Exercise 1 (Expectation Maximization algorithm) In the case where we are interested in estimating unknown parameters $\theta \in \mathbb{R}^m$ characterizing a model with missing data, the Expectation Maximization (EM) algorithm (Dempster et al. 1977) can be used when the joint distribution of the missing data Y and the observed data X is explicit. For all $\theta \in \mathbb{R}^m$, let p_θ be the probability density function of (X, Y) when the model is parameterized by θ with respect to a given reference measure μ. The EM algorithm aims at computing iteratively an approximation of the maximum likelihood estimator which maximizes the observed data loglikelihood:

$$\ell(\theta; X) = \log f_\theta(X) = \log \int p_\theta(X, y)\mu(\mathrm{d}y).$$

As this quantity cannot be computed explicitly in general cases, the EM algorithm finds the maximum likelihood estimator by iteratively maximizing the expected complete data loglikelihood. Start with an inital value $\theta^{(0)}$ and let $\theta^{(t)}$ be the estimate at the t-th iteration for $t \geqslant 0$, then the next iteration of EM is decomposed into two steps.

E step. Compute the expectation of the complete data loglikelihood, with respect to the conditional distribution of the missing data given the observed data parameterized by $\theta^{(t)}$:

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_{\theta^{(t)}}\left[\log p_\theta(X, Y) | X\right] .$$

M step Determine $\theta^{(t+1)}$ by maximizing the function Q:

$$\theta^{(t+1)} \in \mathrm{argmax}_\theta Q(\theta, \theta^{(t)})\,.$$

1. Prove the following crucial property, that motivates the EM algorithm. For all $\theta, \theta^{(t)}$,

$$\ell(\theta, X) - \ell(\theta^{(t)}, X) \geqslant Q(\theta, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)})\,.$$

Solution.

This may be proved by noting that

$$\ell(\theta, X) = \log\left(\frac{p_\theta(X, Y)}{p_\theta(Y|X)}\right)\,.$$

Considering the conditional expectation of both terms given X when the parameter value is $\theta^{(t)}$ yields

$$\ell(\theta, X) = Q(\theta, \theta^{(t)}) - \mathbb{E}_{\theta^{(t)}}\left[\log p_\theta(Y|X) | X\right].$$

Then,

$$\ell(\theta, X) - \ell(\theta^{(t)}, X) = Q(\theta, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)}) + H(\theta, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)})\,,$$

where

$$H(\theta, \theta^{(t)}) = -\mathbb{E}_{\theta^{(t)}}\left[\log p_\theta(Y|X) | X\right].$$

The proof is completed by noting that

$$H(\theta, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)}) \geqslant 0\,,$$

as this difference is a Kullback-Leibler divergence. □

Therefore, we straightforwardly have that the EM algorithm produces a non decreasing sequence of loglikelihoods $\left(\ell(X; \theta^{(t)})\right)_t$.

Mixture of Gaussians. In the following, $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ where $\{(X_i, Y_i)\}_{1 \leqslant i \leqslant n}$ are i.i.d. in $\mathbb{R}^d \times \{-1, 1\}$. For $k \in \{-1, 1\}$, write $\pi_k = \mathbb{P}(Y_1 = k)$. Assume that, conditionally on the event $\{Y_1 = k\}$, X_1 has a Gaussian distribution with mean $\mu_k \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. In this case, the parameter $\theta = (\pi_1, \mu_1, \mu_{-1}, \Sigma)$ belongs to the set $\Theta = [0, 1] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d \times d}$.

2. Write the complete data loglikelihood.
Solution.

The complete data loglikelihood is given by

$$\log p_\theta(X,Y) = -\frac{nd}{2}\log(2\pi) + \sum_{i=1}^n \sum_{k\in\{-1,1\}} \mathbb{1}_{\{Y_i=k\}}\left(\log \pi_k - \frac{\log \det\Sigma}{2} - \frac{1}{2}(X_i-\mu_k)^T\Sigma^{-1}(X_i-\mu_k)\right),$$

$$= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det\Sigma + \left(\sum_{i=1}^n \mathbb{1}_{\{Y_i=1\}}\right)\log \pi_1 + \left(\sum_{i=1}^n \mathbb{1}_{\{Y_i=-1\}}\right)\log(1-\pi_1)$$

$$-\frac{1}{2}\sum_{i=1}^n \mathbb{1}_{\{Y_i=1\}}(X_i-\mu_1)^T\Sigma^{-1}(X_i-\mu_1) - \frac{1}{2}\sum_{i=1}^n \mathbb{1}_{\{Y_i=-1\}}(X_i-\mu_{-1})^T\Sigma^{-1}(X_i-\mu_{-1}).$$

□

3. Let $\theta^{(t)}$ be the current parameter estimate. Compute $\theta \mapsto Q(\theta,\theta^{(t)})$ (tips: use $\omega_t^i = \mathbb{P}_{\theta^{(t)}}(Y_i=1|X_i)$)
Solution.

Write $\omega_t^i = \mathbb{P}_{\theta^{(t)}}(Y_i=1|X_i)$. The intermediate quantity of the EM algorithm is given by

$$Q(\theta,\theta^{(t)}) = -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det\Sigma + \left(\sum_{i=1}^n \omega_t^i\right)\log\pi_1 + \sum_{i=1}^n\left(1-\omega_t^i\right)\log(1-\pi_1)$$

$$-\frac{1}{2}\sum_{i=1}^n \omega_t^i(X_i-\mu_1)^T\Sigma^{-1}(X_i-\mu_1) - \frac{1}{2}\sum_{i=1}^n(1-\omega_t^i)(X_i-\mu_{-1})^T\Sigma^{-1}(X_i-\mu_{-1}).$$

□

4. Compute $\theta^{(t+1)}$.
Solution.

The gradient of $Q(\theta,\theta^{(t)})$ with respect to θ is therefore given by

$$\frac{\partial Q(\theta,\theta^{(t)})}{\partial \pi_1} = \frac{\sum_{i=1}^n \omega_t^i}{\pi_1} - \frac{n-\sum_{i=1}^n\omega_t^i}{1-\pi_1},$$

$$\nabla_{\mu_1}Q(\theta,\theta^{(t)}) = \sum_{i=1}^n \omega_t^i\left(2\Sigma^{-1}X_i - 2\Sigma^{-1}\mu_1\right),$$

$$\nabla_{\mu_{-1}}Q(\theta,\theta^{(t)}) = \sum_{i=1}^n(1-\omega_t^i)\left(2\Sigma^{-1}X_i - 2\Sigma^{-1}\mu_{-1}\right),$$

$$\nabla_{\Sigma^{-1}}Q(\theta,\theta^{(t)}) = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^n\omega_t^i(X_i-\mu_1)(X_i-\mu_1)^T - \frac{1}{2}\sum_{i=1}^n(1-\omega_t^i)(X_i-\mu_{-1})(X_i-\mu_{-1})^T.$$

Then, $\theta^{(t+1)}$ is defined as the only parameter such that all these equations are set to 0. It is given by

$$\widehat{\pi}_1^{(t+1)} = \frac{1}{n}\sum_{i=1}^n\omega_t^i,$$

$$\widehat{\mu}_1^{(t+1)} = \frac{1}{\sum_{i=1}^n\omega_t^i}\sum_{i=1}^n\omega_t^i X_i, \quad \widehat{\mu}_{-1}^{(t+1)} = \frac{1}{n-\sum_{i=1}^n\omega_t^i}\sum_{i=1}^n(1-\omega_t^i)X_i,$$

$$\widehat{\Sigma}^{(t+1)} = \frac{1}{n}\sum_{i=1}^n\omega_t^i\left(X_i-\widehat{\mu}_1^{(t+1)}\right)\left(X_i-\widehat{\mu}_1^{(t+1)}\right)^T + \frac{1}{n}\sum_{i=1}^n(1-\omega_t^i)\left(X_i-\widehat{\mu}_{-1}^{(t+1)}\right)\left(X_i-\widehat{\mu}_{-1}^{(t+1)}\right)^T.$$

□

Exercise 2 Let M_n^+ the space of real-valued $n\times n$ symmetric positive matrices. We show

1. Show that the function $X\mapsto \log\det X$ is concave on M_n^+.
Solution.

Let $X,Y\in M_n^+$ and $\lambda\in[0,1]$. Since $X^{-1/2}YX^{-1/2}\in M_n^+$, it is diagonalisable in some orthonormal basis and write $\mu_1,\ldots,\mu_n$ the (possibly repeated) entries of the diagonal. Note in particular that $\det\left(X^{-1/2}YX^{-1/2}\right) = \prod_{i=1}^n \mu_i$.

Then,

$$\begin{aligned}
\log\det\left((1-\lambda)X+\lambda Y\right) &= \log\det\left[X^{1/2}\left((1-\lambda)I+\lambda X^{-1/2}YX^{-1/2}\right)X^{1/2}\right] \\
&= \log\det X + \log\det\left((1-\lambda)I+\lambda X^{-1/2}YX^{-1/2}\right) \\
&= \log\det X + \sum_{i=1}^{n}\log(1-\lambda+\lambda\mu_i) \\
&\geqslant \log\det X + \sum_{i=1}^{n}(1-\lambda)\underbrace{\log(1)}_{=0}+\lambda\log(\mu_i) := D
\end{aligned}$$

where the last inequality follows from the concavity of the $\log$. Now, rewrite the rhs D as:

$$\begin{aligned}
D &= (1-\lambda)\log\det X + \lambda\left(\log\det X^{1/2} + \log\det X^{-1/2}YX^{-1/2} + \log\det X^{1/2}\right) \\
&= (1-\lambda)\log\det X + \lambda\log\det Y
\end{aligned}$$

□

2. The derivative of the real valued function $\Sigma \mapsto \log\det(\Sigma)$ defined on $\mathbb{R}^{d\times d}$ is given at a Σ which is symmetric positive by:

$$\partial_\Sigma\{\log\det(\Sigma)\} = \Sigma^{-1}\,,$$

where, for all real valued function f defined on $\mathbb{R}^{d\times d}$, $\partial_\Sigma f(\Sigma)$ denotes the $\mathbb{R}^{d\times d}$ matrix such that for all $1 \leqslant i,j \leqslant d$, $\{\partial_\Sigma f(\Sigma)\}_{i,j}$ is the partial derivative of f with respect to $\Sigma_{i,j}$.
Solution.

Recall that for all $i \in \{1,\ldots,d\}$ we have $\det(\Sigma) = \sum_{k=1}^d \Sigma_{i,k}\Delta_{i,k}$ where $\Delta_{i,j}$ is the (i,j)-cofactor associated to Σ. For any fixed i,j, the component $\Sigma_{i,j}$ does not appear in anywhere in the decomposition $\sum_{k=1}^d \Sigma_{i,k}\Delta_{i,k}$, except for the term $k=j$. This implies

$$\frac{\partial\log\det(\Sigma)}{\partial\Sigma_{i,j}} = \frac{1}{\det\Sigma}\frac{\partial\det(\Sigma)}{\partial\Sigma_{i,j}} = \frac{\Delta_{i,j}}{\det\Sigma}$$

Recalling the identity $\Sigma\,[\Delta_{j,i}]_{1\leqslant i,j\leqslant d} = (\det\Sigma)\,I_d$ so that $\Sigma^{-1} = \frac{[\Delta_{i,j}]^T_{1\leqslant i,j\leqslant d}}{\det\Sigma}$, we finally get

$$\left[\frac{\partial\log\det(\Sigma)}{\partial\Sigma_{i,j}}\right]_{1\leqslant i,j\leqslant d} = (\Sigma^{-1})^T = \Sigma^{-1}$$

where the last equality follows from the fact that Σ is symmetric. □

□