EM Algorithm

- an iterative method for numerically calculating MLE
- increases the likelihood at each step

$x[n] \rightarrow$ real observed data

$y[n] \rightarrow$ latent / imaginary data $\rightarrow$ used if $y[n]$ makes MLE calculation easier

Example: Let $x[n] = \sum_{i=1}^{p} \cos 2\pi f_i n + w[n]$ $\quad n = 0, 1, \cdots, N-1$

$w[n] \hookrightarrow N(0, \sigma^2)$

$\theta = [f_1 \cdots f_p]^T \rightarrow$ to be estimated.

Normally MLE would req. minimization of

$$J(f) = \sum_{n=0}^{N-1} \left(x[n] - \sum_{i=1}^{p} \cos 2\pi f_i n \right)^2$$

$f = \begin{bmatrix} f_1 \\ \vdots \\ f_p \end{bmatrix}$ $\rightarrow$ min over p dimensions.

$\rightarrow p$ coupled equations.

On the other hand, if we had different data, say $y_i[n] = \cos 2\pi f_i n + w_i[n]$

$i = 1, 2, \cdots, p$

$n = 0, 1, \cdots, N-1$

where $w_i[n] \sim N(0, \sigma_i^2)$, w_i & w_j are ind.

& $\sum_{i=1}^{p} \sigma_i^2 = \sigma^2$

Then, we can calculate $\hat{f}_i$ by minimizing

$$J(f_i) = \sum_{n=0}^{N-1} \left(y_i[n] - \cos 2\pi f_i n \right)^2$$

separately for each $i = 1, \cdots, p$.

$\rightarrow$ p - decoupled equations; one for each p.

$\Rightarrow$ Easier problem

$\{y_1[n], y_2[n], \cdots, y_p[n]\}$ $\Big\}$ → Complete / Latent Data
$(n = 0, 1, \cdots, N-1)$

$\{x[0], \cdots, x[N-1]\}$ → Incomplete Data

Clearly $\quad x[n] = \sum_{i=1}^{p} y_i[n]$ if $w[n] = \sum_{i=1}^{p} w_i[n]$

Note: This decomposition is not unique: e.g.

$y_1[n] = \sum_{i=1}^{p} \cos 2\pi f_i n \qquad y_2[n] = w[n]$ $\Big\}$ is also a possible decomp.
$x[n] = y_1[n] + y_2[n]$

In general: $\quad x = g(y_1, \cdots, y_p) = g(y)$

Main problem: $y_i[n]$ are not known in reality, whereas $x[n]$ is observed.

Recall: The mse estimate of $\ln p(y;\theta)$ given x is:

$$E_{y|x}[\ln p_y(y;\theta)] = \int \ln p(y;\theta)\, p(y|x;\theta)\, dy$$

unknown known unknown

Strategy: 1) Replace θ with the current guess of θ & then $\max_{\theta} E_{y|x}[\ln p_y(y;\theta)]$.

2) Iterate

E.M. Algo: (Expectation – Maximization)

E-Step: $\quad U(\theta, \theta_k) = \int \ln p_y(y;\theta)\, p(y|x;\theta_k)\, dy$

M-Step: $\quad \theta_{k+1} = \arg\max_{\theta} U(\theta, \theta_k)$

Calulation of the conditional prob $p(y/x;\theta)$ is often very difficult.

For our example above:

$$\ln p(y;\theta) = \sum_{i=1}^{P} \ln p(y_i;\theta_i)$$

$$= \sum_{i=1}^{P} \ln \left\{ \frac{1}{(2\pi\sigma_i^2)^{N/2}} \exp\left[-\frac{1}{2\sigma_i^2} \sum_{n=0}^{N-1} \left(y_i[n] - \cos 2\pi f_i n \right)^2 \right] \right\}$$

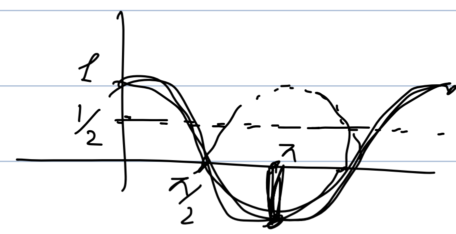
$$= \underbrace{c}_{\text{constant}} - \sum_{i=1}^{P} \frac{1}{2\sigma_i^2} \sum_{n=0}^{N-1} \left(y_i[n] - \cos 2\pi f_i n \right)^2$$

$$= \underbrace{g(y)}_{\text{ind. of } f_i} + \sum_{i=1}^{P} \frac{1}{\sigma_i^2} \sum_{n=0}^{N-1} \left(y_i[n] \cos 2\pi f_i n - \frac{1}{2} \cos^2 2\pi f_i n \right)$$

Approximate $\sum_{n=0}^{N-1} \cos^2 2\pi f_i n \approx \frac{N}{2}$ for f_i not near 0 or $\frac{1}{2}$

$$\begin{cases} \cos 0 = 1 \Rightarrow \sum_{n=0}^{N-1} 1 = N \\ \cos 2\pi \frac{1}{2} n = \cos \pi n = -N \end{cases}$$

Otherwise $\approx \sum_{n=0}^{N-1} \frac{1}{2} \approx \frac{N}{2}$



Then $\ln p(y;\theta) = h(y) + \sum_{i=1}^{P} \frac{1}{\sigma_i^2} \sum_{n=0}^{N-1} y_i[n] \cos 2\pi f_i n$

Let $c_i = \left[1 \;\; \cos 2\pi f_i \;\; \ldots \;\; \cos 2\pi f_i (N-1) \right]^T$

Then $\ln p(y;\theta) = h(y) + \sum_{i=1}^{P} \frac{1}{\sigma_i^2} c_i^T y_i$, $\quad y_i = \begin{bmatrix} y_i[0] \\ \vdots \\ y_i[N-1] \end{bmatrix}$

Further let $c = \begin{bmatrix} \frac{1}{\sigma_1^2} c_1 \\ \vdots \\ \frac{1}{\sigma_p^2} c_p \end{bmatrix}$ $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$

Then $\ln p(y;\theta) = h(y) + c^T y$

The E-step : $U(\theta;\theta_k) = E[\ln p(y;\theta)/x;\theta_k]$

$$= \underbrace{E[h(y)/x;\theta_k]}_{\substack{\text{ind. of } \theta \\ \text{hence ignored} \\ \text{in next M-step}}} + c^T \underbrace{E(y/x;\theta_k)}_{\text{Recall } y \text{ \& } x \text{ are jointly Gaussian}}$$

since $\underset{N\times 1}{x} = \sum y_i = \underbrace{[\underset{N\times N}{I}\ I \cdots I]}_{N \times Np} \underset{Np\times 1}{y}$

Since x & y are jointly Gaussian:

$$E(y/x;\theta_k) = E(y) + C_{yx} C_{xx}^{-1} (x - E(x))$$

Recall : $E(y) = \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix}_{Np\times 1}$ $E(x) = \sum_{i=1}^{p} c_i$

$C_{xx} = \sigma^2 I$; $C_{yx} = E\left(\begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} w^T \right)$

$$= E\left(\begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} \left\{ [I\ I \cdots I] \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} \right\}^T \right)$$

$$= \begin{bmatrix} \sigma_1^2 I & & \\ & \ddots & \\ & & \sigma_p^2 I \end{bmatrix} \begin{bmatrix} I \\ I \\ \vdots \\ I \end{bmatrix} = \begin{bmatrix} \sigma_1^2 I \\ \vdots \\ \sigma_p^2 I \end{bmatrix}_{Np \times N}$$

$N \times N$

$$E(y/x; \theta_k) = \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix} + \frac{1}{\sigma^2} \begin{bmatrix} \sigma_1^2 I \\ \vdots \\ \sigma_p^2 I \end{bmatrix} \left(x - \sum_{i=1}^{p} c_i\right)$$

$$= \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix} + \begin{bmatrix} \frac{\sigma_1^2}{\sigma^2}\left(x - \sum_{i=1}^{p} c_i\right) \\ \vdots \\ \frac{\sigma_p^2}{\sigma^2}\left(x - \sum_{i=1}^{p} c_i\right) \end{bmatrix}$$

$$\Leftrightarrow \underbrace{E(y_i^\circ/x; \theta_k)}_{\text{Recall this is the best estimate of } y_i^\circ} = c_i^\circ + \frac{\sigma_i^2}{\sigma^2}\left(x - \sum_{i=1}^{p} c_i\right) \quad i = 1, \cdots, p$$

$=: \hat{y}_i^\circ$

$$\hat{y}_i^\circ[n] = \cos 2\pi f_{i_k}^\circ n + \frac{\sigma_i^2}{\sigma^2}\left(x[n] - \sum_{i=1}^{p} \cos 2\pi f_{i_k}^\circ n\right)$$

Relevant part of the E-step:

$$U'(\theta, \theta_k) = \underbrace{\sum_{i=1}^{p} c_i^{\circ T} \hat{y}_i^\circ}$$

Note that this sum can be maximized by maximizing each term separately

i.e. $\underbrace{f_{i_{k+1}}^\circ}_{\text{iteration no}} = \arg \max_{f_i^\circ} c_i^{\circ T} \hat{y}_i^\circ$

In effect: the E-M Steps are:

E-Step: For $i = 1, 2, \cdots, P$

$$\overbrace{y_i^\circ[n] = \cos 2\pi f_{i_k}^\circ n + \beta_i\left[x[n] - \sum_{i=1}^{P} \cos 2\pi f_{i_k} n\right)}^{\text{Est. of } w_i[n]}$$

M-Step: For $i = 1, 2, \cdots, P$

$$f_{i_{k+1}}^\circ = \arg\max_{f_i^\circ} \sum_{n=0}^{N-1} y_i^\circ[n] \cos 2\pi f_i^\circ n$$

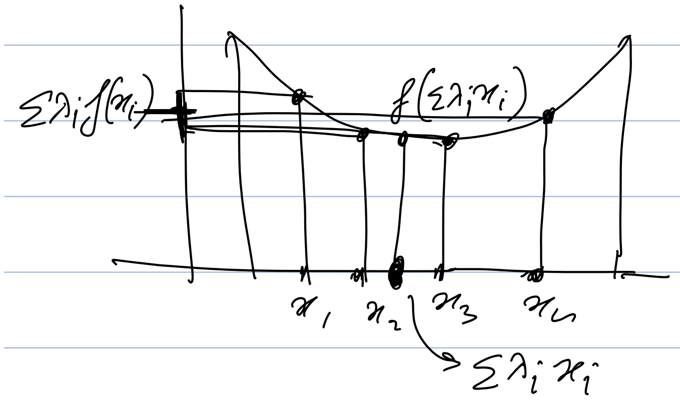
β_i's are arbitrary as long as $\sum_{i=1}^{P} \beta_i^\circ = 1$?

σ_i^2 are not unique, they can be chosen arbitrarily as long as $\sum_{i=1}^{P} \sigma_i^2 = \sigma^2$ or $\sum_{i=1}^{P} \beta_i = \sum_{i=1}^{P} \frac{\sigma_i^2}{\sigma^2} = 1$

Q. Why does it work?

Jensen's Inequality: Let f be a convex function defined on an interval I. If $x_1, \cdots, x_n \in I$ and $\lambda_1, \lambda_2, \cdots, \lambda_n \geq 0$ with $\sum_{i=1}^{n} \lambda_i = 1$, then

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$



Proof: $n = 1$ trivial

$n = 2$ defn of convexity

Induction: assume true for n.

$$f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) = f\left(\lambda_{n+1} x_{n+1} + \sum_{i=1}^{n} \lambda_i x_i\right)$$

$$= f\left(\lambda_{n+1} x_{n+1} + (1-\lambda_{n+1})\left(\frac{1}{1-\lambda_{n+1}}\right)\sum \lambda_i x_i\right)$$

$$\leqslant \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) f\left(\frac{1}{1-\lambda_{n+1}} \sum \lambda_i x_i\right)$$

$$= \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) f\left(\sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} x_i\right)$$

$$\leqslant \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) \sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} f(x_i)$$

$$= \lambda_{n+1} f(x_{n+1}) + \sum_{i=1}^{n} \lambda_i f(x_i) \quad \left[\because \lambda_1 + \cdots + \lambda_{n+1} = 1\right]$$

$$= \sum_{i=1}^{n+1} \lambda_i f(x_i)$$

\# Since $\ln(x)$ is concave,

$$\ln\left(\sum_{i=1}^{n} \lambda_i x_i\right) \geqslant \sum_{i=1}^{n} \lambda_i \ln(x_i)$$

\# Technical Assumption on y (the latent variable)

$\theta \to Y \to X$ is a Markov Chain

$\Leftrightarrow$ $P(x/y)$ is independent of θ.

Define $L(\theta) = \ln p(x;\theta)$ i.e. $L(\theta_k) := \ln p(x;\theta_k)$

Then, $L(\theta) - L(\theta_k) = \ln\left(\frac{p(x;\theta)}{p(x;\theta_k)}\right)$

$$= \ln \int \frac{p(y,x;\theta)\,dy}{p(x;\theta_k)} = \ln \int \frac{P(y,x;\theta)}{\frac{P(y,x;\theta_k)}{P(y/x;\theta_k)}}\,dy$$

$$= \ln \int \frac{P(y,x;\theta)}{P(y,x;\theta)} p(y/x;\theta_k)\,dy$$

$$= \ln \underbrace{\int \frac{P(y;\theta)\cancel{P(x/y)}}{P(y;\theta_k)\cancel{P(x/y)}} p(y/x;\theta_k)\,dy}$$

due to Markov assumption.

$$\geq \int \left\{ \ln \left(\frac{P(y;\theta)}{P(y;\theta_k)} \right) \right\} p(y/x;\theta_k)\,dy \quad \left[\text{By Jensen's inequality} \right] \rightarrow (\mathcal{L})$$

$$= \underbrace{\int \ln p(y;\theta)\, p(y/x;\theta_k)\,dy}$$

$$- \underbrace{\int [\ln p(y;\theta_k)]\, p(y/x;\theta_k)\,dy}_{\text{Ind of } \theta \rightarrow \text{Ignore}}$$

Recall that this is the quantity we maximize in the E-M step.

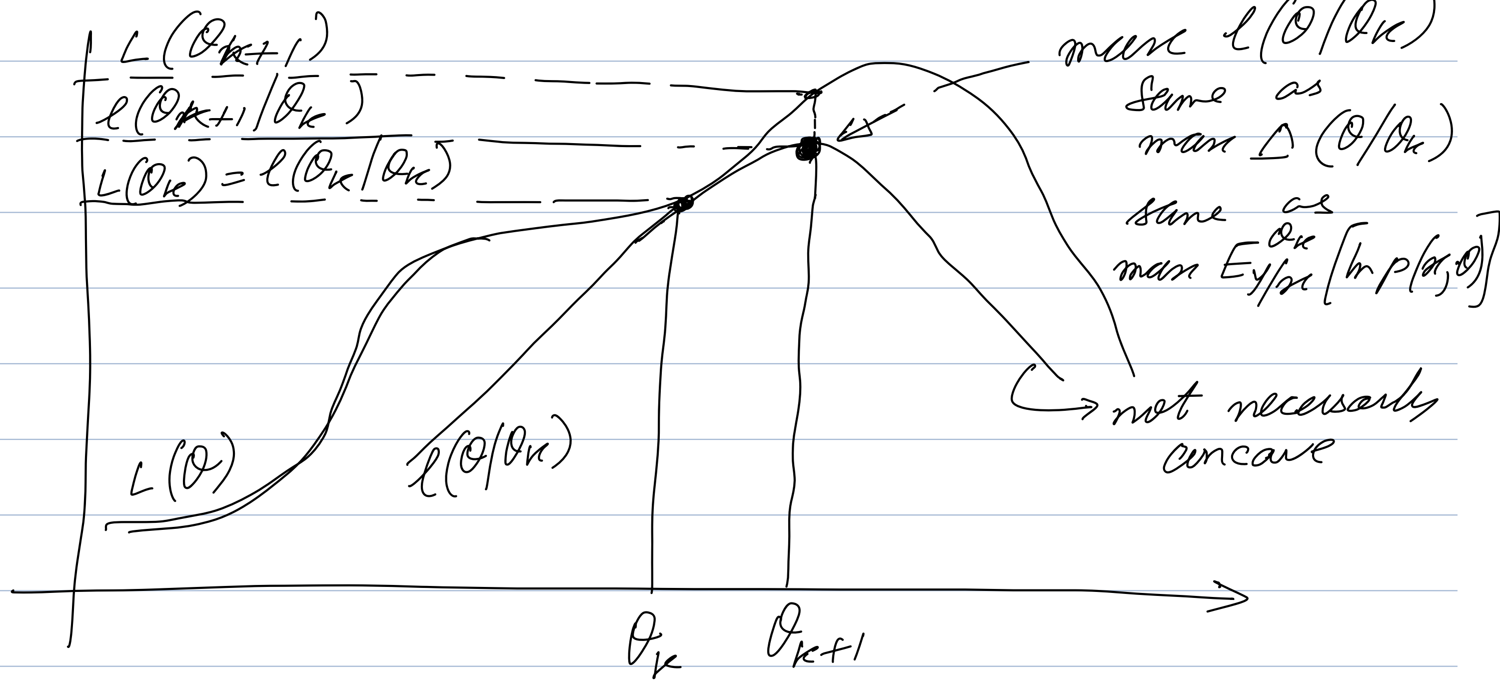
Define $\Delta(\theta/\theta_k) := \int \left\{ \ln \left(\frac{P(y;\theta)}{P(y;\theta_k)} \right) \right\} p(y/x;\theta_k)\,dy$

From $(\mathcal{L})$, $L(\theta) - L(\theta_k) \geq \Delta(\theta/\theta_k)$

$$\Leftrightarrow L(\theta) \geq \underbrace{L(\theta_k) + \Delta(\theta/\theta_k)}_{=: l(\theta/\theta_k)}$$

Clearly, $\Delta(\theta_k/\theta_k) = \int \ln 1\, p(y/x;\theta_k)\,dy = 0$

$$\Rightarrow l(\theta_k/\theta_k) = L(\theta_k)$$



From fig. it is 'clear' that value of
$L(\theta_{k+1}) \geq L(\theta_k) \quad \forall k.$

This also follows from the fact that we maximize $l(\theta/\theta_k)$ at each iteration

$$\boxed{L(\theta_{k+1}) \geq L(\theta_k) \quad \forall k}$$

Monotonicity Property of EM.