1 The EM Algorithm

1.1 Suffient Statistics and Exponential Distributions

Let $p(y|\theta)$ be a family of density functions parameterized by $\theta \in \Omega$, and let Y be a random object with a density function from this family.

- Definition: A *statistic* is any function $T(Y)$ of the data Y.
- Definition: We say that a statistic $T(Y)$ is a *sufficient statistic* for θ if there exist functions $g(\cdot, \cdot)$ and $h(\cdot)$ such that

$$p(y|\theta) = h(y)\, g(T(y), \theta) \tag{1}$$

for all $y \in I\!R^N$ and $\theta \in \Omega$.

If $T(Y)$ is a sufficient statistic for θ where θ parameterizes the distribution for Y, then the ML estimator of θ must be a function of $T(Y)$. To see this, notice that

$$\begin{aligned}
\hat{\theta}_{ML} &= \arg\max_{\theta \in \Omega} p(y|\theta) \\
&= \arg\max_{\theta \in \Omega} \log p(y|\theta) \\
&= \arg\max_{\theta \in \Omega} \{\log h(y) + \log g(T(y), \theta)\} \\
&= \arg\max_{\theta \in \Omega} \log g(T(y), \theta) \\
&= f(T(y))
\end{aligned}$$

for some function $f(\cdot)$.

Example 1: Let $\{Y_n\}_{n=0}^{N-1}$ be i.i.d. random variables with distribution $N(\mu, 1)$. Define the following statistic corresponding to the sample mean of the random variables.

$$t_1 = \sum_{n=0}^{N-1} y_n$$

By writing the density function for the sequence Y as

$$p(y|\mu) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y_n - \mu)^2\right\}$$

$$
\begin{aligned}
&= \frac{1}{\left(\sqrt{2\pi}\right)^N} \exp\left\{ -\frac{1}{2} \sum_{n=0}^{N-1} (y_n - \mu)^2 \right\} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^N} \exp\left\{ -\frac{1}{2} \sum_{n=0}^{N-1} (y_n^2 - 2y_n\mu + \mu^2) \right\} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^N} \exp\left\{ -\frac{1}{2} \sum_{n=0}^{N-1} y_n^2 \right\} \exp\left\{ \frac{1}{2}(2t_1\mu - N\mu^2) \right\} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^N} \exp\left\{ -\frac{1}{2} \left(\sum_{n=0}^{N-1} y_n^2 + t_1^2/N \right) \right\} \exp\left\{ -\frac{N}{2}(t_1/N - \mu)^2 \right\} ,
\end{aligned}
$$

where we can see that it has the form of equation (1). Therefore, t_1 is a sufficient statistic for the parameter μ. Computing the ML estimate yeilds the following.

$$
\begin{aligned}
\hat{\mu}_{ML} &= \arg\max_{\mu} \log p(y|\mu) \\
&= \arg\max_{\mu} \left\{ -\frac{N}{2}(t_1/N - \mu)^2 \right\} \\
&= \arg\min_{\mu} (t_1/N - \mu)^2 \\
&= \frac{t_1}{N}
\end{aligned}
$$

Many commonly used distributions such as Gaussian, exponential, Poisson, Bernoulli, and binomial have a structure which makes them particularly useful. These distributions are known as exponential families and have the following special property.

- Definition: A family of density functions $p(y|\theta)$ for $y \in R^N$ and $\theta \in \Omega$ is said to be a *k-parameter exponential family* if there exist functions $g(\theta) \in I\!R^k$, $s(y)$, $d(\theta)$ and statistic $T(y) \in I\!R^k$ such that

  $$p(y|\theta) = \exp\{< g(\theta), T(y) > + d(\theta) + s(y)\} \tag{2}$$

  for all $y \in I\!R^N$ and $\theta \in \Omega$ where $< \cdot, \cdot >$ denotes the inner product. We refer to $T(y)$ as the *natural sufficient statistic* or *natural statistic* for the exponential distribution.

Example 2: Let $\{Y_n\}_{n=0}^{N-1}$ be i.i.d. random variables with distribution $N(\mu, \sigma^2)$. Define the following statistics corresponding to the sample mean and variance of the random variables.

$$\begin{aligned}
t_1 &= \sum_{n=0}^{N-1} y_n \\
t_2 &= \sum_{n=0}^{N-1} y_n^2
\end{aligned}$$

Then we may write the density function for Y in the following form.

$$\begin{aligned}
p(y|\mu,\sigma^2) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_n-\mu)^2 \right\} \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} \exp\left\{ -\sum_{n=0}^{N-1} \frac{1}{2\sigma^2}(y_n-\mu)^2 \right\} \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (y_n^2 - 2y_n\mu + \mu^2) \right\} \\
&= \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} \exp\left\{ -\frac{1}{2\sigma^2} t_2 + 2\frac{\mu}{2\sigma^2} t_1 - \frac{N}{2\sigma^2}\mu^2 \right\} \\
&= \exp\left\{ -\frac{1}{2\sigma^2} t_2 + 2\frac{\mu}{2\sigma^2} t_1 - \frac{N}{2\sigma^2}\mu^2 - \frac{N}{2}\log(2\pi\sigma^2) \right\} \\
&= \exp\left\{ \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right] \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} - \frac{N}{2\sigma^2}\mu^2 - \frac{N}{2}\log(2\pi\sigma^2) \right\}
\end{aligned}$$

Using the following definitions

$$\begin{aligned}
g(\theta) &= \left[\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right] \\
T(y) &= \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \\
d(\theta) &= -\frac{N}{2\sigma^2}\mu^2 - \frac{N}{2}\log(2\pi\sigma^2) \\
s(y) &= 0
\end{aligned}$$

we can see that $p(y|\mu,\sigma^2)$ has the form of equation (2) with suffient statistic $T(y)$. With some calculations it may be easily shown that the ML estimates of μ and σ^2 are given by

$$\begin{aligned}
\hat{\mu}_{ML} &= \frac{t_1}{N} \\
\hat{\sigma}^2_{ML} &= \frac{t_2}{N} - \left(\frac{t_1}{N}\right)^2
\end{aligned}$$

1.2 General Formulation of EM Update

One reason that the EM algorithm is so useful is that for many practical situations the distributions are exponential, and in this case the EM updates have a particularly simple form. Let Y is the observed or incomplete data and let X be the unobserved data, and assume that the joint density of (Y, X) is from and exponential family with parameter vector θ. Then we know that

$$p(y, x|\theta) = \exp\{< g(\theta), T(y, x) > +d(\theta) + s(y, x)\}$$

for some sufficient statistic $T(y, x)$. Assuming the ML estimate of θ exists, then it is given by

$$\begin{aligned} \theta_{ML} &= \arg\max_{\theta\in\Omega}\{< g(\theta), T(y, x) > +d(\theta)\} && (3)\\ &= f(T(y, x)) && (4) \end{aligned}$$

where $f(\cdot)$ is some function of the k dimensional suffient statistic for the exponential density.

Recalling the form of the Q function, we have

$$Q(\theta', \theta) = E\left[\log p(y, X|\theta')|Y = y, \theta\right]$$

where Y is the observed data and X is the unknown data. Since our objective is to maximize Q with respect to θ', we only need to know the function Q within a constant that is not dependent on θ'. Therefore, we have

$$\begin{aligned} Q(\theta', \theta) &= E\left[\log p(y, X|\theta')|Y = y, \theta\right]\\ &= E\left[< g(\theta'), T(y, X) > +d(\theta') + s(y, X)|Y = y, \theta\right]\\ &= < g(\theta'), \bar{T} > +d(\theta') + constant \end{aligned}$$

were

$$\bar{T} = E\left[T(y, X)|Y = y, \theta\right]$$

is the conditional expectation of the sufficient statistic $T(y, x)$. A single update of the EM algorithm is then given by the recursion

$$\begin{aligned} \theta'' &= \arg\max_{\theta'\in\Omega} Q(\theta', \theta) && (5)\\ &= \arg\max_{\theta'\in\Omega} < g(\theta'), \bar{T} > +d(\theta')\\ &= f(\bar{T}) \end{aligned}$$

Intuitively, we see that the EM update has the same form as the computation of the ML estimate, but with the expected value of the statistic replacing the actual statistic.

Example 3: Let $\{X_n\}_{n=0}^{N-1}$ be i.i.d. random variables with $P\{X_n = 0\} = \pi_0$ and $P\{X_n = 1\} = \pi_1 = 1 - \pi_0$. Let $\{Y_n\}_{n=0}^{N-1}$ be conditionally i.i.d random variables given X, and let the conditional distribution of Y_n given X_n be Gaussian $N(\mu_{X_n}, \sigma_{X_k}^2)$ where μ_0, μ_1, σ_0, and σ_1 are parameters of the distribution. Then the complete set of parameters for the density of (Y, X) are given by

$$\theta = [\mu_0, \mu_1, \sigma_0, \sigma_1, \pi_0] \ .$$

Define the statistics

$$\begin{aligned} N_k &= \sum_{n=0}^{N-1} \delta(x_n - k) \\ t_{1,k} &= \sum_{n=0}^{N-1} y_n \delta(x_n - k) \\ t_{2,k} &= \sum_{n=0}^{N-1} y_n^2 \delta(x_n - k) \end{aligned}$$

where $k \in \{0, 1\}$ and $\delta(\cdot)$ is a Kroniker delta function. We know that if both Y and X are known then the ML estimates are given by

$$\begin{aligned} \hat{\mu}_k &= \frac{t_{1,k}}{N_k} && (6) \\ \hat{\sigma}_k^2 &= \frac{t_{2,k}}{N_k} - \left(\frac{t_{1,k}}{N_k}\right)^2 && (7) \\ \hat{\pi}_k &= \frac{N_k}{N} \ . && (8) \end{aligned}$$

We can express the density function for $p(y|x, \theta)$ by starting with the expressions derived in example 2 for each of the two classes corresponding to $X_n = 0$ and $X_n = 1$.

$$\begin{aligned} p(y|x, \theta) &= \prod_{k=0}^{1} \exp\left\{ \left[\frac{\mu_k}{\sigma_k^2}, -\frac{1}{2\sigma_k^2}\right] \left[\begin{array}{c} t_{1,k} \\ t_{2,k} \end{array}\right] - \frac{N_k}{2\sigma_k^2}\mu_k^2 - \frac{N_k}{2}\log(2\pi\sigma_k^2) \right\} \\ &= \prod_{k=0}^{1} \exp\left\{ \left[\frac{\mu_k}{\sigma_k^2}, -\frac{1}{2\sigma_k^2}, -\frac{\mu_k^2}{2\sigma_k^2}\right], \left[\begin{array}{c} t_{1,k} \\ t_{2,k} \\ N_k \end{array}\right] - \frac{N_k}{2}\log(2\pi\sigma_k^2) \right\} \end{aligned}$$

$$
\begin{aligned}
&= \prod_{k=0}^{1} \exp\left\{ \left[\frac{\mu_k}{\sigma_k^2}, -\frac{1}{2\sigma_k^2}, -\frac{\mu_k^2}{2\sigma_k^2} - \frac{1}{2}\log(2\pi\sigma_k^2) \right], \left[\begin{array}{c} t_{1,k} \\ t_{2,k} \\ N_k \end{array} \right] \right\} \\
&= \exp\left\{ \sum_{k=0}^{1} \left[\frac{\mu_k}{\sigma_k^2}, -\frac{1}{2\sigma_k^2}, -\frac{\mu_k^2}{2\sigma_k^2} - \frac{1}{2}\log(2\pi\sigma_k^2) \right], \left[\begin{array}{c} t_{1,k} \\ t_{2,k} \\ N_k \end{array} \right] \right\}
\end{aligned}
$$

The distribution for X also has exponential form with

$$
\begin{aligned}
p(x|\theta) &= \pi_0^{N_0}\pi_1^{N_1} \\
&= \exp\left\{ \sum_{k=0}^{1} N_k \log \pi_k \right\}
\end{aligned}
$$

This yields to joint density for (Y, X) with the following form.

$$
\begin{aligned}
p(y, x|\theta) &= p(y|x, \theta)p(x|\theta) \\
&= \exp\left\{ \sum_{k=0}^{1} \left[\frac{\mu_k}{\sigma_k^2}, -\frac{1}{2\sigma_k^2}, -\frac{\mu_k^2}{2\sigma_k^2} - \frac{1}{2}\log(2\pi\sigma_k^2) + \log \pi_k \right], \left[\begin{array}{c} t_{1,k} \\ t_{2,k} \\ N_k \end{array} \right] \right\}
\end{aligned}
$$

Therefore, we can see that (Y, X) have an exponential density function. Using the result of equation (5), we know that the EM update must have the form of equations (6), (7), and (8). Where the statistic $T(Y, X)$ is replace with its conditional expectation.

$$
\begin{aligned}
\hat{\mu}_k &\leftarrow \frac{\bar{t}_{1,k}}{\bar{N}_k} && (9) \\
\hat{\sigma}_k &\leftarrow \frac{\bar{t}_{2,k}}{\bar{N}_k} - \left(\frac{\bar{t}_{1,k}}{\bar{N}_k} \right)^2 && (10) \\
\hat{\pi}_k &\leftarrow \frac{\bar{N}_k}{N} && (11)
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{N}_k &= \sum_{n=0}^{N-1} P\{X_n = k|Y = y, \theta\} \\
\bar{t}_{1,k} &= \sum_{n=0}^{N-1} y_n P\{X_n = k|Y = y, \theta\}
\end{aligned}
$$

$$\bar{t}_{2,k} = \sum_{n=0}^{N-1} y_n^2 P\{X_n = k | Y = y, \theta\}$$