

EM Algorithm Implementation Details

Assumptions and derivations

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

We assume independent samples $(X_1, Y_1), \dots, (X_n, Y_n)$ are drawn from the bivariate normal pdf with means μ_X and μ_Y , variances σ_X^2 and σ_Y^2 , and correlation ρ .

The bivariate normal pdf can be expressed as:

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

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:

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

where θ is a vector of parameters

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

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

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 form¹. We solve the system of k equations:

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

¹Chapter 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, μ_X , μ_Y , σ_X^2 , σ_Y^2 , and ρ .

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

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:

$$\begin{aligned} A &= \left[\sum_{i=1}^n x_i \quad \sum_{i=1}^n y_i \quad \sum_{i=1}^n x_i^2 \quad \sum_{i=1}^n y_i^2 \quad \sum_{i=1}^n x_i y_i \right] \\ B &= \left[A_1 \quad \dots \quad \dots \quad A_4 \quad \sum_{i=1}^m x_i y_i \right] \end{aligned}$$

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, θ , 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 distributions².

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 μ_X , ..., σ_Y from the un-matched terms only. This is to enforce a constraint that the resulting estimate of ρ 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 σ'_X , σ'_Y , μ'_X , μ'_Y are MLEs obtained from un-matched samples only.

In the **M-Step**, we find the values of θ that maximize the likelihood given our partially-observed data. Since the MLEs of μ_X and σ_X^2 depend only on $\sum x_i$ and $\sum x_i^2$ (and likewise for μ_Y and σ_Y^2), the only parameter to update in this step is ρ .

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

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