EM algorithm for a toy case

Aude Sportisse

October 2021

We consider $X \sim \mathcal{N}(\mu, \Sigma)$, with

$$\mu = \begin{pmatrix} 5 \\ -1 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

We consider that X_2 contain some MCAR values. We want to estimate the parameters (μ, Σ).

1 First option: directly maximize the observed likelihood

In this simple case, we can directly maximize the observed likelihood. As the missing values are MCAR, we can ignore the missing-data mechanism and just consider the following optimization problem:

$$\hat{\theta} = \mathrm{argmax}_\theta \, \ell_{\mathrm{ign}}(\theta; X^{\mathrm{obs}}) = \log(p(X^{\mathrm{obs}}; \theta)),$$

with $\hat{\theta} = (\hat{\mu}, \hat{\Sigma})$ and $\theta = (\mu, \Sigma)$.

We have seen (cf slides) that the observed likelihood is

$$\begin{aligned}
\ell(\mu, \Sigma; X_{.1}, X_{.2}) = &-\frac{n}{2}\log(\sigma_{11}^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(X_{i1}-\mu_1)^2}{\sigma_{11}^2} \\
&- \frac{r}{2}\log\left(\sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}}\right)^2 - \frac{1}{2}\sum_{i=1}^{r}\frac{(X_{i2}-\mu_2+\frac{\sigma_{21}}{\sigma_{11}}(X_{i1}-\mu_1))^2}{\left(\sigma_{22}-\frac{\sigma_{21}^2}{\sigma_{11}}\right)^2}
\end{aligned}$$

Let us compute the gradients of $\ell(\mu, \Sigma; X_{.1}, X_{.2})$. **To simplify, we just derive the computations for the estimation of μ_1 and μ_2.**

$$\begin{aligned}
\nabla_{\mu_1}\ell(\mu, \Sigma; X_{.1}, X_{.2}) &= \sum_{i=1}^{n}\frac{X_{i1}-\mu_1}{\sigma_{11}^2} \\
\nabla_{\mu_2}\ell(\mu, \Sigma; X_{.1}, X_{.2}) &= \sum_{i=1}^{r}\frac{X_{i2}-\mu_2+\frac{\sigma_{21}}{\sigma_{11}}(X_{i1}-\mu_1)}{\left(\sigma_{22}-\frac{\sigma_{21}^2}{\sigma_{11}}\right)^2}
\end{aligned}$$

$$\nabla_{\mu_1}\ell(\mu,\Sigma;X_{.1},X_{.2}) = 0 \Longleftrightarrow \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{i1}$$

$$\nabla_{\mu_2}\ell(\mu,\Sigma;X_{.1},X_{.2}) = 0 \Longleftrightarrow \hat{\mu}_2 = \frac{1}{r}\sum_{i=1}^{r} X_{i2} + \frac{\sigma_{21}}{\sigma_{11}}\sum_{i=1}^{r} X_{i1} - \frac{\sigma_{21}}{\sigma_{11}}\hat{\mu}_1,$$

where we have used plug-in to get an estimator $\hat{\mu}_2$ which only involves known quantities. (We also skip the proof of concavity for ℓ.)

2 Second option: derive an EM algorithm

It is an iterative algorithm, starting from an initial point θ^0, there are two steps iteratively proceeded:

- **E-step:** computation of the expected full log-likelihood **over the distribution of the missing data given the observed data** and a current value of the parameters.

$$Q(\theta;\theta^r) = \mathbb{E}[\ell_{\text{full}}(X;\theta)|X^{\text{obs}},\theta^r] = \int \ell_{\text{full}}(X;\theta)p(X^{\text{mis}}|X^{\text{obs}};\theta^r)dX^{\text{mis}}$$

- **M-step:** maximization of $Q(\theta;\theta^r)$ over θ.

$$\theta^{r+1} = \text{argmax}_\theta\, Q(\theta;\theta^r)$$

2.1 E-step

First, we have to write the full log-likelihood (it is an easy thing to do) $\ell_{\text{full}}(X;\theta)$.
As $X \sim \mathcal{N}(\mu,\Sigma)$, we have

$$\begin{aligned}\ell_{\text{full}}(X;\theta) &= -\frac{n}{2}\log(\det(\Sigma)) - \frac{1}{2}\sum_{i=1}^{n}\begin{pmatrix}x_{i1}-\mu_1 & x_{i2}-\mu_2\end{pmatrix}\Sigma^{-1}\begin{pmatrix}x_{i1}-\mu_1 & x_{i2}-\mu_2\end{pmatrix}^T \\ &= \frac{n}{2}\log(\det(\Sigma)) - \frac{1}{2}\sum_{i=1}^{n}(x_{i1}-\mu_1)^2\tilde{\sigma}_{11} + 2(x_{i1}-\mu_1)(x_{i2}-\mu_2)\tilde{\sigma}_{12}^2 + (x_{i2}-\mu_2)^2\tilde{\sigma}_{22},\end{aligned}$$

with $\Sigma^{-1} = \begin{pmatrix}\tilde{\sigma}_{11} & \tilde{\sigma}_{12}\\ \tilde{\sigma}_{12} & \tilde{\sigma}_{22}\end{pmatrix}$.

If we develop this expression, we obtain that $\ell_{\text{full}}(X;\theta)$ is a linear function of some terms, usually called *sufficient statistics* which are $\sum_{i=1}^n x_{i1}$, $\sum_{i=1}^n x_{i1}^2$, $\sum_{i=1}^n x_{i2}$, $\sum_{i=1}^n x_{i2}^2$ and $\sum_{i=1}^n x_{i1}x_{i2}$.

Thus, the E-step just calculates the conditional expectation of these terms over the distribution of the missing data given the observed data: $\mathbb{E}\left[\sum_{i=1}^n x_{i1}|X^{\text{obs}},\theta^r\right]$,

$\mathbb{E}\left[\sum_{i=1}^n x_{i1}^2|X^{\text{obs}},\theta^r\right]$, $\mathbb{E}\left[\sum_{i=1}^n x_{i2}|X^{\text{obs}},\theta^r\right]$, $\mathbb{E}\left[\sum_{i=1}^n x_{i2}^2|X^{\text{obs}},\theta^r\right]$ and $\mathbb{E}\left[\sum_{i=1}^n x_{i1}x_{i2}|X^{\text{obs}},\theta^r\right]$. Remark that the observed variables X^{obs} is here the first variable X_1. We have

$$\mathbb{E}\left[\sum_{i=1}^n x_{i1}|X_1,\theta^r\right] = \sum_{i=1}^n \int x_{i1}p(x_{i2}^{\text{mis}}|x_{i1};\theta^r)dx_{i2}^{\text{mis}} = \sum_{i=1}^n x_{i1},$$

because x_{i1} can be taken out of the integral and the integral is equal to 1 (by definition, a probability density function must integrate to one). Similarly, we have

$$\mathbb{E}\left[\sum_{i=1}^n x_{i1}^2|X_1,\theta^r\right] = \sum_{i=1}^n x_{i1}^2,$$

For the other sufficient statistics, it is more difficult !

$$\begin{aligned}\mathbb{E}\left[x_{i2}|X_1,\theta^r\right] &= \int x_{i2}p(x_{i2}^{\text{mis}}|x_{i1};\theta^r)dx_{i2}^{\text{mis}} \\ &= \begin{cases} x_{i2} & \text{if } x_{i2} \text{ is observed} \\ \int x_{i2}^{\text{mis}}p(x_{i2}^{\text{mis}}|x_{i1};\theta^r)dx_{i2}^{\text{mis}} & \text{otherwise.}\end{cases}\end{aligned}$$

For the first case (x_{i2} is observed), we use the fact that x_{i2} can be taken out of the integral and the integral is equal to 1. For the second case, it is just the expectation of the conditional distribution of X_2 given X_1. For this, we can use the classical formulae for a bivariate Gaussian variable given as follows:

$$X_{i2}|X_{i1} \sim \mathcal{N}(\mathbb{E}[X_{i2}|X_{i1}], \text{Var}(X_{i2}|X_{i1}))$$

with

$$\begin{aligned}\mathbb{E}[X_{i2}|X_{i1}] &= \mu_2 + \frac{\sigma_{21}}{\sigma_{11}}(X_{i1}-\mu_1) \\ \text{Var}(X_{i2}|X_{i1}) &= \sigma_{22} - \frac{\sigma_{21}^2}{\sigma_{11}}\end{aligned}$$

Thus, we have

$$\begin{aligned}\mathbb{E}\left[x_{i2}|X_1,\theta^r\right] &= \int x_{i2}p(x_{i2}^{\text{mis}}|x_{i1};\theta^r)dx_{i2}^{\text{mis}} \\ &= \begin{cases} x_{i2} & \text{if } x_{i2} \text{ is observed} \\ \mu_2^r + \frac{\sigma_{21}^r}{\sigma_{11}^r}(x_{i1}-\mu_1^r) & \text{otherwise.}\end{cases}\end{aligned}$$

The same strategy is used to compute the other sufficient statistics. We obtain

$$\begin{aligned}\mathbb{E}\left[x_{i2}^2|X_1,\theta^r\right] &= \begin{cases} x_{i2}^2 & \text{if } x_{i2} \text{ is observed} \\ (\mathbb{E}[x_{i2}|x_{i1}])^2 + \text{Var}(X_{i2}|X_{i1}) & \text{otherwise.}\end{cases} \\ &= \begin{cases} x_{i2}^2 & \text{if } x_{i2} \text{ is observed} \\ \left(\mu_2^r + \frac{\sigma_{21}^r}{\sigma_{11}^r}(x_{i1}-\mu_1^r)\right)^2 + \sigma_{22}^r - \frac{(\sigma_{21}^r)^2}{\sigma_{11}^r} & \text{otherwise.}\end{cases}\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}\left[x_{i1}x_{i2}|X_1,\theta^r\right] &= \begin{cases} x_{i1}x_{i2} & \text{if } x_{i2} \text{ is observed} \\ x_{i1}\mathbb{E}[x_{i2}|x_{i1}] & \text{otherwise.} \end{cases} \\
&= \begin{cases} x_{i1}x_{i2} & \text{if } x_{i2} \text{ is observed} \\ x_{i1}\left(\mu_2^r + \frac{\sigma_{21}^r}{\sigma_{11}^r}(x_{i1}-\mu_1^r)\right) & \text{otherwise.} \end{cases}
\end{aligned}
$$

Let us assume to simplify the computations that the m first values in X_2 are missing. Finally, we have

$$
\begin{aligned}
s_1 &= \mathbb{E}\left[\sum_{i=1}^{n} x_{i1}|X_1,\theta^r\right] = \sum_{i=1}^{n} x_{i1} \\
s_{11} &= \mathbb{E}\left[\sum_{i=1}^{n} x_{i1}^2|X_1,\theta^r\right] = \sum_{i=1}^{n} x_{i1}^2
\end{aligned}
$$

$$
s_2 = \mathbb{E}\left[\sum_{i=1}^{n} x_{i2}|X_1,\theta^r\right] = \sum_{i=m+1}^{n} x_{i2} + \sum_{i=1}^{m}\left(\mu_2^r + \frac{\sigma_{21}^r}{\sigma_{11}^r}\left(x_{i1}-\mu_1^r\right)\right)
$$

$$
s_{22} = \mathbb{E}\left[\sum_{i=1}^{n} x_{i2}^2|X_1,\theta^r\right] = \sum_{i=m+1}^{n} x_{i2}^2 + \sum_{i=1}^{m}\left(\left(\mu_2^r + \frac{\sigma_{21}^r}{\sigma_{11}^r}(x_{i1}-\mu_1^r)\right)^2 + \sigma_{22}^r - \frac{(\sigma_{21}^r)^2}{\sigma_{11}^r}\right)
$$

$$
s_{12} = \mathbb{E}\left[\sum_{i=1}^{n} x_{i1}x_{i2}|X_1,\theta^r\right] = \sum_{i=m+1}^{n} x_{i1}x_{i2} + \sum_{i=1}^{m} x_{i1}\left(\mu_2^r + \frac{\sigma_{21}^r}{\sigma_{11}^r}\left(x_{i1}-\mu_1^r\right)\right)
$$

2.2 M-step

It is easy to maximize $Q(\theta;\theta^r)$! We do not detail the computations because it is really the usual ML estimates.

$$
\begin{aligned}
\mu_1^{r+1} &= \frac{s_1}{n} \\
\mu_2^{r+1} &= \frac{s_2}{n} \\
\sigma_{11}^{r+1} &= \frac{s_{11}}{n} - (\mu_1^{r+1})^2 \\
\sigma_{22}^{r+1} &= \frac{s_{22}}{n} - (\mu_2^{r+1})^2 \\
\sigma_{12}^{r+1} &= \frac{s_{12}}{n} - (\mu_1^{r+1}\mu_2^{r+1})
\end{aligned}
$$