Addendum: The EM Algorithm

In many problems MLE based on observed data $X$ would be greatly simplified if we had additionally observed another piece of data $Y$. $Y$ is called the hidden or latent data.

**Example 1** $X \sim \mathcal{N}(H\theta, I)$ can be modeled as:

$$
Y_{k \times 1} = \theta + W_1
$$

$$
X_{n \times 1} = H_{n \times k} Y + W_2
$$

such that $HW_1 + W_2 \sim \mathcal{N}(0, I)$.

If we just have $X$, then we must solve a system of equations to obtain the MLE. If the dimension is large, then computing the MLE is quite expensive (i.e. the inversion is at least $O(\max(nk^2, k^3))$). But if we also have $Y$, then the MLE can be computed with $O(k)$ as we know $\hat{\theta} = Y$.

**Example 2**

$x_i \overset{iid}{\sim} p\mathcal{N}(\mu_0, \sigma_0^2) + (1-p)\mathcal{N}(\mu_1, \sigma_1^2)$

$y_i \overset{iid}{\sim} \text{Bernoulli}(p) = p^{1-y_i} (1-p)^{y_i}$

$x_i | y_i = l \sim \mathcal{N}(\mu_l, \sigma_l^2)$

Given $\{(x_i, y_i)\}_{i=1}^{n}$, we have:

$$
\hat{\mu}_l = \frac{1}{\sum 1_{y_i = l}} \sum 1_{y_i = l} x_i
$$

$$
\hat{\sigma}_l = \frac{1}{\sum 1_{y_i = l}} \sum 1_{y_i = l} (x_i - \hat{\mu}_l)^2
$$

$$
\hat{p} = \frac{1}{n} \sum 1_{y_i = l}
$$

MLE’s are easy to compute here. However, if we only have $\{x_i\}_{i=1}^{n}$, the computation of MLE is a complicated, non-convex optimization, where we can apply EM algorithm to compute. The application of EM algorithm in this situation is shown in Example 4.

**Main Idea**

Let $L(\theta) = \log p(x|\theta)$ and also define the complete data log-like:

$$L_c(\theta) = \log p(x, y|\theta) = \log p(y|x|\theta) p(x|\theta)$$

$$= \log p(y|x|\theta) + \log p(x|\theta) = \log p(y|x|\theta) + L(\theta)$$

Suppose our current guess of $\theta$ is $\theta^{(t)}$ and that we would like to improve this guess. Consider

$$L(\theta) - L(\theta^{(t)}) = L_c(\theta) - L_c(\theta^{(t)}) + \log \frac{p(y|x|\theta^{(t)})}{p(y|x|\theta)}$$
Now take expectation of both sides with respect to \( y \sim p(y|x|\theta(t)) \), we have:
\[
L(\theta) - L(\theta(t)) = E_y[L_c(\theta)] - E_y[L_c(\theta(t))] + D(p(y|x|\theta(t)) \| p(y|x))
\]
As \( D(p(y|x|\theta(t)) \| p(y|x)) \geq 0 \), we have the following inequality:
\[
L(\theta) - L(\theta(t)) \geq E_y[L_c(\theta)] - E_y[L_c(\theta(t))] = Q(\theta, \theta(t)) - Q(\theta(t), \theta(t))
\]
Note: \( Q(\theta, \theta') = E_{y}(y|\theta') \log \frac{p(y|x)}{p(y|x)} \) is the expectation of complete data log-likelihood. We choose \( \theta(t+1) \) as answer of the following optimization problem:
\[
\theta(t+1) = \arg \max_{\theta} Q(\theta, \theta(t))
\]
The relationship between \( \log p(x, \theta) \), \( Q(\theta, \theta(t)) \), \( \theta^t \) and \( \theta^{t+1} \) are showed in the following graph:

![Graphical show of EM algorithm](image)

The process of EM algorithm is as follows:

**Init:** \( t = 0, \theta^{(0)} = 0 \) or random value

**Loop:**

- **E step:** Compute
  \[
  Q(\theta, \theta^{(t)}) = E_y[y|x|\theta^{(t)}] \log p(x, y|\theta)
  \]

- **M step:**
  \[
  \theta^{(t+1)} = \arg \max_{\theta} Q(\theta, \theta^{(t)})
  \]

**End**

The E-step and M-step repeat until convergence.

**Properties of EM algorithm:**

1. \( \log p(x|\theta^{(0)}) \leq \log p(x|\theta^{(1)}) \leq \ldots \)

2. It converges to stationary point (e.g., local max)

**Example 3** Original model \( X = H\theta + W \):

**Complete model:**

\[
\begin{align*}
Y &= \theta + W_1 \quad W_1 \sim \mathcal{N}(0, \alpha^2 I_{k \times k}) \\
X &= H_{n \times k}Y + W_2 \quad W_2 \sim \mathcal{N}(0, I_{n \times n} - \alpha^2 HH^T)
\end{align*}
\]
Then we construct the complete log-likelihood:
\[
\log p(x, y | \theta) = \log p(x | y, \theta) + \log p(y | \theta) \\
= \text{constant} - \frac{\|y - \theta\|^2}{2\alpha^2} \\
= \frac{1}{2\alpha^2} (2\theta^T y - \theta^T \theta - y^T y) + \text{constant} \\
= \frac{1}{2\alpha^2} (2\theta^T y - \theta^T \theta) + \text{constant}
\]

As the part left after taking away the constant is proportional to \(y\), so we only need to calculate \(\mathbb{E}_{p(y|x|\theta(t))}[y]\).

Introduce \(Z_1 = Y, Z_2 = X - HY\), then we have the joint distribution of \(Z_1, Z_2\) as:
\[
\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \mathcal{N}(\begin{bmatrix} \theta \\ \alpha^2 I_{k \times k} \\ 0 \end{bmatrix}, \begin{bmatrix} \alpha^2 H H^T \end{bmatrix})
\]

As we know \(\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} H & I_{n \times n} \\ I_{k \times k} & 0 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}\), we know:
\[
\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \theta \\ \alpha^2 H \end{bmatrix}, \begin{bmatrix} I_{n \times n} & \alpha^2 H \\ \alpha^2 H^T & \alpha^2 I_{k \times k} \end{bmatrix})
\]

Make a linear transformation, we have:
\[
\begin{bmatrix} X \\ Y - \alpha^2 H^T X \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \theta \\ \alpha^2 H \theta \\ \alpha^2 H^T H \theta \\ \alpha^2 I_{k \times k} - \alpha^2 H^T H \end{bmatrix})
\]

So we have:
\[
\mathbb{E}_{p(y|x|\theta(t))}[y] = \alpha^2 H^T x + \theta(t) - \alpha^2 H^T H \theta(t) = y(t)
\]

As \(Q(\theta, \theta(t)) = \frac{1}{2\alpha^2} (2\theta^T y(t) - \theta^T \theta) + \text{constant}\), set \(\frac{\partial Q}{\partial \theta} = 0\), we have:
\[
\theta(t+1) = y(t)
\]

It is easy to calculate the stationary point in this iteration, let \(\theta(t+1) = \theta(t)\), we have:
\[
\theta_{\text{stationary}} = (H^T H)^{-1} H^T x
\]
which is the answer we are familiar with.

Example 4 Suppose:
\[
X_1, X_2, \ldots, X_n \overset{iid}{\sim} \sum_{j=1}^{m} p_j \mathcal{N}(\mu_j, \sigma_j^2)
\]
We have:
\[
p(x, y | \theta) = \prod_{i=1}^{n} \sum_{j=1}^{m} p_j \frac{1}{\sqrt{2\pi\sigma_j}} e^{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}} 1_{y_i = j}
\]
Thus,
\[
\log p(x, y | \theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \log(\frac{p_j}{\sqrt{2\pi\sigma_j}} e^{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}}) 1_{y_i = j}
\]
\[
\mathbb{E}_{p(y|x|\theta(t))}[\log p(x, y | \theta)] = \sum_{i=1}^{n} \sum_{j=1}^{m} \log(\frac{p_j}{\sqrt{2\pi\sigma_j}} e^{-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}}) \mathbb{E}_{p(y|x|\theta(t))}[1_{y_i = j}] 
\]
\[
\begin{align*}
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \log \left( \frac{p_j}{\sqrt{2\pi}\sigma_j} \right) - \frac{(x_i - \mu_j)^2}{2\sigma_j^2} - \frac{p_j^{(t)} N(x_i; \mu_j^{(t)}, (\sigma_j^{(t)})^2)}{\sum_{l=1}^{m} p_l^{(t)} N(x_i; \mu_l^{(t)}, (\sigma_l^{(t)})^2)} \\
\text{Denote } &p^{(t)}(y_i = j) = \frac{p_j^{(t)} N(x_i; \mu_j^{(t)}, (\sigma_j^{(t)})^2)}{\sum_{l=1}^{m} p^{(t)}_l N(x_i; \mu_l^{(t)}, (\sigma_l^{(t)})^2)}, \text{ we have the expression of } Q(\theta, \theta^{(t)}):
\end{align*}
\]
\[
Q(\theta, \theta^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} p^{(t)}(y_i = j) \log(p_j^{(t)} N(x_i; \mu_j, \sigma_j^2))
\]
\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} p^{(t)}(y_i = j) \log(N(x_i; \mu_j, \sigma_j^2)) + \text{constant}
\]

Set $\frac{\partial Q}{\partial \theta} = 0$, we have:
\[
\begin{align*}
\mu_j^{(t+1)} &= \frac{\sum_{i=1}^{n} p^{(t)}(y_i = j)x_i}{\sum_{i=1}^{n} p^{(t)}(y_i = j)} \\
(\sigma_j^{(t+1)})^2 &= \frac{\sum_{i=1}^{n} (x_i - \mu_j^{(t+1)})^2 p^{(t)}(y_i = j)}{\sum_{i=1}^{n} p^{(t)}(y_i = j)}
\end{align*}
\]