# The use of Expectation Maximization (EM) for finding the parameters of a 2D Gaussian given missing data (Example 2, from DHS)

Extra notes for MAS622J/1.126J by Rosalind W. Picard

Let $D$ be a set of data, here containing four samples. One sample has an element that is either missing, known to be corrupted, or is otherwise "bad", $D_b = x_{41} =$ " $*$ ".

$$D = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} = \{\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} * \\ 4 \end{bmatrix}\}$$

A 2-D Gaussian has four parameters, $\theta$, to estimate, for which we make the reasonable starting guess $\theta^0$:

$$\theta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \sigma_1^2 \\ \sigma_2^2 \end{bmatrix} \qquad \theta^0 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Our goal is to iteratively estimate $\theta^i$ until its values converge.

The estimation process consists of two steps: (1) E-step and (2) M-step. In the E-step we wish to formulate the log likelihood for all the data, marginalizing over the possible values for the unknown data. We assume there is a known set of parameters: for this, we use our current best guess of the parameter vector.

$$\begin{aligned} Q(\theta;\theta^0) &= E_{x_{41}}[\ln p(x_g, x_b; \theta) | D_g; \theta^0] \\ &= \int_{-\infty}^{\infty} \left[ \sum_{k=1}^{3} \ln p(\mathbf{x}_k|\theta) + \ln p(\mathbf{x}_4|\theta) \right] p(x_{41}|x_{42} = 4; \theta^0) dx_{41} \end{aligned}$$

Note that $x_{41}$ is independent of all the other $x_{ij} \in D_g$ except for possibly $x_{42}$.
Now let $c = \int p(x_{41}, x_{42}|\theta^0) dx_{41} = p(x_{42}|\theta^0)$. Applying Bayes rule, we get:

$$\begin{aligned} Q(\theta;\theta^0) &= \int_{-\infty}^{\infty} \left[ \sum_{k=1}^{3} \ln p(\mathbf{x}_k|\theta) + \ln p(\mathbf{x}_4|\theta) \right] \frac{p(x_{41}, x_{42}; \theta^0)}{c} dx_{41} \\ &= \sum_{k=1}^{3} \ln p(\mathbf{x_k}|\theta) + \int_{-\infty}^{\infty} \ln p \left( \begin{bmatrix} x_{41} \\ 4 \end{bmatrix} |\theta \right) \frac{p(x_{41}, x_{42}|\theta^0)}{c} dx_{41} \end{aligned}$$

Also, since $c$ is not a function of the bad data, $x_{41}$, we can pull it out of the integral:

$$
\begin{aligned}
&Q(\theta;\theta^0)\\
&= \sum_{k=1}^{3}\ln p(\mathbf{x_k}|\theta) + \frac{1}{c}\int_{-\infty}^{\infty}\ln p\left(\left[\begin{array}{c} x_{41} \\ 4 \end{array}\right]|\theta\right)\frac{1}{2\pi^{d/2}\left|\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right|^{1/2}}e^{-\frac{(x_{41}-0)^2}{2}}e^{-\frac{(4-0)^2}{2}}dx_{41}\\
&= \sum_{k=1}^{3}\ln p(\mathbf{x_k}|\theta) + \frac{1}{c}\int_{-\infty}^{\infty}\ln\left[\frac{1}{2\pi\left|\begin{array}{cc} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{array}\right|^{1/2}}e^{-\frac{1}{2}\left(\begin{array}{c} x_{41}-\mu_1 \\ 4-\mu_2 \end{array}\right)^T\left[\begin{array}{cc} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{array}\right]\left(\begin{array}{c} x_{41}-\mu_1 \\ 4-\mu_2 \end{array}\right)}\right]\\
&\quad * \left[\frac{1}{2\pi}e^{-\frac{(x_{41}^2+16)}{2}}\right]dx_{41}
\end{aligned}
$$

Rewrite the log term in the integral above:

$$-\ln 2\pi\sigma_1\sigma_2 - \frac{(x_{41}-\mu_1)^2}{2\sigma_1^2} - \frac{(4-\mu_2)^2}{2\sigma_2^2}$$

or

$$-\ln 2\pi\sigma_1\sigma_2 - \frac{(x_{41}^2-2x_{41}\mu_1+\mu_1^2)}{2\sigma_1^2} - \frac{(4-\mu_2)^2}{2\sigma_2^2}$$

Now we can separate out the terms that involve the bad data, $x_{41}$, and rewrite $Q$ as:

$$
\begin{aligned}
&Q(\theta;\theta^0) =\\
&\sum_{k=1}^{3}\ln p(\mathbf{x_k}|\theta) + \left[-\ln 2\pi\sigma_1\sigma_2 - \frac{\mu_1^2}{2\sigma_1^2} - \frac{(4-\mu_2)^2}{2\sigma_2^2}\right]\left[\frac{1}{c}\int_{-\infty}^{\infty}\frac{1}{2\pi}e^{\frac{-(x_{41}^2+16)}{2}}dx_{41}\right] + III
\end{aligned}
$$

where $III$ contains the remaining two terms involving $-x_{41}^2$ and $2x_{41}\mu_1$. Note first that $\frac{1}{c}$ multiplying the integral causes that whole term to go to 1. Subsequently,

$$
\begin{aligned}
Q(\theta;\theta^0) &= I + II + III\\
&= \sum_{k=1}^{3}\ln p(\mathbf{x_k}|\theta) + \left[-\ln 2\pi\sigma_1\sigma_2 - \frac{\mu_1^2}{2\sigma_1^2} - \frac{(4-\mu_2)^2}{2\sigma_2^2}\right] + III
\end{aligned}
$$

Now, let's look at that last term:

$$III = \frac{1}{c}\int_{-\infty}^{\infty} -\frac{x_{41}^2}{2\sigma_1^2}\frac{1}{2\pi}e^{\frac{-(x_{41}^2+16)}{2}}dx_{41} + \frac{1}{c}\int_{-\infty}^{\infty} -\frac{2x_{41}\mu_1}{2\sigma_1^2}\frac{1}{2\pi}e^{\frac{-(x_{41}^2+16)}{2}}dx_{41}$$

The second integral above consists of an odd multiplier, $\frac{2x_{41}\mu_1}{2\sigma_1^2}$, times an even exponential (even with respect to $x_{41}$). Thus, the integral is odd and integrates to zero. This leaves only the first integral:

$$
\begin{aligned}
III &= \frac{1}{c}\int_{-\infty}^{\infty} -\frac{x_{41}^2}{2\sigma_1^2}\frac{1}{2\pi}e^{\frac{-(x_{41}^2+16)}{2}}dx_{41} \\
&= \frac{-\frac{e^{-8}}{2\sigma_1^2}\int_{-\infty}^{\infty} -\frac{x_{41}^2}{2\pi}e^{-\frac{x_{41}^2}{2}}dx_{41}}{e^{-8}\int_{-\infty}^{\infty}\frac{1}{2\pi}e^{-\frac{x_{41}^2}{2}}dx_{41}}
\end{aligned}
$$

Let $z = \frac{x_{41}}{\sqrt{2}}$ so $dx_{41} = \sqrt{2}dz$, and cancel the $\frac{e^{-8}}{2\pi}$ in the numerator and denominator. Then:

$$
III = \frac{-\frac{1}{2\sigma_1^2}\int_{-\infty}^{\infty} 2z^2e^{-z^2}\sqrt{2}dz}{\int_{-\infty}^{\infty} e^{-z^2}\sqrt{2}dz}
$$

Integral tables come in handy at this point. From standard tables, we find these:

$$
\int_0^{\infty} x^2e^{-x^2}dx = \sqrt{\frac{\pi}{16}} \qquad \int_0^{\infty} e^{-x^2}dx = \sqrt{\frac{\pi}{4}}
$$

Using these relationships, we simplify:

$$
\begin{aligned}
III &= \frac{\frac{-1}{\sigma_1^2}(2\sqrt{2}\sqrt{\frac{\pi}{16}})}{2\sqrt{2}\sqrt{\frac{\pi}{4}}} \\
&= \frac{-1}{\sigma_1^2}\sqrt{\frac{4}{16}} = -\frac{1}{2\sigma_1^2}
\end{aligned}
$$

Now, return to (1) substituting in this simplified $III$ to write Q:

$$
Q(\theta;\theta^0) = \sum_{k=1}^{3} \ln p(\mathbf{x_k}|\theta) - \ln 2\pi\sigma_1\sigma_2 - \frac{1+\mu_1^2}{2\sigma_1^2} - \frac{(4-\mu_2)^2}{2\sigma_2^2}
$$

Equation (2) is called the E-step, the first step of Expectation-Maximization.

Let's expand the first term in (2) and then we will be ready for the M-step, in which we'll maximize Q:

$$
Q(\theta;\theta^0) = \sum_{k=1}^{3}\left(-\ln 2\pi\sigma_1\sigma_2 - \frac{(x_{k1}-\mu_1)^2}{2\sigma_1^2} - \frac{(x_{k2}-\mu_2)^2}{2\sigma_2^2}\right) - \ln 2\pi\sigma_1\sigma_2 - \frac{1+\mu_1^2}{2\sigma_1^2} - \frac{(4-\mu_2)^2}{2\sigma_2^2}
$$

Now we're ready for the Maximization (M-step):

$$
\theta^1 = \arg\max_{\theta} Q(\theta;\theta^0)
$$

We need to take derivatives of Q w.r.t. each of its four parameters, set the derivatives equal to zero, and solve for the new parameter values.

$$\begin{aligned}
\frac{dQ(\theta;\theta^0)}{d\mu_1} &= \frac{2}{2\sigma_1^2}\sum_{k=1}^{3}(x_{k1}-\mu_1) - \frac{2\mu_1}{2\sigma_1^2} = 0 \\
\sum_{k=1}^{3} x_{k1} &= 4\mu_1 \\
\mu_1 &= \frac{0+1+2}{4} = \frac{3}{4}
\end{aligned}$$

$$\begin{aligned}
\frac{dQ(\theta;\theta^0)}{d\mu_2} &= \frac{1}{\sigma_2^2}\sum_{k=1}^{3}(x_{k2}-\mu_2) - \frac{(4-\mu_2)}{\sigma_2^2} = 0 \\
\mu_2 &= \frac{1}{4}(2+0+2+4) = 2
\end{aligned}$$

$$\begin{aligned}
\frac{dQ(\theta;\theta^0)}{d\sigma_1} &= \frac{-4(2\pi\sigma_2)}{2\pi\sigma_1\sigma_2} + \sum_{k=1}^{3}\frac{(x_{k1}-\mu_1)^2}{\sigma_1^3} + \frac{(1+\mu_1^2)}{\sigma_1^3} = 0 \\
\frac{4\sigma_1^3}{\sigma_1} &= (0-\mu_1)^2 + (1-\mu_1)^2 + (2-\mu_1)^2 + (1+\mu_1^2) \\
\sigma_1^2 &= \frac{1}{4}\left(\frac{9}{16}+\frac{1}{16}+\frac{25}{16}+\frac{16}{16}+\frac{9}{16}\right) = \frac{15}{16} = 0.938
\end{aligned}$$

Similarly, solving $\frac{dQ(\theta;\theta^0)}{d\sigma_2} = 0$ leads to the solution $\sigma_2^2 = 2$.

Now we have completed the M-Step and we have the next iteration of our estimate of the parameters:

$$\theta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \sigma_1^2 \\ \sigma_2^2 \end{bmatrix} \qquad \theta^0 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \qquad \theta^1 = \begin{bmatrix} 3/4 \\ 2 \\ 0.938 \\ 2 \end{bmatrix}$$

We use this to form $Q(\theta;\theta^1)$ and go back to the beginning of the E-step. Repeat this process until the parameters converge.

The E-M algorithm guarantees that the log-likelihood of the good data (with the bad data marginalized) will increase monotonically.

Fun facts: E-M is a Maximum Likelihood method, not a fully Bayesian method. Also, the Baum-Welch (forward-backward) algorithm used for training HMM's is an example of the E-M method.