4F13: Machine Learning

Lecture 5: Unsupervised Learning: ICA and EM

Zoubin Ghahramani

zoubin@eng.cam.ac.uk

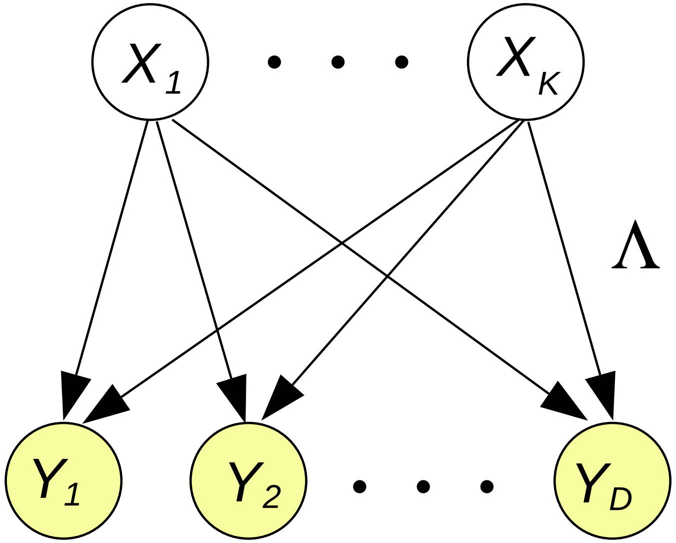
Department of Engineering
University of Cambridge

Michaelmas, 2006

http://learning.eng.cam.ac.uk/zoubin/ml06/

Factor Analysis

Factor analysis models high dimensional data $\mathbf{y}$ in terms of a linear transformation of some smaller number of latent factors, $\mathbf{x}$.



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

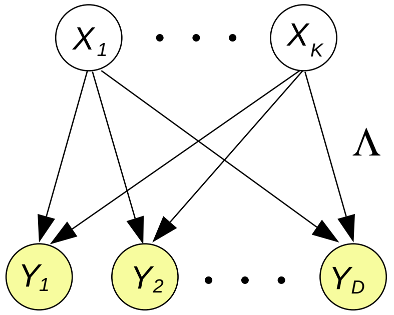
- x_k are independent $\mathcal{N}(0,1)$ Gaussian factors
- ϵ_d are independent $\mathcal{N}(0,\Psi_{dd})$ Gaussian noise
- $K < D$

Properties:

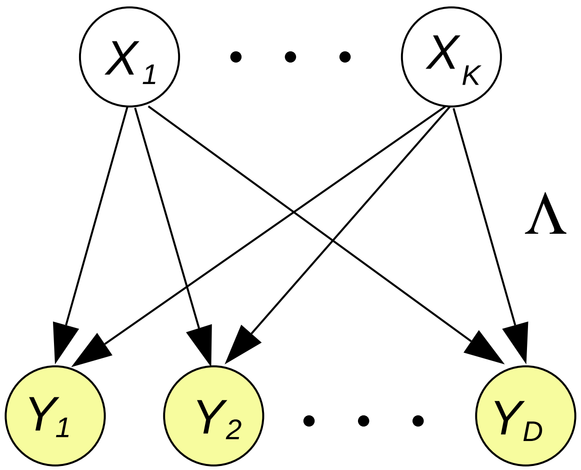
- $p(\mathbf{x}) = \mathcal{N}(0, I)$ and $\mathbf{y} = \Lambda\mathbf{x} + \epsilon$
- Since $p(\epsilon) = \mathcal{N}(0, \Psi)$, we get that $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\Lambda\mathbf{x}, \Psi)$
- $p(\mathbf{y}) = \int p(\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{x} = \mathcal{N}(0, \Lambda\Lambda^\top + \Psi)$ where Λ is a $D \times K$ matrix, and Ψ is diagonal.

latent = hidden = unobserved = missing

Blind Source Separation



Independent Components Analysis



- Just like Factor Analysis, hidden factors in ICA are *independent*: $p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k)$

- **But**, their distribution $p(x_k)$ is *non-Gaussian*:

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

- We can call the special case of $K = D$, with invertible Λ and zero observation noise, standard ICA. This was the originally proposed model (analogous to PCA) and has been studied extensively[1]:

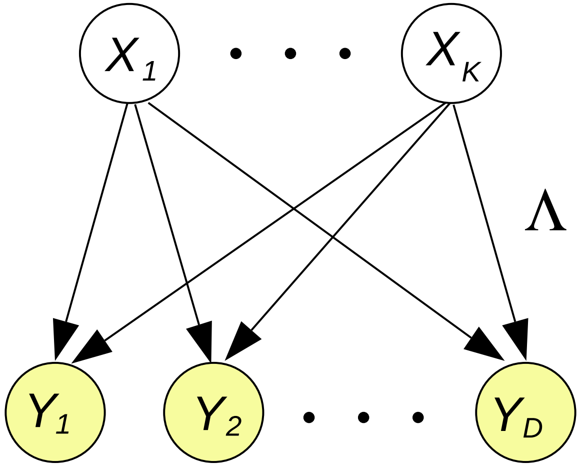
$$\mathbf{y} = \Lambda \mathbf{x} \quad \text{which implies} \quad \mathbf{x} = W\mathbf{y} \quad \text{where} \quad W = \Lambda^{-1}$$

  where $\mathbf{x}$ are the independent components (factors), $\mathbf{y}$ are the observations, Λ is the mixing matrix, and W is the unmixing matrix.

- Inferring $\mathbf{x}$ given $\mathbf{y}$ and learning Λ is easy in standard ICA.

[1] See: `http://www.cnl.salk.edu/~tony/ica.html`

ICA: Choosing non-Gaussian hidden factor densities



- Just like Factor Analysis, hidden factors in ICA are *independent*: $p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k)$
- **But**, their distribution $p(x_k)$ is *non-Gaussian*:

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

There are many possible continuous non-Gaussian densities for the hidden factors $p(x_k)$ from which we can choose.

A major distinction between univariate distributions is whether they are heavy tailed or light tailed.

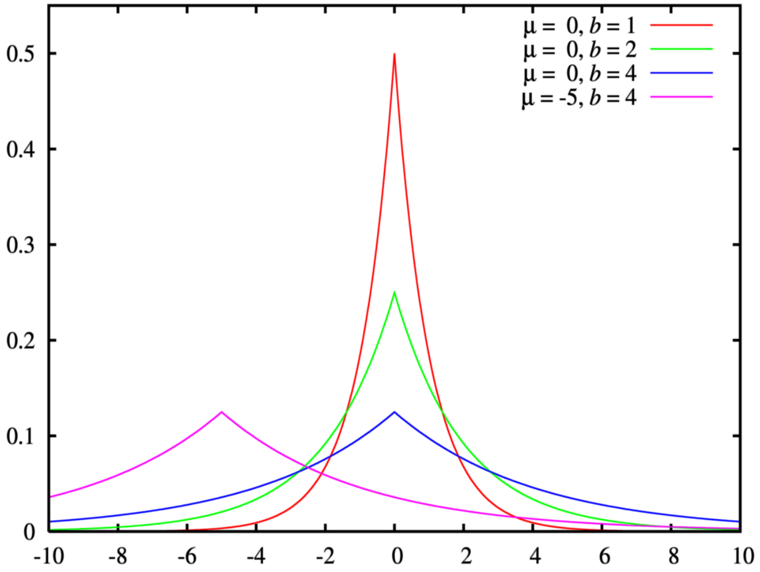
This is defined in terms of the kurtosis.

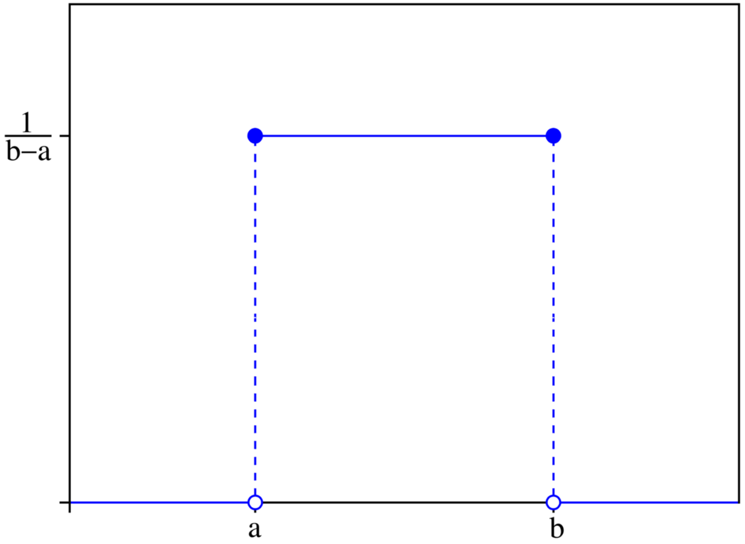
Kurtosis

The kurtosis (or excess kurtosis) measures how “peaky” or “heavy-tailed” a distribution is.

$$K = \frac{E((x-\mu)^4)}{E((x-\mu)^2)^2} - 3$$

where $\mu = E(x)$ is the mean of x.
Gaussian distributions have zero kurtosis.





Heavy tailed distributions have positive kurtosis (leptokurtic).

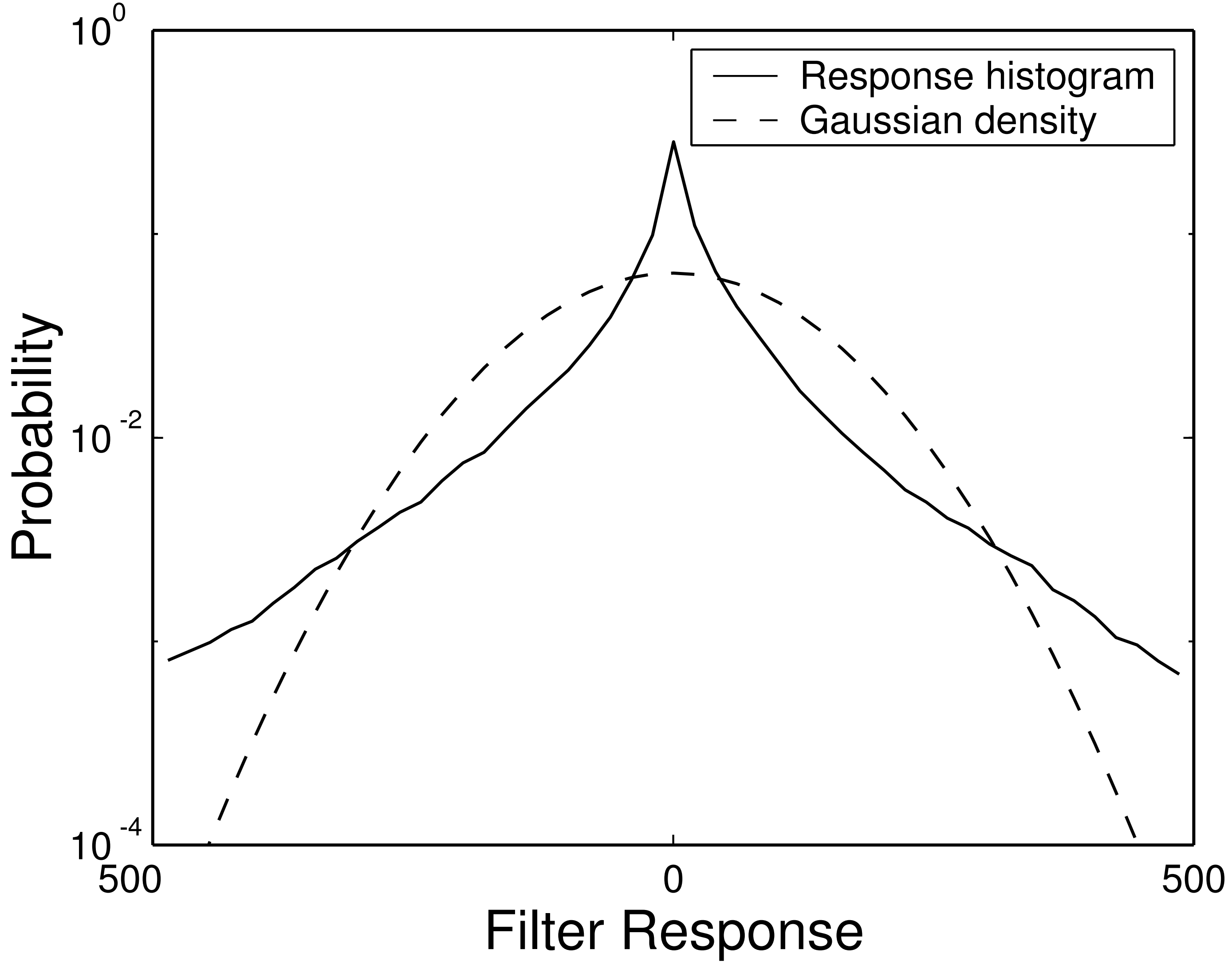
Light tailed distributions have negative kurtosis (platykurtic).

ICA models often use heavy-tailed distributions.

Why are heavy-tailed distributions interesting?

Natural Scenes and Sounds

Experiment: take some local linear filter (e.g. Gabor wavelet) and run it on some natural sounds or images. Measure filter output.



Natural Scenes

Interesting fact: ICA models seem to learn representations ($\mathbf{x}$ given $\mathbf{y}$) that look very similar to responses of neurons in primary visual cortex of the brain.

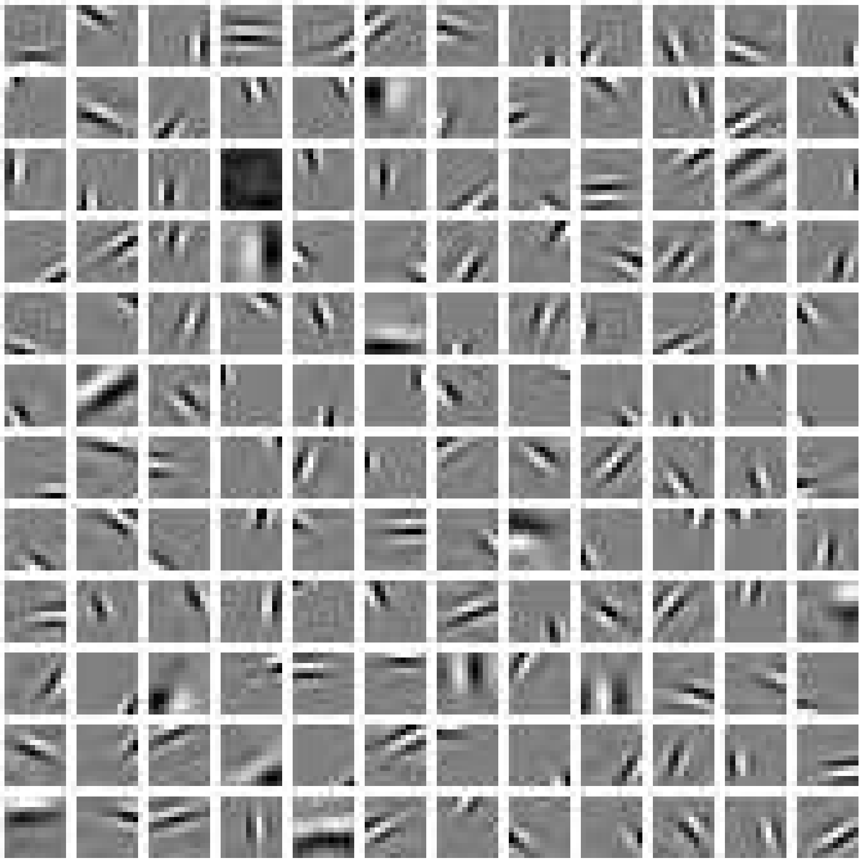


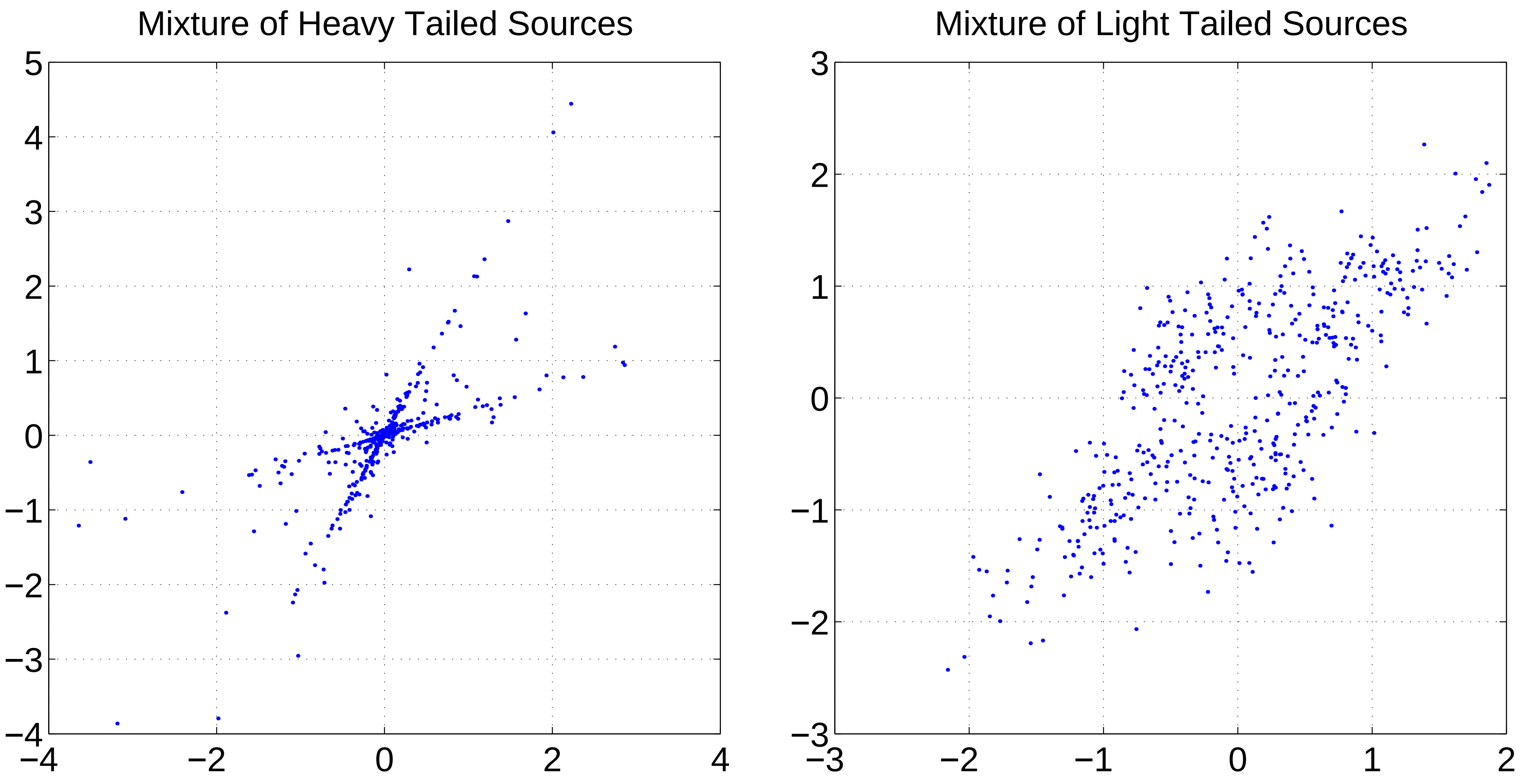
Figure 7: Example basis functions derived using sparseness criterion see (Olshausen & Field 1996).

Applications of ICA and Related Methods

- Separating auditory sources
- Analysis of EEG data
- Analysis of functional MRI data
- Natural scene analysis
- ...

Generating data from an ICA model

To understand how ICA works it's useful to show data generated from it ($K = D = 2$).



ICA (with heavy tailed noise) tries to find the directions with outliers.

How ICA Relates to Factor Analysis and Other Models

- **Factor Analysis (FA)**: Linear latent variable model which assumes that the factors are Gaussian, and Gaussian observation noise.

- **Probabilistic Principal Components Analysis (pPCA)**: Assumes isotropic observation noise: $\Psi = \sigma^2 I$ (PCA: $\Psi = \lim_{\sigma^2 \to 0} \sigma^2 I$).

- **Independent Components Analysis (ICA)**: Assumes that the factors are non-Gaussian.

- **Mixture of Gaussians**: A single discrete-valued "factor": $x_k = 1$ and $x_j = 0$ for all $j \neq k$.

- **Linear Gaussian State-space Model (Linear Dynamical System)**: Time series model in which the factor at time t depends linearly on the factor at time $t-1$, with added Gaussian noise.

ICA can and has been extended in several ways: fewer sources than "microphones", time varying mixing matrices, combining with convolution with linear filters, discovering number of sources...

The EM Algorithm

- Latent variable models:[2] model data $\mathbf{y}_n$ in terms of latent variables $\mathbf{x}_n$.

- Data set $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, likelihood: $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{y}_n|\boldsymbol{\theta}) = \prod_n \int p(\mathbf{y}_n, \mathbf{x}_n|\boldsymbol{\theta}) d\mathbf{x}_n$

- Goal: learn maximum likelihood (ML) parameter values

- The maximum likelihood procedure finds parameters $\boldsymbol{\theta}$ such that:

$$\boldsymbol{\theta}_{\mathrm{ML}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}|\boldsymbol{\theta})$$

- Because of the integral (or sum) over latent variables, the likelihood can be a very complicated, and hard to optimize function of $\boldsymbol{\theta}$.

- The Expectation–Maximization (EM) algorithm is a method for ML learning of parameters in latent varible models.

- Basic intuition of EM: iterate between inferring latent variables and fitting parameters.

[2] Examples of latent variable models: factor analysis, probabilistic PCA, ICA, mixture models, hidden Markov models, linear-Gaussian state-space models...

The Expectation Maximisation (EM) algorithm

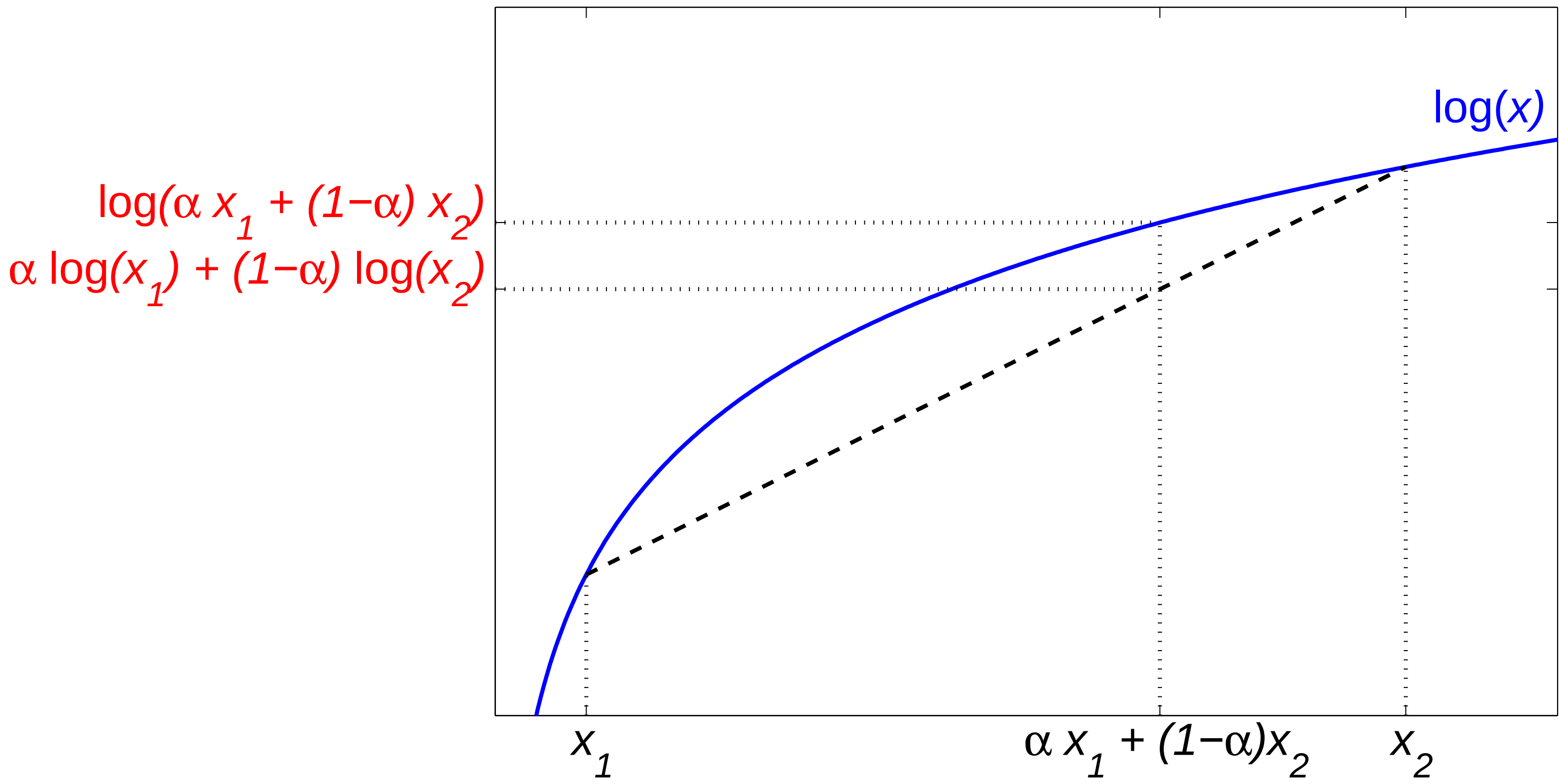
The EM algorithm finds a (local) maximum of a latent variable model likelihood. It starts from arbitrary values of the parameters, and iterates two steps:

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. FA turns into linear regression).
- 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).

Lower Bounding the Log Likelihood

Observed data $\mathcal{D} = \{\mathbf{y}_n\}$; Latent variables $\mathcal{X} = \{\mathbf{x}_n\}$; Parameters $\boldsymbol{\theta}$.

Goal: Maximize the log likelihood (i.e. ML learning) wrt $\boldsymbol{\theta}$:

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

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:

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

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

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

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

Notation and Terminology

$$\langle f(x) \rangle_{p(x)} \overset{\text{def}}{=} \int f(x) p(x) dx$$

$$\mathbf{H}[p] = -\int p(x) \log p(x) dx$$

Links between statistical physics and machine learning:

- negative log probabilities correspond to the "energy" of a system
- $-\langle \log P(\mathcal{X}, \mathcal{D}|\boldsymbol{\theta}) \rangle_{q(\mathcal{X})}$ is the average energy
- $\mathcal{F}(q, \boldsymbol{\theta})$ is the negative free energy

Physical systems tend to converge to a distribution of states with low free energy $\approx$ Learning systems should find a distribution of parameters and hidden variables with low free energy

The E and M steps of EM

The lower bound on the log likelihood is given by:

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

EM alternates between:

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

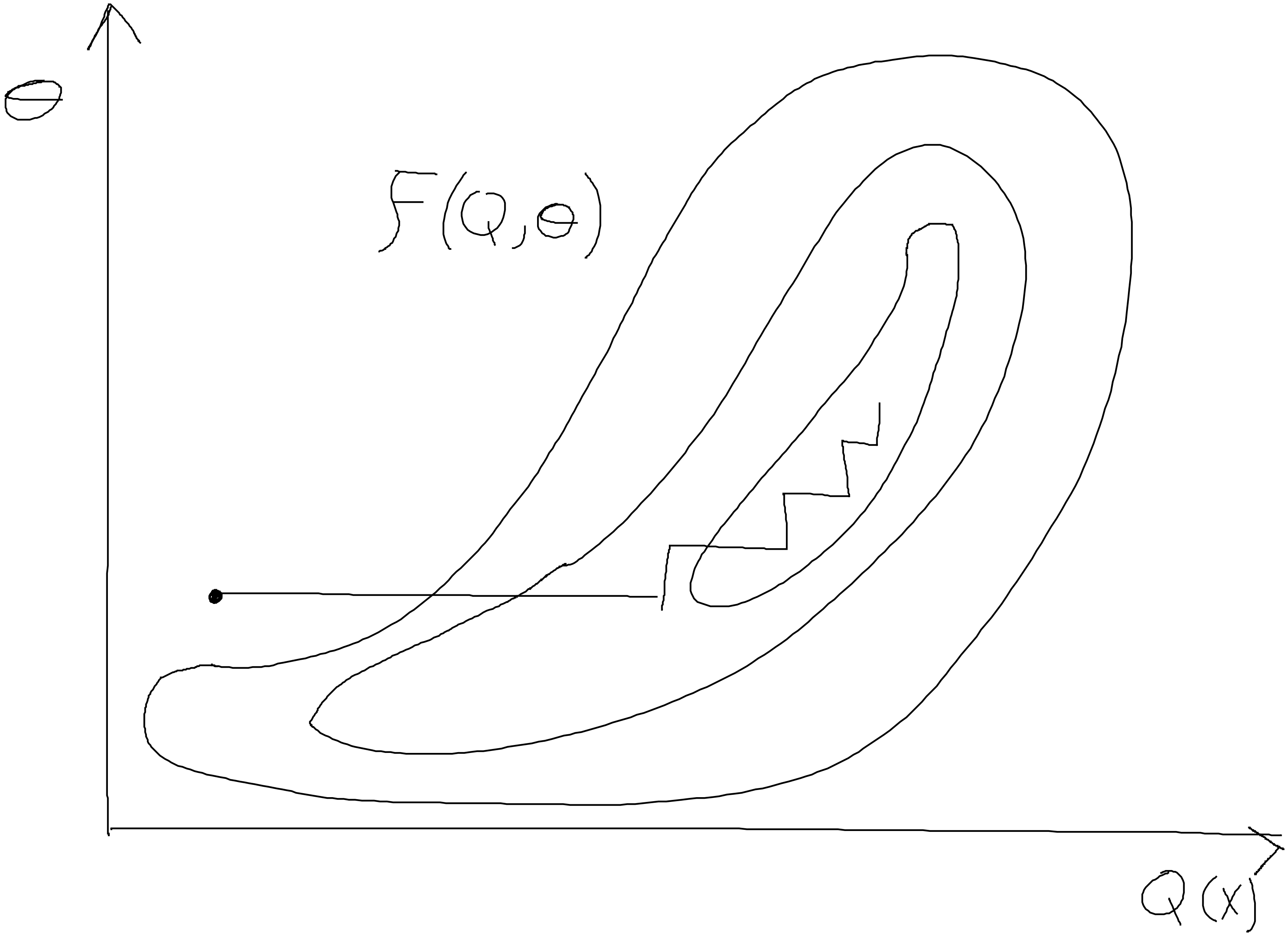
$$q^{(k)}(\mathcal{X}) := \underset{q(\mathcal{X})}{\operatorname{argmax}} \ \mathcal{F}\big(q(\mathcal{X}), \boldsymbol{\theta}^{(k-1)}\big).$$

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

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

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

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



The E Step

The free energy can be re-written

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

The second term is the Kullback-Leibler divergence.

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

But $\mathbf{KL}[q \| p]$ is zero if and only if $q = p$.

So, the E step simply sets

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

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

The M Step

$$\mathcal{F}(q, \boldsymbol{\theta}) = \int q(\mathcal{X}) \log \frac{P(\mathcal{X}, \mathcal{D}|\boldsymbol{\theta})}{q(\mathcal{X})}\, d\mathcal{X}$$

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

$$\begin{aligned} \boldsymbol{\theta}^{(k)} &:= \operatorname*{argmax}_{\boldsymbol{\theta}} \; \mathcal{F}\big(q^{(k)}(\mathcal{X}), \boldsymbol{\theta}\big) && (1) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \; \int q^{(k)}(\mathcal{X}) \log P(\mathcal{X}, \mathcal{D}|\boldsymbol{\theta})\, d\mathcal{X} && (2) \end{aligned}$$

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

The specific form of the M step depends on the model.

Often the maximum wrt $\boldsymbol{\theta}$ can be found analytically. See the appendix for the M step for factor analysis.

Appendices

Appendix: Matlab Code for Standard ICA

```
% ICA using tanh nonlinearity and batch covariant algorithm
% (c) Zoubin Ghahramani
%
% function [W, Mu, LL]=ica(X,cyc,eta,Winit);
%
% X - data matrix (each row is a data point),   cyc - cycles of learning (default = 200)
% eta - learning rate (default = 0.2),          Winit - initial weight
%
% W - unmixing matrix,  Mu - data mean,         LL - log likelihoods during learning

function [W, Mu, LL]=ica(X,cyc,eta,Winit);

if nargin<2,  cyc=200; end;
if nargin<3,  eta=0.2; end;
[N D]=size(X);                  % size of data
Mu=mean(X); X=X-ones(N,1)*Mu;   % subtract mean
if nargin>3,    W=Winit;        % initialize matrix
else,   W=rand(D,D);  end;
LL=zeros(cyc,1);                % initialize log likelihoods

for i=1:cyc,
  U=X*W';
  logP=N*log(abs(det(W)))-sum(sum(log(cosh(U))))-N*D*log(pi);
  W=W+eta*(W-tanh(U')*U*W/N);                   % covariant algorithm
  % W=W+eta*(inv(W)-X'*tanh(U)/N)';             % standard algorithm
  LL(i)=logP; fprintf('cycle %g log P= %g\n',i,logP);
end;
```

Appendix: The $\textbf{KL}[q(x)\|p(x)]$ 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 $\textbf{KL}[q\|p]$ 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:

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

Appendix: 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.

Appendix: EM Never Decreases the Likelihood

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

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

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt $\boldsymbol{\theta}$.
- $\mathcal{F} \leq \mathcal{L}$ by Jensen – or, equivalently, from the non-negativity of KL

If the M-step is executed so that $\boldsymbol{\theta}^{(k)} \neq \boldsymbol{\theta}^{(k-1)}$ iff $\mathcal{F}$ increases, then the overall EM iteration will step to a new value of $\boldsymbol{\theta}$ iff the likelihood increases.

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

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

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

Now,

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

so,

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

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

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

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

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

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

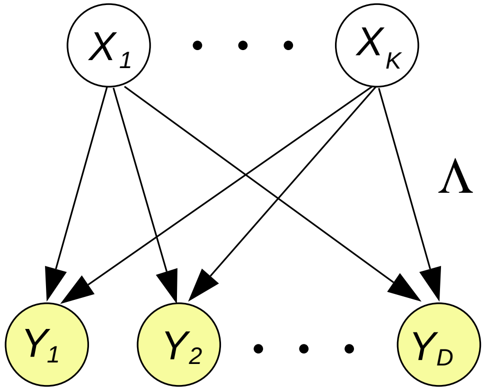
Differentiating the previous expression wrt $\boldsymbol{\theta}$ again we find

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

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

$\boldsymbol{\theta}^*$ is a maximum of $\mathcal{L}$.

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

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

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

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

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.)

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

Taking derivatives w.r.t. Λ and Ψ^{-1}, using $\frac{\partial \mathrm{Tr}[AB]}{\partial B} = A^\top$ 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: $\theta^{(k)} := \underset{\theta}{\operatorname{argmax}} \ \mathcal{F}(q, \theta)$

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

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

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