigma_m^2} = 0 \Rightarrow \quad \mu_m = \frac{\sum_i r_{im} y_i}{\sum_i r_{im}}, \\
\frac{\partial E}{\partial \sigma_m} &= \sum_i r_{im} \Big[-\frac{1}{\sigma_m} + \frac{(y_i - \mu_m)^2}{\sigma_m^3} \Big] = 0 \Rightarrow \quad \sigma_m^2 = \frac{\sum_i r_{im} (y_i - \mu_m)^2}{\sum_i r_{im}}, \\
\frac{\partial E}{\partial \pi_m} &= \sum_i r_{im} \frac{1}{\pi_m}, \qquad \frac{\partial E}{\partial \pi_m} + \lambda = 0 \Rightarrow \quad \pi_m = \frac{1}{n} \sum_i r_{im},
\end{aligned}
$$

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

Factor Analysis



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

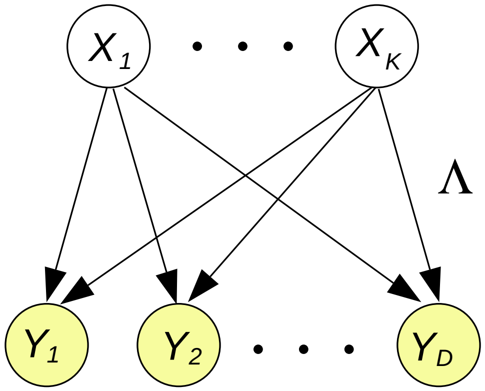
- x_k are independent $\mathcal{N}(0, 1)$ Gaussian factors
- ϵ_d are independent $\mathcal{N}(0, \Psi_{dd})$ Gaussian noise
- $K < D$

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

where Λ is a $D \times K$ matrix, and Ψ 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**:
$p(\mathbf{y}|\theta) = \int p(\mathbf{x}|\theta)p(\mathbf{y}|\mathbf{x},\theta)d\mathbf{x} = \mathcal{N}(0, \Lambda\Lambda^\top + \Psi)$
Model parameters: $\theta = \{\Lambda, \Psi\}$.

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

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

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

The E step for Factor Analysis

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

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

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

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

The M step for Factor Analysis

M step: Find θ_{t+1} maximising $\mathcal{F} = \sum_n \int q_n(\mathbf{x}) \left[\log p(\mathbf{x}|\theta) + \log p(\mathbf{y}_n|\mathbf{x},\theta)\right] d\mathbf{x} + \mathrm{c}$

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

Taking expectations over $q_n(\mathbf{x})\ldots$

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

Note that we don't need to know everything about q, just the expectations of $\mathbf{x}$ and $\mathbf{x}\mathbf{x}^\top$ under q (i.e. the expected sufficient statistics).

The M step for Factor Analysis (cont.)

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

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

$$\frac{\partial\mathcal{F}}{\partial\Lambda} = \Psi^{-1}\sum_n \mathbf{y}_n\mu_n{}^\top - \Psi^{-1}\Lambda\left(N\Sigma + \sum_n \mu_n\mu_n{}^\top\right) = 0$$

$$\hat{\Lambda} = \left(\sum_n \mathbf{y}_n\mu_n{}^\top\right)\left(N\Sigma + \sum_n \mu_n\mu_n{}^\top\right)^{-1}$$

$$\frac{\partial\mathcal{F}}{\partial\Psi^{-1}} = \frac{N}{2}\Psi - \frac{1}{2}\sum_n\left[\mathbf{y}_n\mathbf{y}_n{}^\top - \Lambda\mu_n\mathbf{y}_n{}^\top - \mathbf{y}_n\mu_n{}^\top\Lambda^\top + \Lambda(\mu_n\mu_n{}^\top + \Sigma)\Lambda^\top\right]$$

$$\hat{\Psi} = \frac{1}{N}\sum_n\left[\mathbf{y}_n\mathbf{y}_n{}^\top - \Lambda\mu_n\mathbf{y}_n{}^\top - \mathbf{y}_n\mu_n{}^\top\Lambda^\top + \Lambda(\mu_n\mu_n{}^\top + \Sigma)\Lambda^\top\right]$$

$$\hat{\Psi} = \Lambda\Sigma\Lambda^\top + \frac{1}{N}\sum_n(\mathbf{y}_n - \Lambda\mu_n)(\mathbf{y}_n - \Lambda\mu_n)^\top \qquad \text{(squared residuals)}$$

Note: we should actually only take derivarives w.r.t. Ψ_{dd} since Ψ 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* $\mathcal{F}$ wrt θ rather than maximize. (Dempster, Laird and Rubin (1977) call this the generalized EM, or GEM, algorithm).

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. 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: $\theta^{(k)} := \operatorname*{argmax}_\theta \; \mathcal{F}(q, \theta)$

$$\begin{aligned} \mathcal{F}(q,\theta) &= \int q(\mathbf{x}) \log p(\mathbf{x}, \mathbf{y}|\theta) d\mathbf{x} - \mathcal{H}(q) \\ &= \int q(\mathbf{x}) [\theta^\top s(\mathbf{z}) - \log \alpha(\theta)] d\mathbf{x} + \text{const} \end{aligned}$$

It is easy to verify that: $\dfrac{\partial \log \alpha(\theta)}{\partial \theta} = E[s(\mathbf{z})|\theta]$

Therefore, M step solves: $\dfrac{\partial \mathcal{F}}{\partial \theta} = E_{q(\mathbf{x})}[s(\mathbf{z})] - E[s(\mathbf{z})|\theta] = 0$

Mixtures of Factor Analysers

Simultaneous clustering and dimensionality reduction.

$$p(\mathbf{y}|\theta) = \sum_k \pi_k \, \mathcal{N}(\mu_k, \Lambda_k \Lambda^\top{}_k + \Psi)$$

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

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

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

E step: Infer joint distribution of latent variables, $p(\mathbf{x}_n, s_n|\mathbf{y}_n, \theta)$

M step: Maximize $\mathcal{F}$ with respect to θ.

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:

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

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:

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