# EM Algorithm Implementation Details

### Assumptions and derivations

*Note: I am using notation from Casella and Berger for convenience.*

We assume independent samples $(X_1, Y_1), ..., (X_n, Y_n)$ are drawn from the bivariate normal pdf with means $\mu_X$ and $\mu_Y$, variances $\sigma_X^2$ and $\sigma_Y^2$, and correlation $\rho$.

The bivariate normal pdf can be expressed as:

$$\begin{aligned} f(x, y|\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) = {} & (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1} \\ & \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left((\frac{x-\mu_X}{\sigma_X})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^2\right)\right\} \end{aligned}$$

We aim to derive the MLEs of $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho$ when all five parameters are unknown.

One method of finding the MLEs is to use properties of exponential family distributions. Exponential families have the following pdf form:

$$f(x|\theta) = h(x)c(\theta)\text{exp}\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right)$$

where $\theta$ is a vector of parameters

We re-arrange the bivariate normal pdf to show exponential family form:

$$\begin{aligned} f(x, y|\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) = {} & (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1}\text{exp}\left\{\frac{-1}{2(1-\rho^2)}\left(\frac{\mu_X^2}{\sigma_X^2} - \frac{2\rho\mu_X\mu_Y}{\sigma_X\sigma_Y} + \frac{\mu_Y^2}{\sigma_Y^2}\right)\right\} \\ & \times \exp\left\{\frac{1}{1-\rho^2}(\frac{\mu_X}{\sigma_X^2} - \frac{\rho\mu_Y}{\sigma_X\sigma_Y})x + \frac{1}{1-\rho^2}(\frac{\mu_Y}{\sigma_Y^2} - \frac{\rho\mu_X}{\sigma_X\sigma_Y})y\right\} \\ & \times \exp\left\{-\frac{1}{2\sigma_X^2(1-\rho^2)}x^2 - -\frac{1}{2\sigma_Y^2(1-\rho^2)}y^2 + \frac{\rho}{\sigma_X\sigma_Y(1-\rho^2)}xy\right\} \end{aligned}$$

where $h(x, y)$ is 1,
$c(\theta)$ is the first line in the above equation,
and $w_i(\theta)$ and $t_i(x, y)$ are clearly found inside the exponent terms on lines 2 and 3

Casella and Berger discuss the ability to derive MLEs directly from pdfs in exponential family form[1]. We solve the system of $k$ equations:

$$\sum_{j=1}^{n} t_i(x_j, y_j) = \mathrm{E}\left(\sum_{j=1}^{n} t_i(x_j, y_j)\right), \text{for } i = 1, ..., k$$

[1] Chapter 7. Point Estimation. Miscellanea pp. 367-368.

The above equations can be solved for $\hat{w}_i(\theta)$, or by invariance, any one-to-one function $g(\hat{w}_i(\theta))$. The Miscellanea discusses the correspondence between MLEs and method of moments estimators in exponential family distributions.

Therefore the $k = 5$ equations below can be solved to find the MLEs of the bivariate Normal parameters, $\mu_X$, $\mu_Y$, $\sigma_X^2$, $\sigma_Y^2$, and $\rho$.

$$\begin{aligned}
\sum_{j=1}^{n} t_1(x_j, y_j) = \sum_{j=1}^{n} x_j = \mathrm{E}\left(\sum_{j=1}^{n} x_j\right) &= n\mu_X \\
\sum_{j=1}^{n} t_2(x_j, y_j) = \sum_{j=1}^{n} y_j = \mathrm{E}\left(\sum_{j=1}^{n} y_j\right) &= n\mu_Y \\
\sum_{j=1}^{n} t_3(x_j, y_j) = \sum_{j=1}^{n} x_j^2 = \mathrm{E}\left(\sum_{j=1}^{n} x_j^2\right) &= n(\sigma_X^2 + \mu_X^2) \\
\sum_{j=1}^{n} t_4(x_j, y_j) = \sum_{j=1}^{n} y_j^2 = \mathrm{E}\left(\sum_{j=1}^{n} y_j^2\right) &= n(\sigma_Y^2 + \mu_Y^2) \\
\sum_{j=1}^{n} t_5(x_j, y_j) = \sum_{j=1}^{n} x_j y_j = \mathrm{E}\left(\sum_{j=1}^{n} x_j y_j\right) &= n(\rho\sigma_X\sigma_Y + \mu_X mu_Y)
\end{aligned}$$

The solutions are found easily for $\hat{\mu}_X$ and $\hat{\mu}_Y$ by dividing the first two equations by $n$. With some algebra, we find the solutions for $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$, and $\hat{\rho}$.

$$\begin{aligned}
\hat{\mu}_X &= \frac{\sum_{j=1}^{n} x_j}{n} = \bar{x} \\
\hat{\mu}_Y &= \frac{\sum_{j=1}^{n} y_j}{n} = \bar{y} \\
\hat{\sigma}_X^2 &= \frac{\sum_{j=1}^{n} (x_j - \bar{x})^2}{n} \\
\hat{\sigma}_Y^2 &= \frac{\sum_{j=1}^{n} (y_j - \bar{y})^2}{n} \\
\hat{\rho} = \frac{1}{n} &\frac{\sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y})}{\hat{\sigma}_X\hat{\sigma}_Y}
\end{aligned}$$

### Algorithm setup

Our missing data problem is that some observations are un-matched. In effect we do not estimate $\sum_{i=1}^{n} x_i y_i$ completely, but only partially for some subset of observations, $m$ ($m < n$).

For the EM algorithm we must choose a representation of our complete data and our partially-observed data. I will use $A$ to represent the vector of complete data quantities, and $B$ to represent the vector of partially-observed data quantities. Formally:

$$\begin{aligned}
A = \begin{bmatrix} \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} y_i^2 & \sum_{i=1}^{n} x_i y_i \end{bmatrix} \\
B = \begin{bmatrix} A_1 & \dots & \dots & A_4 & \sum_{i=1}^{m} x_i y_i \end{bmatrix}
\end{aligned}$$

Thus the only difference between $A$ and $B$ is the last element; in $B$ we only observe the sum of the product of $m$ terms, rather than $n$ terms.

The following vector, $\theta$, represents all unknown parameters for which we aim to find the MLEs given our partially-observed data.

$$\theta = \begin{bmatrix} \mu_X & \mu_Y & \sigma_X^2 & \sigma_Y^2 & \rho \end{bmatrix}$$

We use the iterative technique of the EM algorithm for finding the MLEs. In each iteration there are two steps, the **E-Step** and the **M-Step**. Note there are convenient properties of the EM algorithm when applied to exponential family distributions[2].

In the **E-Step**, we find the expected value of the sufficient statistics, $E[A]$, given our partially-observed data, $B$, and some set of parameters, $\theta^{(p)}$, at iteration $p$.

$$E[A|B,\theta^{(p)}] = \begin{bmatrix} \sum_{i=1}^n x_i & \sum_{i=1}^n y_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n y_i^2 & E[\sum_{i=1}^n x_i y_i] \end{bmatrix}$$

Note that the expected value of the missing data quantity, $\sum_{i=1}^n x_i y_i$, is simply the sum of the product of $m$ matched terms added to the expected value of the sum of the product of $n-m$ un-matched terms, using estimates of $\mu_X$, ..., $\sigma_Y$ from the un-matched terms only. This is to enforce a constraint that the resulting estimate of $\rho$ does not exceed the bounds of $[-1,1]$.

$$E\left[\sum_{i=1}^n x_i y_i | B, \theta^{(p)}\right] = \sum_{i=1}^m x_i y_i + (n-m)\left(\rho^{(p)}\sigma_X^{'}\sigma_Y^{'} + \mu_X^{'}\mu_Y^{'}\right)$$

where $\sigma_X^{'}$, $\sigma_Y^{'}$, $\mu_X^{'}$, $\mu_Y^{'}$ are MLEs obtained from un-matched samples only.

In the **M-Step**, we find the values of $\theta$ that maximize the likelihood given our partially-observed data. Since the MLEs of $\mu_X$ and $\sigma_X^2$ depend only on $\sum x_i$ and $\sum x_i^2$ (and likewise for $\mu_Y$ and $\sigma_Y^2$), the only parameter to update in this step is $\rho$.

$$\rho^{(p+1)} = \frac{1}{n}\left(\frac{E[\sum_{i=1}^n x_i y_i] - n\bar{x}\bar{y}}{\sigma_X \sigma_Y}\right)$$

[2] Nan Laird. Handbook of Statistics (1993). Chapter 14: The EM Algorithm. pp. 512-513.

**Examples - Issues resolved as of 11/10**

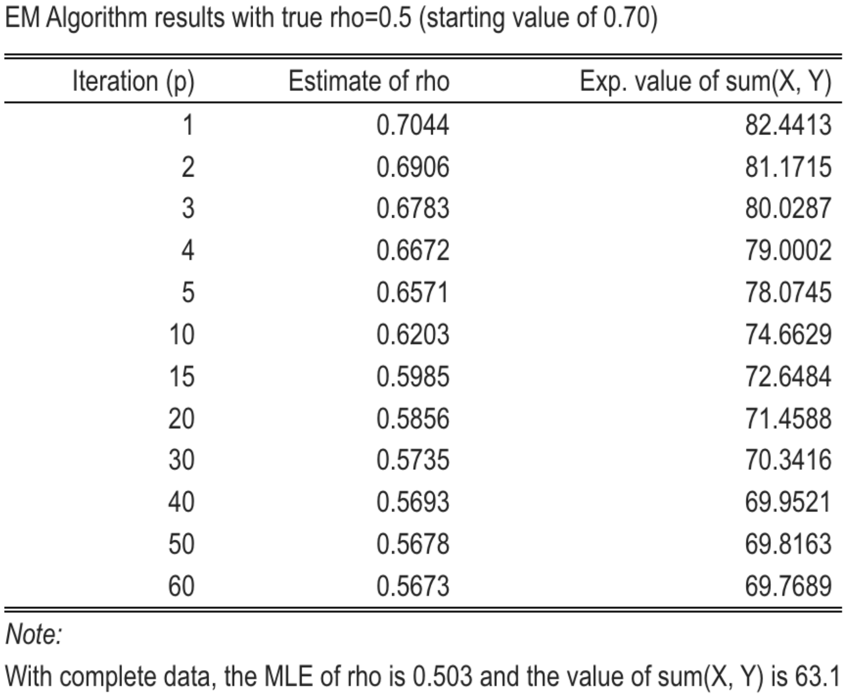
In the following example I simulated data from bivariate normal with the following parameters.

$$\begin{aligned} \mu_X &= 0 \\ \mu_Y &= 0.5 \\ \sigma_X^2 = \sigma_Y^2 &= 1 \\ \rho &= 0.5 \end{aligned}$$

I limited the number of matched samples to 10% of the observations ($m = 10$, $n = 100$).

For starting values of $\theta$, I used the MLEs of $\mu_X$, $\mu_Y$, $\sigma_X^2$, $\sigma_Y^2$, and the sample correlation of the $m$ matched samples.

The table below presents the iterative estimates of $\hat{\rho}$ and the expected value of $\sum_{i=1}^{n} x_i y_i$.



EM Algorithm results with true rho=0.5 (starting value of 0.70)

| Iteration (p) | Estimate of rho | Exp. value of sum(X, Y) |
|---:|---:|---:|
| 1 | 0.7044 | 82.4413 |
| 2 | 0.6906 | 81.1715 |
| 3 | 0.6783 | 80.0287 |
| 4 | 0.6672 | 79.0002 |
| 5 | 0.6571 | 78.0745 |
| 10 | 0.6203 | 74.6629 |
| 15 | 0.5985 | 72.6484 |
| 20 | 0.5856 | 71.4588 |
| 30 | 0.5735 | 70.3416 |
| 40 | 0.5693 | 69.9521 |
| 50 | 0.5678 | 69.8163 |
| 60 | 0.5673 | 69.7689 |

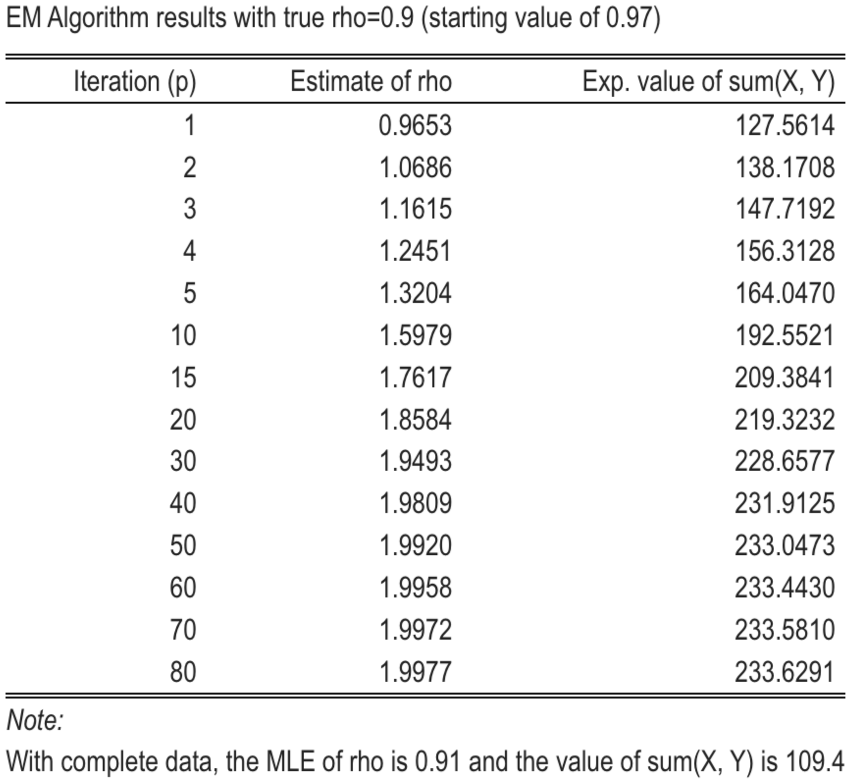
*Note:*
With complete data, the MLE of rho is 0.503 and the value of sum(X, Y) is 63.1

After about 60 iterations, we see that the estimate of $\hat{\rho}$ eventually converges to 0.5673; this is a more conservative estimate compared to the sample correlation of the $m$ matched samples (0.70). However it does not quite achieve the MLE under the complete data scenario, which would be $\hat{\rho} = 0.503$ (we would not expect to achieve the complete data MLE).

Unfortunately, the algorithm breaks under different conditions, for example when I simulated data with the following parameters.

$$\begin{aligned} \mu_X &= 0 \\ \mu_Y &= 0.5 \\ \sigma_X^2 = \sigma_Y^2 &= 1 \\ \rho &= 0.9 \end{aligned}$$

This time the true correlation was more extreme, as was the starting value (the sample correlation in the $m$ matched samples was 0.97).



EM Algorithm results with true rho=0.9 (starting value of 0.97)

| Iteration (p) | Estimate of rho | Exp. value of sum(X, Y) |
|---:|---:|---:|
| 1 | 0.9653 | 127.5614 |
| 2 | 1.0686 | 138.1708 |
| 3 | 1.1615 | 147.7192 |
| 4 | 1.2451 | 156.3128 |
| 5 | 1.3204 | 164.0470 |
| 10 | 1.5979 | 192.5521 |
| 15 | 1.7617 | 209.3841 |
| 20 | 1.8584 | 219.3232 |
| 30 | 1.9493 | 228.6577 |
| 40 | 1.9809 | 231.9125 |
| 50 | 1.9920 | 233.0473 |
| 60 | 1.9958 | 233.4430 |
| 70 | 1.9972 | 233.5810 |
| 80 | 1.9977 | 233.6291 |

*Note:*

With complete data, the MLE of rho is 0.91 and the value of sum(X, Y) is 109.4

We see that the algorithm quickly leads to solutions outside the parameter space, since $\rho > 1$ is not possible. I discuss this in the issues below.

**Current issues with the approach**

There is likely a mistake in my implementation of the algorithm, since the EM algorithm should yield successive iterates that lie in the parameter space, provided the starting values lie in the interior of the parameter space[3].

Secondly, the current approach does not provide an estimate of the standard error of $\hat{\theta}$, which would be required to conduct any hypothesis testing involving the MLEs. Louis's method may present an option for finding the observed information matrix, $I_{obs}(\hat{\theta})$, under the EM algorithm[4].

[3] Nan Laird. Handbook of Statistics (1993). Chapter 14: The EM Algorithm. p. 513.

[4] TA Louis. Finding the observed information matrix when using the EM algorithm (1982). Journal of the Royal Statistical Society.