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 $\theta, \theta^{(t)}$,

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

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

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^+.
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 $\Sigma_{i,j}$.