Statistical Approaches to Learning and Discovery

The EM Algorithm

Zoubin Ghahramani & Teddy Seidenfeld
`zoubin@cs.cmu.edu` & `teddy@stat.cmu.edu`

CALD / CS / Statistics / Philosophy
Carnegie Mellon University
Spring 2002

The Expectation Maximization (EM) algorithm

Given a set of observed (visible) variables V, a set of unobserved (hidden / latent / missing) variables H, and model parameters θ, optimize the log likelihood:

$$\mathcal{L}(\theta) = \log p(V|\theta) = \log \int p(H, V|\theta) dH, \tag{1}$$

where we have written the marginal for the visibles in terms of an integral over the joint distribution for hidden and visible variables.

Using *Jensen's inequality* for any distribution of hidden states $q(H)$ we have:

$$\mathcal{L} = \log \int q(H) \frac{p(H, V|\theta)}{q(H)} dH \geq \int q(H) \log \frac{p(H, V|\theta)}{q(H)} \, dH = \mathcal{F}(q, \theta), \tag{2}$$

defining the $\mathcal{F}(q, \theta)$ functional, which is a lower bound on the log likelihood.

In the EM algorithm, we alternately optimize $\mathcal{F}(q, \theta)$ wrt q and θ, and we can prove that this will never decrease $\mathcal{L}$.

The E and M steps of EM

The lower bound on the log likelihood:

$$\mathcal{F}(q,\theta) = \int q(H) \log \frac{p(H,V|\theta)}{q(H)} dH = \int q(H) \log p(H,V|\theta) dH + \mathcal{H}(q), \tag{3}$$

where $\mathcal{H}(q) = -\int q(H) \log q(H) dH$ is the entropy of q. We iteratively alternate:

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

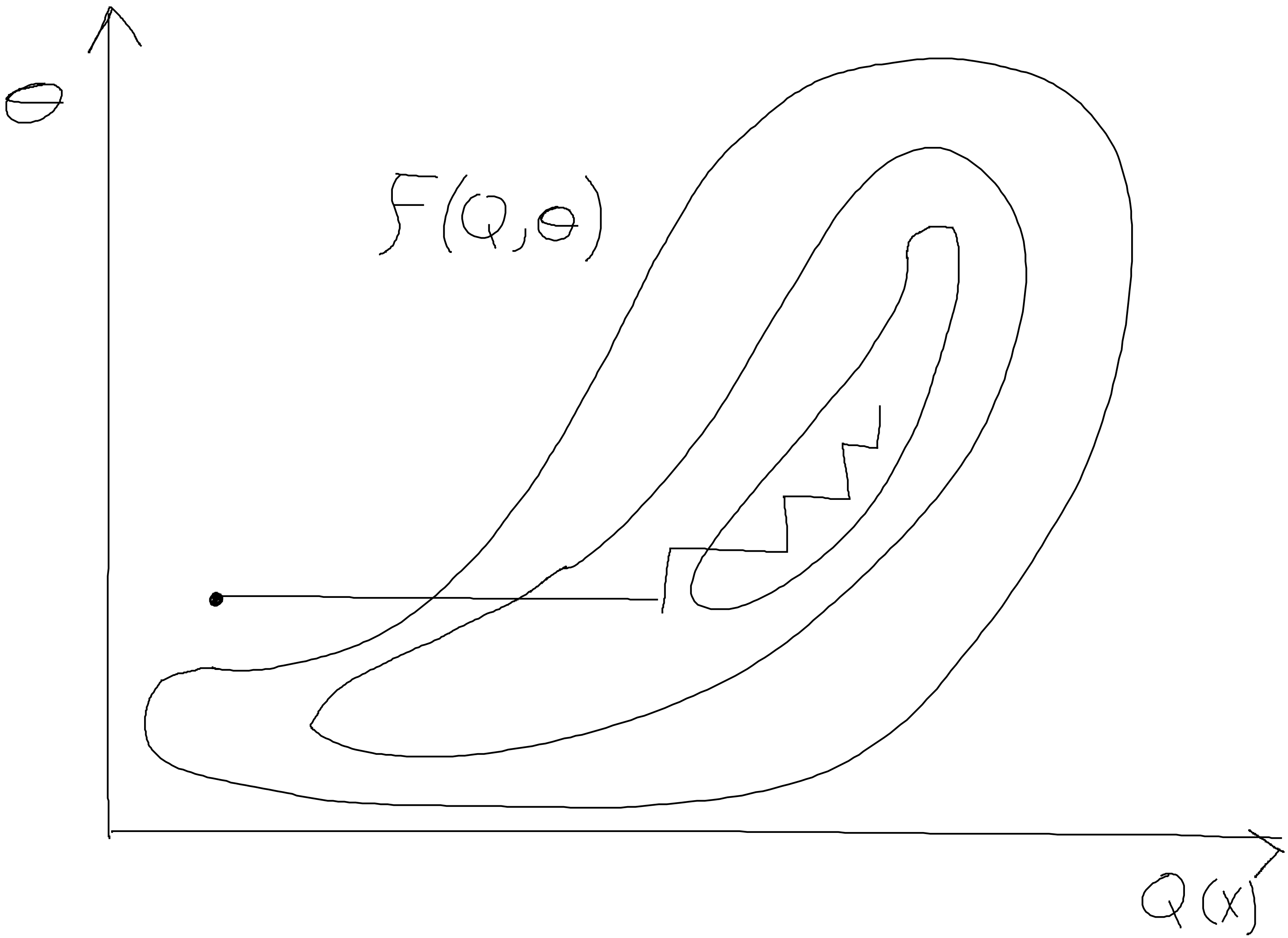
$$q^{(k)}(H) := \underset{q(H)}{\operatorname{argmax}} \; \mathcal{F}\Big(q(H), \theta^{(k-1)}\Big). \tag{4}$$

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

$$\theta^{(k)} := \underset{\theta}{\operatorname{argmax}} \; \mathcal{F}\Big(q^{(k)}(H), \theta\Big) = \underset{\theta}{\operatorname{argmax}} \; \int q^{(k)}(H) \log p(H,V|\theta) dH, \tag{5}$$

which is equivalent to optimizing the expected complete-data likelihood $p(H,V|\theta)$, since the entropy of $q(H)$ does not depend on θ.

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



The EM algorithm never decreases the log likelihood

The difference between the cost functions:

$$\begin{aligned}\mathcal{L}(\theta) - \mathcal{F}(q, \theta) &= \log p(V|\theta) - \int q(H) \log \frac{p(H, V|\theta)}{q(H)} dH \\ &= \log p(V|\theta) - \int q(H) \log \frac{p(H|V, \theta) p(V|\theta)}{q(H)} dH \\ &= -\int q(H) \log \frac{p(H|V, \theta)}{q(H)} dH \;=\; \mathcal{KL}\big(q(H), p(H|V, \theta)\big),\end{aligned}$$

is called the Kullback-Liebler divergence; it is non-negative and only zero if and only if $q(H) = p(H|V, \theta)$ (thus this is the E step). Although we are working with the wrong cost function, the likelihood is still increased in every iteration:

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

where the first equality holds because of the E step, and the first inequality comes from the M step and the final inequality from Jensen. Usually EM converges to a local optimum of $\mathcal{L}$ (although there are exceptions).

The $\mathcal{KL}\big(p(x), q(x)\big)$ **is non-negative and zero iff** $\forall x:\ p(x) = q(x)$

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

$$\mathcal{KL}(p, q) = \sum_i q_i \log \frac{q_i}{p_i}.$$

To find the distribution q which minimizes $\mathcal{KL}(p, q)$ we add a lagrange multiplier to enforce the normalization:

$$E = \mathcal{KL}(p, q) + \lambda(1 - \sum_i q_i) = \sum_i q_i \log \frac{q_i}{p_i} + \lambda(1 - \sum_i q_i).$$

We then take partial derives and set to zero:

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

Why $\mathcal{KL}(p,q)$ is . . .

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 $\mathcal{KL}(p,p) = 0$.

A similar proof can be done for continuous distributions, the partial derivatives being substituted by functional derivatives.

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. (DLR 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 $X = (H, V)$ if it can be written:

$$p(X|\theta) = b(X) \exp\{\theta^\top s(X)\}/\alpha(\theta)$$

where $\alpha(\theta) = \int b(X) \exp\{\theta^\top s(X)\} dX$

E step: $q(H) = p(H|V, \theta)$

M step: $\theta^{(k)} := \underset{\theta}{\operatorname{argmax}} \; \mathcal{F}(q, \theta)$

$$\begin{aligned} \mathcal{F}(q, \theta) &= \int q(H) \log p(H, V|\theta) dH - \mathcal{H}(q) \\ &= \int q(H)[\theta^\top s(X) - \log \alpha(\theta)] dH + \text{const} \end{aligned}$$

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

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

The Gaussian mixture model (E-step)

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

$$p(x|\theta) = \sum_{k=1}^{K} p(H=k|\theta)p(x|H=k,\theta) \propto \sum_{k=1}^{K} \frac{\pi_k}{\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2),$$

where θ is the collection of parameters: means μ_k, variances σ_k^2 and mixing proportions $\pi_k = p(H=k|\theta)$.
The hidden variables $H^{(c)}$ indicate which component observation $x^{(c)}$ belongs to.
In the E-step, compute the posterior for $H^{(c)}$ given the current parameters:

$$\begin{aligned} q(H^{(c)}) &= p(H^{(c)}|x^{(c)},\theta) \propto p(x^{(c)}|H^{(c)},\theta)p(H^{(c)}|\theta) \\ r_k^{(c)} \equiv q(H^{(c)}=k) &\propto \frac{\pi_k}{\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x^{(c)}-\mu_k)^2) \quad \text{(responsibilities)} \end{aligned}$$

with the normalization such that $\sum_k r_k^{(c)} = 1$.

The Gaussian mixture model (M-step)

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

$$E \;=\; \sum q(H)\log[p(H|\theta)\;p(x|H,\theta)] \;=\; \sum_{c,k} r_k^{(c)}\big[\log\pi_k - \log\sigma_k - \frac{1}{2\sigma_k^2}(x^{(c)}-\mu_k)^2\big].$$

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

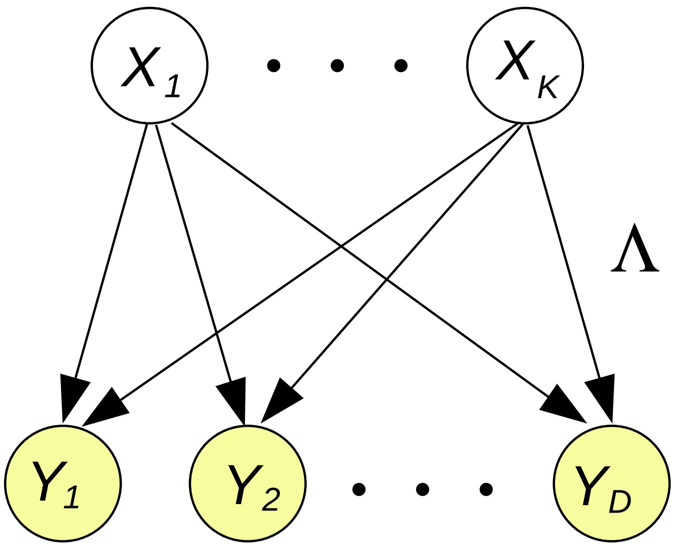
$$\frac{\partial E}{\partial \mu_k} \;=\; \sum_c r_k^{(c)}\frac{(x^{(c)}-\mu_k)}{2\sigma_k^2} = 0 \Rightarrow \quad \mu_k = \frac{\sum_c r_k^{(c)} x^{(c)}}{\sum_c r_k^{(c)}},$$

$$\frac{\partial E}{\partial \sigma_k} \;=\; \sum_c r_k^{(c)}\big[-\frac{1}{\sigma_k} - \frac{(x^{(c)}-\mu_k)}{\sigma_k^3}\big] = 0 \Rightarrow \quad \sigma_k^2 = \frac{\sum_c r_k^{(c)}(x^{(c)}-\mu_k)^2}{\sum_c r_k^{(c)}},$$

$$\frac{\partial E}{\partial \pi_k} \;=\; \sum_c r_k^{(c)}\frac{1}{\pi_k}, \qquad \frac{\partial E}{\partial \pi_k} + \lambda = 0 \Rightarrow \quad \pi_k = \frac{1}{n}\sum_c r_k^{(c)},$$

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

Factor Analysis



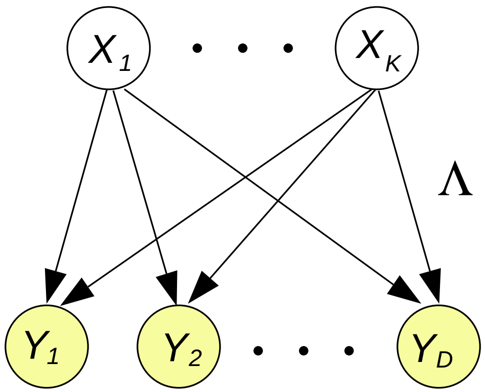
Linear generative model: $y_d = \sum_{k=1}^{K} \Lambda_{dk}\, x_k + \epsilon_d$

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

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

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 $\mathbf{x}$?

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

So $\Sigma = (I + \Lambda^\top\Psi^{-1}\Lambda)^{-1} = I - \beta\Lambda$ and $\mu = \Sigma\Lambda^\top\Psi^{-1}\mathbf{y}_n = \beta\mathbf{y}_n$.
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 $\mathcal{F} = \sum_n \int q_n(\mathbf{x}) \left[\log p(\mathbf{x}|\theta) + \log p(\mathbf{y}_n|\mathbf{x}, \theta)\right] d\mathbf{x} + \mathsf{c}$

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

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

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

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.)

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

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

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

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

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

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

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

When $\Sigma \to 0$ these become the equations for linear regression!

Mixtures of Factor Analysers

Simultaneous clustering and dimensionality reduction.

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

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. $\theta = \{\{\pi_k, \mu_k, \Lambda_k\}_{k=1\dots K}, \Psi\}$
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 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:

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

Readings

- David MacKay's Textbook http://wol.ra.phy.cam.ac.uk/itprnn/book.ps.gz, chapter 21, pages 295-306, draft 2.2.4, August 31, 2001

- Ghahramani, Z. and Hinton, G.E. (1996) The EM Algorithm for Mixtures of Factor Analyzers. University of Toronto Technical Report CRG-TR-96-1 http://www.gatsby.ucl.ac.uk/∼zoubin/papers/tr-96-1.ps.gz

- Minka, T. Tutorial on linear algebra. http://www-white.media.mit.edu/∼tpminka/papers/matrix.html

- Roweis, S.T. and Ghahramani, Z. (1999) A Unifying Review of Linear Gaussian Models. Neural Computation 11(2). Sections 1-5.3 and 6-6.1. See also Appendix A.1-A.2. http://www.gatsby.ucl.ac.uk/∼zoubin/papers/lds.ps.gz

- Welling, M. (2000) Linear models. class notes. http://www.gatsby.ucl.ac.uk/∼zoubin/course01/PCA.ps