

## Probabilistic & Unsupervised Learning

### Expectation Maximisation

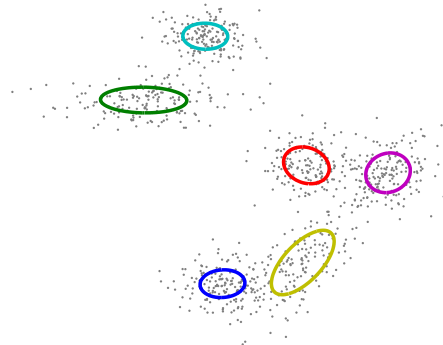
Maneesh Sahani

maneesh@gatsby.ucl.ac.uk

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

Term 1, Autumn 2014

### Example: mixture of Gaussians



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

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

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

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

Log-likelihood:

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

### Log-likelihoods

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

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

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

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

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

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

### 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:

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

- Conversely, if we knew  $\boldsymbol{\theta}$ , 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?

## 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:

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

Now,

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

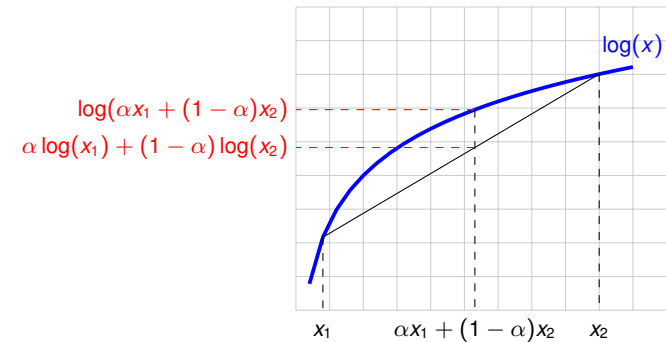
where  $\mathbf{H}[q]$  is the entropy of  $q(\mathcal{Y})$ .

So:

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

## 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  $\{x_i > 0\}$ ):

For probability measure  $\alpha$  and **concave**  $f$

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

$$f(\mathbb{E}_\alpha[x]) \geq \mathbb{E}_\alpha[f(x)]$$

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 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:

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

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

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

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

## The E Step

The free energy can be re-written

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

The second term is the Kullback-Leibler divergence.

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

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

So, the E step sets

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

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

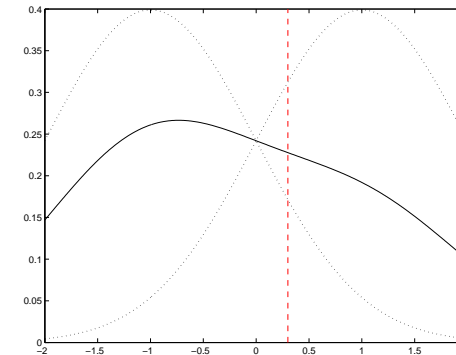
## Coordinate Ascent in $\mathcal{F}$ (Demo)

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

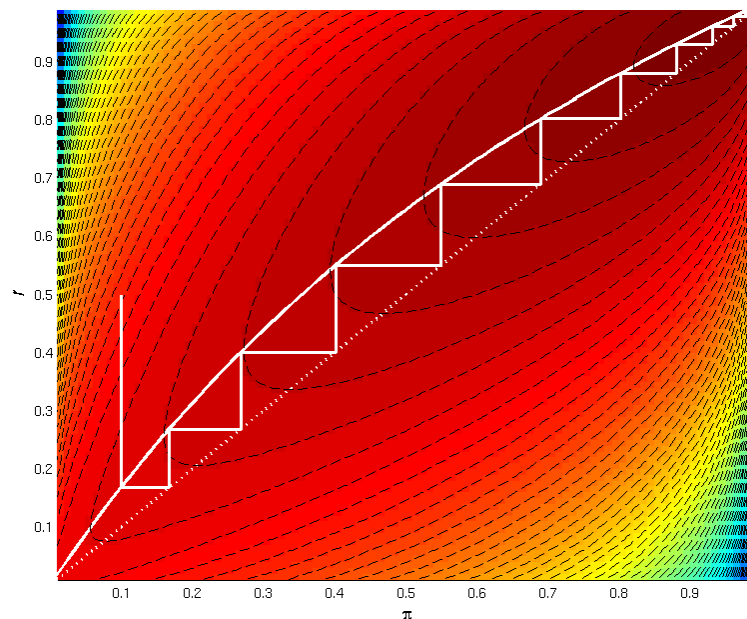
$$\begin{aligned}s &\sim \text{Bernoulli}[\pi] \\ x|s = 0 &\sim \mathcal{N}[-1, 1] \quad x|s = 1 \sim \mathcal{N}[1, 1].\end{aligned}$$

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:

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

- ▶ 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).

## EM Summary

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

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

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

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

- **E step:**

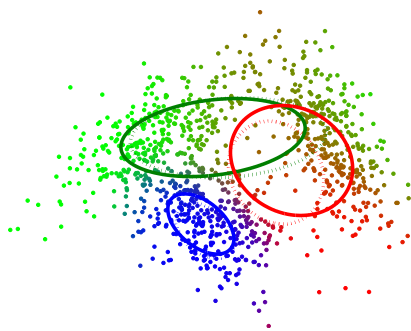
$$q^{(k)}(\mathcal{Y}) := \operatorname{argmax}_{q(\mathcal{Y})} \mathcal{F}(q(\mathcal{Y}), \theta^{(k-1)}) = P(\mathcal{Y}|\mathcal{X}, \theta^{(k-1)})$$

- **M step:**

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

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

## EM for MoGs



- Evaluate responsibilities

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

- Update parameters

$$\begin{aligned} \mu_m &\leftarrow \frac{\sum_i r_{im} \mathbf{x}_i}{\sum_i r_{im}} \\ \Sigma_m &\leftarrow \frac{\sum_i r_{im} (\mathbf{x}_i - \mu_m)(\mathbf{x}_i - \mu_m)^T}{\sum_i r_{im}} \\ \pi_m &\leftarrow \frac{\sum_i r_{im}}{N} \end{aligned}$$

## 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

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

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

**Partial E steps:** We can also just *increase*  $\mathcal{F}$  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:

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

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:

$$\begin{aligned} q(s_i) &= p(s_i|x_i, \theta) \propto p(x_i|s_i, \theta) p(s_i|\theta) \\ r_{im} &\stackrel{\text{def}}{=} q(s_i=m) \propto \frac{\pi_m}{\sigma_m} \exp \left\{ -\frac{1}{2\sigma_m^2} (x_i - \mu_m)^2 \right\} \quad (\text{responsibilities}) \quad \leftarrow \langle \delta_{s_i=m} \rangle_q \end{aligned}$$

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):

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

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

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

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

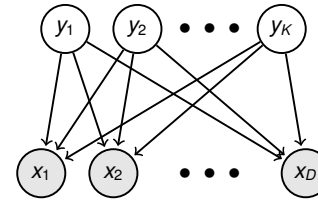
**Tactic:** write  $p(\mathbf{y}_n, \mathbf{x}_n|\theta)$ , consider  $\mathbf{x}_n$  to be fixed. What is this as a function of  $\mathbf{y}_n$ ?

$$\begin{aligned} p(\mathbf{y}_n, \mathbf{x}_n) &= p(\mathbf{y}_n)p(\mathbf{x}_n|\mathbf{y}_n) \\ &= (2\pi)^{-\frac{K}{2}} \exp\left\{-\frac{1}{2}\mathbf{y}_n^T \mathbf{y}_n\right\} |2\pi\Psi|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_n - \Lambda\mathbf{y}_n)^T \Psi^{-1}(\mathbf{x}_n - \Lambda\mathbf{y}_n)\right\} \\ &= c \times \exp\left\{-\frac{1}{2}[\mathbf{y}_n^T \mathbf{y}_n + (\mathbf{x}_n - \Lambda\mathbf{y}_n)^T \Psi^{-1}(\mathbf{x}_n - \Lambda\mathbf{y}_n)]\right\} \\ &= c' \times \exp\left\{-\frac{1}{2}[\mathbf{y}_n^T (I + \Lambda^T \Psi^{-1} \Lambda) \mathbf{y}_n - 2\mathbf{y}_n^T \Lambda^T \Psi^{-1} \mathbf{x}_n]\right\} \\ &= c'' \times \exp\left\{-\frac{1}{2}[\mathbf{y}_n^T \Sigma^{-1} \mathbf{y}_n - 2\mathbf{y}_n^T \Sigma^{-1} \boldsymbol{\mu}_n + \boldsymbol{\mu}_n^T \Sigma^{-1} \boldsymbol{\mu}_n]\right\} \end{aligned}$$

So  $\Sigma = (I + \Lambda^T \Psi^{-1} \Lambda)^{-1} = I - \beta \Lambda$  and  $\boldsymbol{\mu}_n = \Sigma \Lambda^T \Psi^{-1} \mathbf{x}_n = \beta \mathbf{x}_n$ . Where  $\beta = \Sigma \Lambda^T \Psi^{-1}$ .

Note that  $\boldsymbol{\mu}_n$  is a linear function of  $\mathbf{x}_n$  and  $\Sigma$  does not depend on  $\mathbf{x}_n$ .

## EM for Factor Analysis



The model for  $\mathbf{x}$ :

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

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  $\theta_{t+1}$  that maximises  $\mathcal{F}(q, \theta)$ :

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

## The M step for Factor Analysis

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

$$\begin{aligned} \log p(\mathbf{y}_n|\theta) + \log p(\mathbf{x}_n|\mathbf{y}_n, \theta) &= c - \frac{1}{2}\mathbf{y}_n^T \mathbf{y}_n - \frac{1}{2} \log |\Psi| - \frac{1}{2}(\mathbf{x}_n - \Lambda\mathbf{y}_n)^T \Psi^{-1}(\mathbf{x}_n - \Lambda\mathbf{y}_n) \\ &= c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[ \mathbf{x}_n^T \Psi^{-1} \mathbf{x}_n - 2\mathbf{x}_n^T \Psi^{-1} \Lambda \mathbf{y}_n + \mathbf{y}_n^T \Lambda^T \Psi^{-1} \Lambda \mathbf{y}_n \right] \\ &= c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[ \mathbf{x}_n^T \Psi^{-1} \mathbf{x}_n - 2\mathbf{x}_n^T \Psi^{-1} \Lambda \boldsymbol{\mu}_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda \boldsymbol{\mu}_n \boldsymbol{\mu}_n^T + \Sigma \right] \right] \end{aligned}$$

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

$$= c' - \frac{1}{2} \log |\Psi| - \frac{1}{2} \left[ \mathbf{x}_n^T \Psi^{-1} \mathbf{x}_n - 2\mathbf{x}_n^T \Psi^{-1} \Lambda \boldsymbol{\mu}_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda (\boldsymbol{\mu}_n \boldsymbol{\mu}_n^T + \Sigma) \right] \right]$$

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

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

$$\mathcal{F} = c' - \frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n \left[ \mathbf{x}_n^T \Psi^{-1} \mathbf{x}_n - 2 \mathbf{x}_n^T \Psi^{-1} \Lambda \boldsymbol{\mu}_n + \text{Tr} \left[ \Lambda^T \Psi^{-1} \Lambda (\boldsymbol{\mu}_n \boldsymbol{\mu}_n^T + \Sigma) \right] \right]$$

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

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

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

As  $\Sigma \rightarrow 0$  these become the equations for ML linear regression

## EM for exponential families

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

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

(with  $Z(\theta) = \int f(\mathbf{z}) \exp\{\theta^T \mathbf{T}(\mathbf{z})\} d\mathbf{z}$ ) but whose marginal  $p(\mathbf{x}) \notin \text{ExpFam}$ .

The free energy dependence on  $\theta$  is given by:

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

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

We also have:

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

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

## Mixtures of Factor Analysers

Simultaneous clustering and dimensionality reduction.

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

where  $\pi_k$  is the mixing proportion for FA  $k$ ,  $\boldsymbol{\mu}_k$  is its centre,  $\Lambda_k$  is its “factor loading matrix”, and  $\Psi$  is a common sensor noise model.  $\theta = \{\{\pi_k, \boldsymbol{\mu}_k, \Lambda_k\}_{k=1 \dots K}, \Psi\}$

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

- ▶ A discrete indicator variable  $s_n \in \{1, \dots, K\}$
- ▶ For each factor analyzer, a continuous 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  $\langle \delta_{s_n=m} \mathbf{y}_n \mathbf{y}_n^T \rangle$ .

**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>

## 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

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

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:

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

Regroup:

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

## Fixed Points of EM are Stationary Points in $\ell$

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

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

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

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

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

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

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

## $\mathbf{KL}[q(x)||p(x)] \geq 0$ , 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

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

Check sign of curvature (Hessian):

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

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

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

## Maxima in $\mathcal{F}$ correspond to maxima in $\ell$

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

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

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

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

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