HMWK2

Lemaire - Bouabid
Master M2 MVA 2018/2019
Probabilistic Graphical Models

November 7, 2018

Exercise 1.1
The implied factorization for any $p \in \mathcal{L}(G)$ is straightforward from the definition :

$$p(t, z, x, y) = p(t|z)p(z|x, y)p(x)p(y)$$

Let's show with a counter-example that $X \perp\!\!\!\perp Y|T$ doesn't hold for any $p \in \mathcal{L}(G)$. We set:

$$\begin{aligned} X &\sim \mathcal{B}(\pi) \\ Y &\sim \mathcal{B}(\pi) \text{ with } X \perp\!\!\!\perp Y \\ Z &= X \oplus Y \\ T &= Z \end{aligned}$$

We then have:

$$\begin{aligned} p(X, Y, Z, T) &\in \mathcal{L}(G) \\ p(X = 1|T = 1) &= \pi \\ p(Y = 1|T = 1) &= \pi \\ p(X = 1, Y = 1|T = 1) &= 0.5 \end{aligned}$$

which shows that $p(X = 1, Y = 1|T = 1) \neq p(X = 1|T = 1)p(Y = 1|T = 1)$ (as long as $\pi \neq \sqrt{0.5}$). Hence we don't always have $X \perp\!\!\!\perp Y|T$.

Exercise 1.2

(a) Assuming that Z is binary variable, we can without loss of generality suppose $Z(\Omega) = \{0, 1\}$. Let $\pi = \mathbb{P}(Z = 0)$, and $x, y \in X(\Omega) \times Y(\Omega)$

$$\begin{aligned} p(x, y) &= p(x, y|Z = 0)\pi + p(x, y|Z = 1)(1 - \pi) \\ &= p(x|Z = 0)p(y|Z = 0)\pi + p(x|Z = 1)p(y|Z = 1)(1 - \pi) \qquad (X \perp\!\!\!\perp Y \mid Z) \end{aligned}$$

Furthermore,

$$\begin{aligned} p(x, y) &= p(x)p(y) \qquad (X \perp\!\!\!\perp Y) \\ &= (p(x|Z = 0)\pi + p(x|Z = 1)(1 - \pi))\,(p(y|Z = 0)\pi + p(y|Z = 1)(1 - \pi)) \end{aligned}$$

Hence, if we note $p_0 = p(\cdot|Z = 0)$ and $p_1 = p(\cdot|Z = 1)$:

$$\begin{aligned} p_0(x)p_0(y)\pi + p_1(x)p_1(y)(1 - \pi) &= (p_0(x)\pi + p_1(x)(1 - \pi))\,(p_0(y)\pi + p_1(y)(1 - \pi)) \\ \Rightarrow p_0(x)p_0(y)\pi(1 - \pi) + p_1(x)p_1(y)\pi(1 - \pi) - (p_0(x)p_1(y) + p_1(x)p_0(y))\,\pi(1 - \pi) &= 0 \\ \Rightarrow p_0(x)(p_0(y) - p_1(y)) + p_1(x)(p_1(y) - p_0(y)) &= 0 \\ \Rightarrow (p_0(x) - p_1(x))(p_0(y) - p_1(y)) &= 0 \end{aligned}$$

Thus, $p(x|Z = 0) = p(x|Z = 1)$ or $p(y|Z = 0) = p(y|Z = 1)$.

We can conclude using the fact that if there exists an y_0 (or an x_0) such that $p(y_0|Z = 0) \neq p(y_0|Z = 1)$, we use what we have just shown for all $x \in X(\Omega)$ and y_0.

(b) We'll construct a counter-example in order to show that the result doesn't hold in the general case.
Let's consider $X, Y \sim \mathcal{B}(\pi)$ with $\pi \in (0, 1)$ such that $X \perp\!\!\!\perp Y$ and set $Z = 2X - Y$.
We have $ImZ = \{-1, 0, 1, 2\}$ and the function Φ defined by

$$\begin{aligned} \Phi\colon ImX \times ImY &\to ImZ \\ (x, y) &\mapsto 2x - y. \end{aligned}$$

is a bijection.
Hence for all f, g bounded measurable functions, we have:

$$\mathbb{E}[f(X)g(Y)|Z] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

since $X \perp\!\!\!\perp Y$ and $\sigma(Z) = \sigma(X, Y)$. This shows that $X \perp\!\!\!\perp Y|Z$.
Still,

$$\mathbb{P}(X = 0, Z = -1) = \mathbb{P}(Y = 1|X = 0)\mathbb{P}(X = 0) = \pi(1 - \pi)$$

$$\mathbb{P}(X = 0)\mathbb{P}(Z = -1) = \mathbb{P}(X = 0)\mathbb{P}(X = 0, Y = -1) = \pi^2(1 - \pi)$$

which shows that $X \not\perp\!\!\!\perp Z$. Similarly $Y \not\perp\!\!\!\perp Z$ which ends the proof.

Exercise 2.1

First let's show that G' is a DAG.

Suppose it's not i.e there exists a cycle that goes through node j (otherwise it would be clear that it is also a cycle for G). Let's denote $(j, b, a_1, ..., a_p, k, j)$ this cycle.

If $b = i$ then $(i, a_1, ..., a_p, k, i)$ is a cycle in G (using the definitions of i and j).

If $b \neq i$ then $(i, j, b, a_1, ..., a_p, k, i)$ is a cycle in G.

In both cases, we have a contradiction. Hence G' is a DAG.

Let $p \in \mathcal{L}(G)$. By definition:

$$p(x) = \prod_{k=1,...,n} p(x_k|x_{\pi_k}) = \prod_{k=1,...,n,k\neq i,j} p(x_k|x_{\pi_k}) * p(x_i|x_{\pi_i}) * p(x_j|x_{\pi_j})$$

If $k \neq i, j$ then clearly $\pi_k(G) = \pi_k(G')$.

But we also have:

$$\begin{aligned}
p(x_i|x_{\pi_i(G)}) * p(x_j|x_{\pi_j(G)}) &= \frac{p(x_i, x_{\pi_i(G)})}{p(x_{\pi_i(G)})} * \frac{p(x_j, x_{\pi_j(G)})}{p(x_{\pi_j(G)})} \\
&= \frac{p(x_i, x_{\pi_i(G)})}{p(x_{\pi_i(G)})} * \frac{p(x_j, x_{\pi_i(G)\cup\{i\}})}{p(x_{\pi_j(G)\cup\{i\}})} \\
&= \frac{p(x_i, x_j, x_{\pi_i(G)})}{p(x_{\pi_i(G)})} \\
&= p(x_i|x_{\pi_i(G')}) * p(x_j|x_{\pi_j(G')}) \qquad \text{expanding it in the similar way}
\end{aligned}$$

Thus $\mathcal{L}(G) \subset \mathcal{L}(G')$. We also have $\mathcal{L}(G') \subset \mathcal{L}(G)$ since the "covered" graph with respect to (i, j) of $\mathcal{G}'$ is $\mathcal{G}$. Hence $\mathcal{L}(G) = \mathcal{L}(G')$

Exercise 2.2 We let r be the root node of the tree and n be the number of nodes. A tree is a DAG in which all the nodes have exactly one parent except the root r that doesn't have any parent. Hence the only cliques of $\mathcal{G}'$ are the singletons and the 2-tuples {parent, child}.
Belonging to $\mathcal{L}(G)$ thus rewrites:

$$p(x) = p(x_r) \prod_{(i,j)\in E} p(x_j|x_i)$$

and belonging to $\mathcal{L}(G')$:

$$p(x) = \frac{1}{Z} \prod_{(i,j)\in E} \psi(x_i, x_j) \prod_{i\in V} \psi(x_i)$$

Let's first show that $\mathcal{L}(G) \subset \mathcal{L}(G')$. Let $p \in \mathcal{L}(G)$. Setting

$$\begin{aligned} \psi(x_r) &= p(x_r) \\ \psi(x_i) &= 1 \quad \forall i \neq r \\ \psi(x_i, x_j) &= p(x_j|x_i) \\ Z &= 1 \end{aligned}$$

we have

$$0 \leq \psi$$

$$p(x) = \frac{1}{Z} \prod_{(i,j)\in E} \psi(x_i, x_j) \prod_{i\in V} \psi(x_i)$$

$$\sum_x \prod_{(i,j)\in E} \psi(x_i, x_j) \prod_{i\in V} \psi(x_i) = \sum_x p(x) = 1 = \frac{1}{Z}$$

which means that $p \in \mathcal{L}(G')$. Hence $\mathcal{L}(G) \subset \mathcal{L}(G')$.

Let's now show that $\mathcal{L}(G') \subset \mathcal{L}(G)$. Let $p \in \mathcal{L}(G')$. We have for all $i, j \in V$ (with the notation $-i = \{i\}^c$)

$$p(x_i) = \sum_{x_{-i}} p(x_i, x_{-i})$$

$$p(x_i, x_j) = \sum_{x_{-(i,j)}} p(x_i, x_j, x_{-(i,j)})$$

Hence we can compute $f_i(x_i, x_j) = p(x_i|x_j) = \frac{p(x_i,x_j)}{p(x_j)}$ and we have:

$$0 \leq f_i \ , \forall i$$

$$\sum_{x_i} f_i(x_i, x_{p_i}) = \sum_{x_i} \frac{\sum_{x_{-(i,p_i)}} \prod_{(k,l)\in E} \psi(x_k, x_l) \prod_{k\in V} \psi(x_k)}{\sum_{x_{-p_i}} \prod_{(k,l)\in E} \psi(x_k, x_l) \prod_{k\in V} \psi(x_k)} = 1, \ \forall i, x_{p_i}$$

$$p(x_r) \prod_{(i,j)\in E} p(x_j|x_i) = 1$$

Thus $p \in \mathcal{L}(G)$.
We can conclude that $\mathcal{L}(G) = \mathcal{L}(G')$

Exercise 3.a As planned by the theory, different random starts leads to slightly different cluster centres and consequently slightly different distortions. The K-means++ initialization leads to similar results, although the distortion is a bit better in average for the same number of iterations. This sheds light on the fact that K-means++ enables to converge a bit faster, and avoid more often local minimas by initially spreading out the cluster centres.

Exercise 3.b (derivation)

Hypothesis : $\forall i \in [\![1,n]\!]\ Z_i \sim \mathcal{M}(k,\pi)$ and $\forall j \in [\![1,k]\!]\ X_i \mid \{Z_i = j\} \sim \mathcal{N}(\mu_j, \nu_j I_d)$ with $\mu_j \in \mathbb{R}^d$, $\nu_j \in \mathbb{R}$

Let θ our set of parameters, and $\tau_i^j = p_\theta(Z_i = j \mid X_i)$.

Under the isotropic covariance matrix hypothesis, the multivariate normal distribution becomes :

$$\mathcal{N}(x \mid \mu_j, \nu_j I_d) = (2\pi)^{\frac{-d}{2}} \nu_j^{\frac{-1}{2}} \exp\left(\frac{-1}{2}\frac{\|x_i - \mu_j\|^2}{\nu_j}\right) \Rightarrow \log \mathcal{N}(x \mid \mu_j, \nu_j I_d) = \frac{-d}{2}\log 2\pi - \frac{d}{2}\log \nu_j - \frac{1}{2}\frac{\|x_i - \mu_j\|^2}{\nu_j}$$

And we recall the E-step formula :

$$\mathbb{E}_{Z|X=x}[\ell_c(\theta)] = \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j \log \pi_j + \tau_i^j \log \mathcal{N}(x_i \mid \mu_j, \Sigma_j)$$

First, we can see this additional hypothesis doesn't affect the derivation step wrt π. It shouldn't affect the derivation wrt μ_j either as $\nabla_{\mu_j}\mathbb{E}_{Z|X=x}[\ell_c(\theta)] = -\sum_{i=1}^{n} \tau_i^j \frac{(x_i - \mu_j)}{\nu_j}$, hence ν_j is not involved in the nul-gradient equation.

But when it comes to ν_j, we get :

$$\forall j \in [\![1,k]\!],\ \nabla_{\nu_j}\mathbb{E}_{Z|X=x}[\ell_c(\theta)] = 0 \Rightarrow \sum_{i=1}^{n} \frac{-\tau_i^j}{2\nu_j} + \frac{\tau_i^j}{2}\frac{\|x_i - u_j\|^2}{\nu_j^2} = 0 \Rightarrow \boxed{\hat{\nu_j} = \frac{\sum_i \tau_{ij}\|x_i - \hat{\mu}_j\|^2}{d\sum_i \tau_{ij}}}$$

Mass coverage ellipses :

Let $X \sim \mathcal{N}(\mu, \Sigma)$. As $\Sigma \in \mathcal{S}_d^{++}(\mathbb{R})$, so is Σ^{-1} and it hence admits a square root and induces a norm $\|\cdot\|_{\Sigma^{-1}}$. The problem is to find the radius of the circle centered on μ with $(1-\eta)\%$ coverage of the pdf wrt the latter norm.

But

$$\|X - \mu\|_{\Sigma^{-1}}^2 \leq R^2 \Leftrightarrow \| \underbrace{\sqrt{\Sigma^{-1}}(X - \mu)}_{\sim \mathcal{N}(0, I_d)} \|_2^2 \leq R^2$$

$\|\cdot\|_2$ being the euclidean norm, this is basically a sum of univariate centered gaussian rv with variance 1 and therefore follows a χ^2-distribution with d degrees of freedom. We can hence set R^2 to be the quantile of level $(1-\eta)$ of such distribution.

Exercise 3.c (estimator covariance)

Covariance estimator is given by : $\hat{\Sigma}_j = \dfrac{\sum_i \tau_{ij}(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T}{\sum_i \tau_{ij}}$

Exercise 3.d

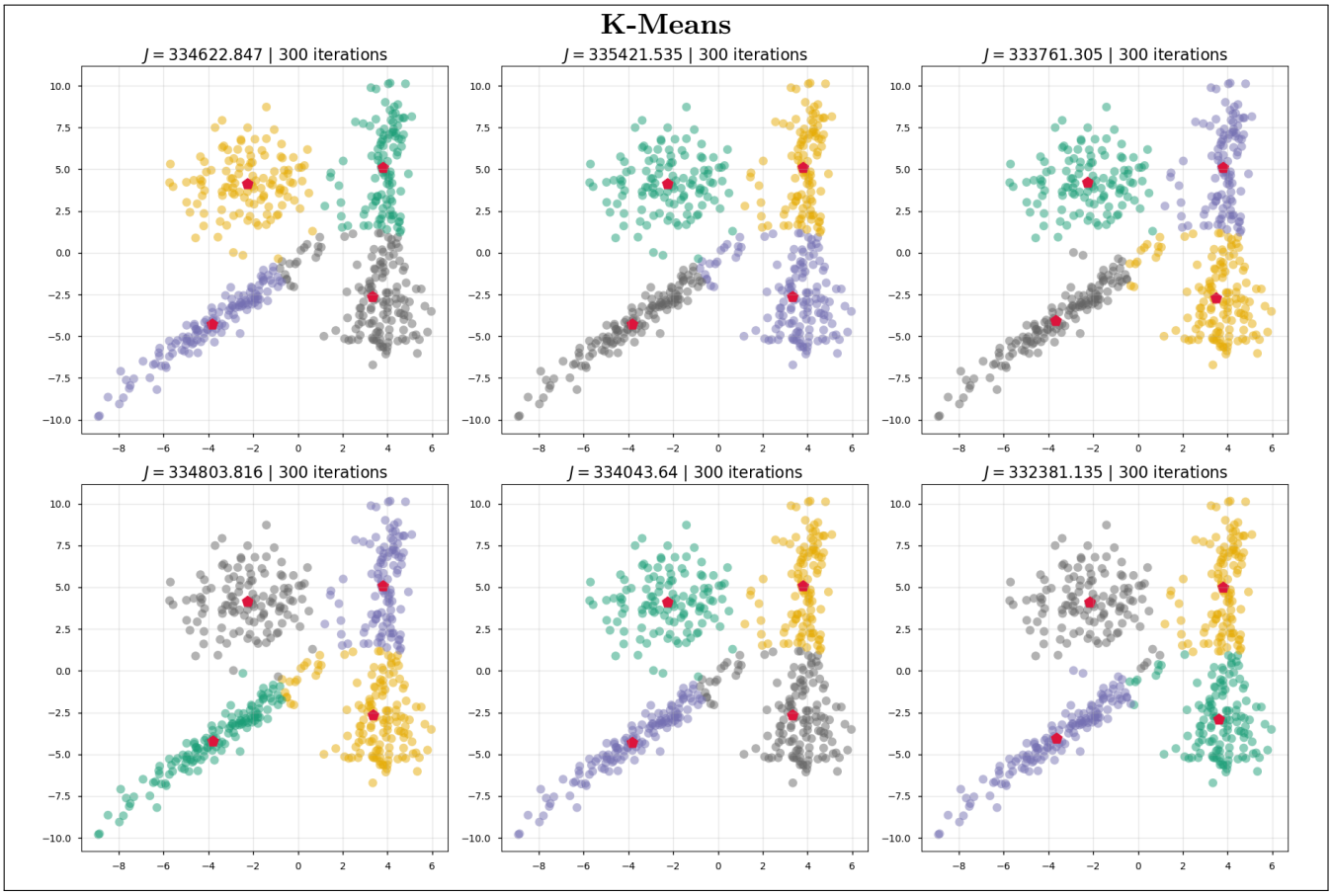
Normalized log-likelihoods

	test	train
General	-6.941	-4.810
Isotropic	-6.979	-5.439

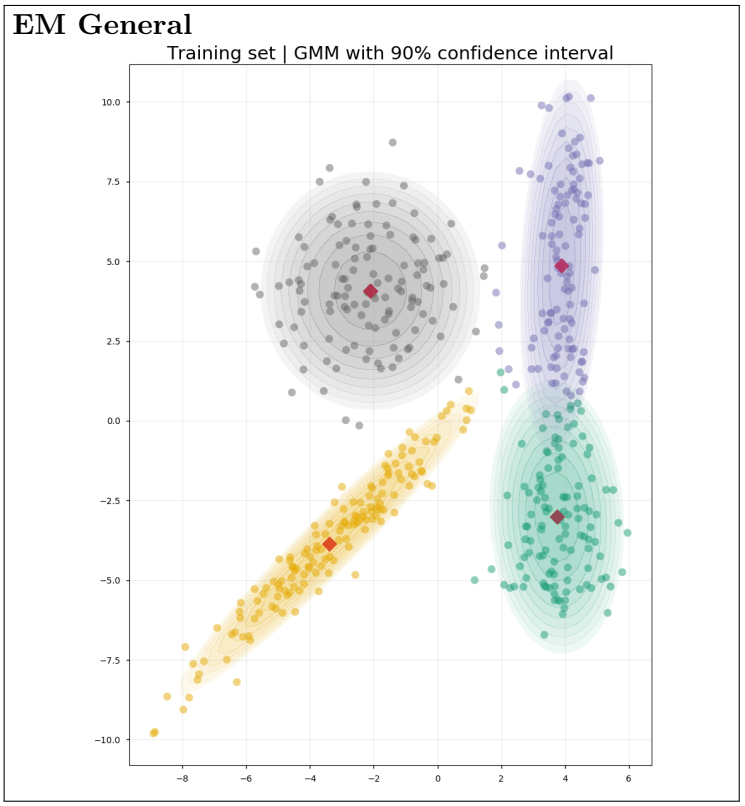
Since the datas used are indeed extracted from a gaussian mixture, the EM algorith clearly captures the gaussian distributions better than a simple K-mean.

Regarding the difference between the isotropic and non-isotropic EM, we also notice on the figures that the non-isotropic fits better the original distribution, which makes sense since the true covariance matrices are indeed not equal.

The better the model, the higher the likelihood, what we can see below. We notice that the likelihood of the testing set is smaller than the one of the training set, which makes sense since the model fits better on the data set it has been trained on.


K-Means
J = 334622.847 | 300 iterations
J = 335421.535 | 300 iterations
J = 333761.305 | 300 iterations
J = 334803.816 | 300 iterations
J = 334043.64 | 300 iterations
J = 332381.135 | 300 iterations


EM isotropic
Training set | Isotropic MM with 90% confidence interval


EM General
Training set | GMM with 90% confidence interval