Probabilistic Graphical Models
Homework 2

Jean-Christophe Corvisier
Mohammed Amine Kheldouni

6 November 2018

Exercice 1

Question 1

Let $p \in \mathcal{L}(G)$, the joint probability p is written down for all x : $p(x,y,z,t) = p(x)p(y)p(z|x,y)p(t|z)$ (noted (1)). The statement $X \perp\!\!\!\perp Y|T$ is false. For instance, if we take X, Y two binairies r.v., Z the binary variable such that $Z = 1$ if $X = Y$ and 0 otherwise, and set $T = Z$. The joint probability in (1) is well defined with those r.v., and we have :

$$p(x,y|t) = \frac{p(x,y,t)}{p(t)} = \frac{p(x)p(y)p(z=0|x,y)p(t|z=0) + p(x)p(y)p(z=1|x,y)p(t|z=1)}{p(t)}$$

if we take $x \neq y$, we then have

$$p(x,y|t) = \frac{p(x,y,t)}{p(t)} = \frac{p(x)p(y)p(z=0|x,y)p(t|z=0)}{p(t)}$$

And taking $t = 1$, we have $p(t=1|z=0) = 0$, Hence

$$p(x,y|t) = 0 \neq p(x,y)$$

Finally, we conclude that X and Y are not independent given Z.

Question 2

1) A concise scheme is elaborated in the following to prove our thoughts. Please refer to the appendix for more detailed arguments. The statement is true. Indeed, assuming $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y|Z$, we have by independence of X and Y :

$$\forall (x,y), \quad P(x,y) = P(x)P(y)$$

Then we have the following formula (denoted (1))

$$P(x)P(y) = P(x,y|z=0)P(z=0) + P(x,y|z=1)P(z=1)$$

We suppose that $P(z=0) > 0$ (the result is trivial if not). Then after some calculation (see appendix) we have :

$$(P(y|z=0) - P(y))(P(x|z=0) - P(x)) = 0$$

and this equality is true for all x, y. As Z is binary, this is a sufficient condition to give that $X \perp\!\!\!\perp Z$ or that $Y \perp\!\!\!\perp Z$.

2) The statement is false. Consider X and Y two r.v. independent, and set $Z = (X,Y)$. We obviously don't have $X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$.

Indeed, since $P(x,y|z) = \frac{P(x,y,x',y')}{P(x',y')} = \frac{P(x,x')P(y,y')}{P(x')P(y')}$ (because $X \perp\!\!\!\perp Y$), we immediately have : $P(x,x') = P(x|x')P(x') = P(x|x',y')P(x')$ and $P(y,y') = P(y|y')P(y') = P(y|y',x')P(y')$.

Finally, we have $P(x,y|z) = P(x|x',y')P(y|x',y') = P(x|z)P(y|z)$, which gives us the conclusion.

Exercise 2

Question 1

By symmetry of the problem on i and j, we can just show the first inclusion, which will complete the proof.

Let p a probability distribution in $\mathcal{L}(G)$, we then have $P(x_1,...,x_n) = \prod_{k=1}^{n} p(x_k|x_{\pi_k})$ and since we have $\pi_j = \pi_i \cup \{i\}$, we conclude that

$$P(x_1,...,x_n) = \prod_{k\neq(i,j)} p(x_k|x_{\pi_k})p(x_i|x_{\pi_i})p(x_j|x_{\pi_j})$$

.
Consequently, $p(x_i|x_{\pi_i})p(x_j|x_{\pi_j}) = p(x_i|x_{\pi_i})p(x_j|x_{\pi_i},x_i) = p(x_i,x_j|x_{\pi_i})$ (1)
But, we also have $p(x_i,x_j|x_{\pi_i}) = p(x_i|x_{\pi_i},x_j)p(x_j|x_{\pi_i})$ (noted (2)) As in the graph $G^{'}$, we have $x_{\pi_i^{'}} = x_{\pi_i} \cup \{j\}$ and $x_{\pi_j^{'}} = x_{\pi_i}$, the previous equality shows that we can have the relation $P(x_1,...,x_n) = \prod_{k=1}^{n} p(x_k|x_{\pi_k^{'}})$ in $G^{'}$. Therefore, p is in $\mathcal{L}(G^{'})$.

Question 2

First, let's show that $\mathcal{L}(G) \subset \mathcal{L}(G^{'})$ Let $p \in \mathcal{L}(G)$ For any x, we have : $p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i})$. But as G is a directed tree, $x_{\pi i}$ is a single node of the tree for x_i, except for the root of the tree. Given all that, the cliques of G are the nodes themselves and the pairs x_i, x_{i}.
We then write the following functions : $\psi_{i,\pi_i}(x_i,x_{\pi_i}) = p(x_i|x_{\pi_i})$ and $\psi_i(x_i) = 1$ if v_i is not the root, and $\psi_i(x_i) = p(x_i)$ if v_i is the root. We then have : $\prod_{i=1}^{n} p(x_i|x_{\pi_i}) = \prod_{c\in(C)} \psi_C(x_C)$ with $\mathcal{C} = (x_i,x_{\pi_i}) \bigcup \{x_i\}$.

Finally, with $Z = \sum_x \prod_{c\in(C)} \psi_C(x_C) = \sum_x \prod_{i=1}^{n} p(x_i|x_{\pi_i}) = 1$ we have : $p(x) = \frac{\prod_{c\in(C)} \psi_C(x_C)}{Z}$ which gives us that $p \in \mathcal{L}(G^{'})$,and so that $\mathcal{L}(G) \subset (L)(G^{'})$.

We then want to show the second inclusion : $\mathcal{L}(G^{'}) \subset \mathcal{L}(G)$. We will proceed by induction on the number of the nodes V of the undirect tree $G^{'}$, noted n. For $n = 1$, the result is trivial. Assuming that the result is true at n, let's consider an undirected tree $G^{'}$, with $n+1$ nodes. Let's write the nodes in an order such that the node $n+1$ is a leaf connected to the node n. Let $p \in \mathcal{L}(G^{'})$. We have $\forall x, \quad p(x) = \frac{1}{Z} \prod_{c\in C} \psi_c(x_c)$ In the undirected graph, the cliques are the nodes and the edges. We can then separate the product and write : $p(x) = \frac{1}{Z} \prod_{c\in C, c\neq v_n+1,(v_n,v_n+1)} \psi_c(x_c)\psi_{v_{n+1}}(x_{n+1})\psi_{v_n,v_{n+1}}(x_{n+1},x_n)$.

Let's note $G^{''}$ the graph created by removing the leaf v_{n+1} and the edge $(n,n+1)$ of $G^{'}$. $G^{''}$ is still an undirected tree (because $n+1 \geq 2$). The probability $p^{'}$ given by :

$$p^{'}(x_1,...x_n) = \sum_{x_{n+1}} \psi_{v_{n+1}}(x_{n+1})\psi_{(n,n+1)}(x_{n+1},x_n)\frac{1}{Z} \prod_{c\in C, c\neq v_n+1,(v_n,v_n+1)} \psi_c(x_c)$$

is a probability of $\mathcal{L}(G^{''})$. We can then use the hypothetis of induction, and we write $p'(x_1,...x_n) = \prod_{i=1}^{n} f_i(x_i,x_{\pi_j})$
Now, if we note : $\phi(x_n) = \sum_{x_{n+1}} \psi_{v_{n+1}}(x_{n+1})\psi_{(n,n+1)}(x_{n+1},x_n)$ and if we note (because v_{n+1} is connected to v_n) :

$$f_{n+1}(x_{n+1},x_{\pi_{n+1}}) = \frac{\psi_{v_{n+1}(x_{n+1}}\psi_{(n,n+1)}(x_{n+1},x_n)}{\phi(x_n)}$$

if $\phi(x_n) \neq 0$ or $f_{n+1}(x_{n+1},x_{\pi_{n+1}}) = \frac{1}{C_{n+1}}$ where C_{n+1} is the cardinal of the set of values taken by X_{n+1}.
We then clearly have : $\forall x \quad p(x) = \prod_{i=1}^{n+1} f_i(x_i,x_{\pi_i})$ And as $\sum_{x_{n+1}} f_{n+1}(x_{n+1}),x_{\pi_{n+1})} = 1$, we then conclude that $p \in \mathcal{L}(G)$, and so that $\mathcal{L}(G^{'}) \subset \mathcal{L}(G)$, which conclude the whole induction.

Exercise 3

a) K-means clustering

Implementing the K-means algorithm for the given dataset, we notice that the training data can be successfully clustered into $K = 4$ clusters by K-means, as the distorsion measure keeps decreasing until converging to stable centroids μ (cf. FIGURE 1). Iterating several random initializations, we obtain different results but most of them minimize the distortion measure as we can see in the distortion measure histogram (cf. FIGURE 2 and 3).

b) EM Algorithm derivations for isotropic Gaussian mixtures

The Maximization step for our algorithm finds the estimators converging towards optimal parameter values by maximizing the expected complete log-likelihood as seen in the corresponding lecture :

$$\begin{aligned}\mathbb{E}_{q^{(t)}}[\widetilde{l}(\theta)] &= \sum_{i,k} q_{ik}^{(t)} log\mathcal{N}(x;\mu_k,\sigma_k^2) + \sum_{i,k} q_{ik}^{(t)} log(\alpha_k) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} q_{ik}^{(t)} log\left(\frac{1}{(2\pi)^{d/2}\sqrt{|\sigma_k^2\mathcal{I}_d|}} exp\left(-\frac{1}{2}\frac{1}{\sigma_k^2}\|x_i-\mu_k\|^2\right)\right) + \sum_{i=1}^{n}\sum_{k=1}^{K} q_{ik}^{(t)} log(\alpha_k)\end{aligned}$$

This yields the following equation for computing the estimator σ_k^2 in the isotropic case (using $|\Sigma_k| = det(\Sigma_k) = \sigma^{2d}$) :

$$\frac{\partial \mathbb{E}_{q^{(t)}}[\widetilde{l}(\theta)]}{\partial \sigma_k^2} = \sum_{i=1}^{n} q_{ik}^{(t)}\left[-\frac{d}{2\sigma^2} + \frac{1}{2}\frac{1}{\sigma_k^4}\|x_i-\mu_k\|^2\right] = 0$$

Therefore, the estimators for our isotropic covariance matrices are :

$$\forall k \in \{0,...,K\}, \qquad (\sigma_k^2)^{(t)} = \frac{\sum_{i=1}^{n} q_{ik}^{(t)}\|x_i-\mu_k^{(t)}\|^2}{d\sum_{i=1}^{n} q_{ik}^{(t)}}$$

c) EM Algorithm estimators for general Gaussian mixtures

Considering the general GMM, the estimator for the K covariance matrix Σ_k are computed by the following formula :

$$\forall k \in \{0,...,K\}, \qquad \Sigma_k^{(t)} = \frac{\sum_{i=1}^{n} q_{ik}^{(t)}(x_i-\mu_k^{(t)})(x_i-\mu_k^{(t)})^T}{\sum_{i=1}^{n} q_{ik}^{(t)}}$$

d) Results and comments

First and foremost, we managed to cluster successfully the datasets into $K = 4$ different clusters which parameters have been computed with the EM algorithm. We notice that the isotropic model keeps the gaussian ellipses horizontal because of the isotropy introduced. Therefore, it doesn't fit the data as well as the general case which allows more flexibility and variability to the model parameters Σ. Indeed, the clusters 1 and 2 in the isotropic training clustering shows the underperformance of the imposed isotropy. This is highlighted by the likelihood tabular, which infers that the likelihood is better for the general case. We also notice a better normalized likelihood for the training set which is a predictable behavior since test data are only used for validating the calibrated parameters whereas the training data have contributed to the computation.

	Train dataset	Test dataset
General GMM EM	-4.74	-4.91
Isotropic GMM EM	-5.45	-5.49

TABLE 1 – Normalized loglikelihood for both GMM models on training and testing sets

Figures for training and test data

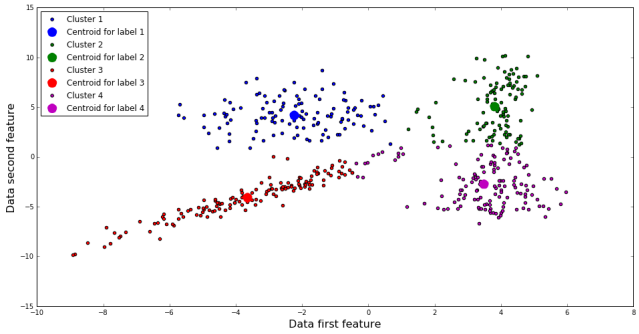

FIGURE 1 – Final clusters built with K-means

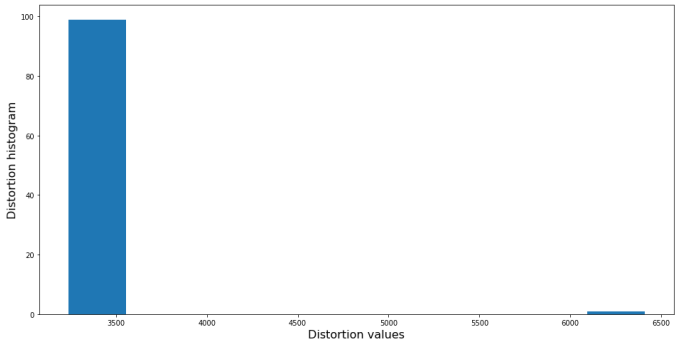

FIGURE 2 – Histogram of final distorsion for multiple K-means iterations

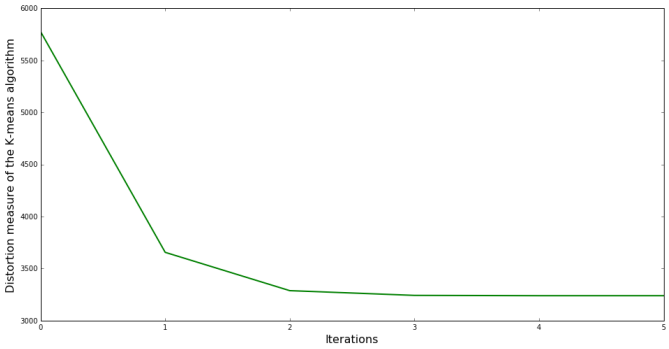

FIGURE 3 – Figure showing the convergence of the distortion using K-means

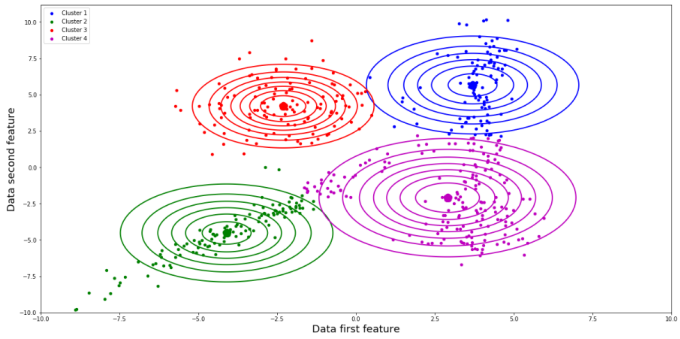

FIGURE 4 – Clusters in training set for an isotropic GMM

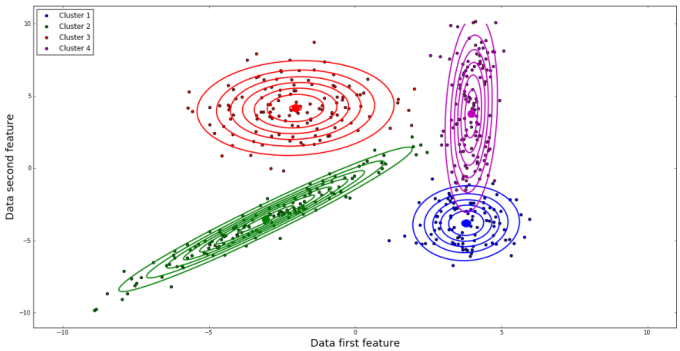

FIGURE 6 – Clusters in traning set for a general GMM

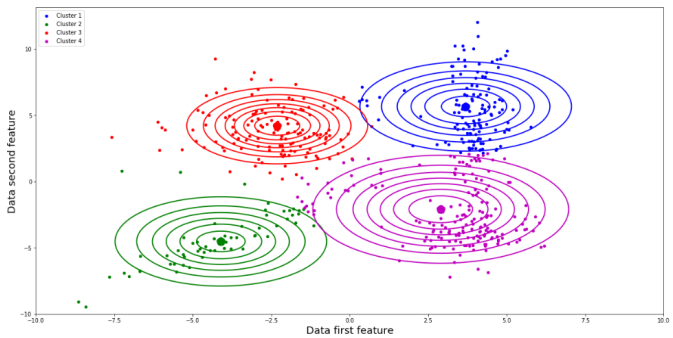

FIGURE 5 – Clusters in test set for an isotropic GMM

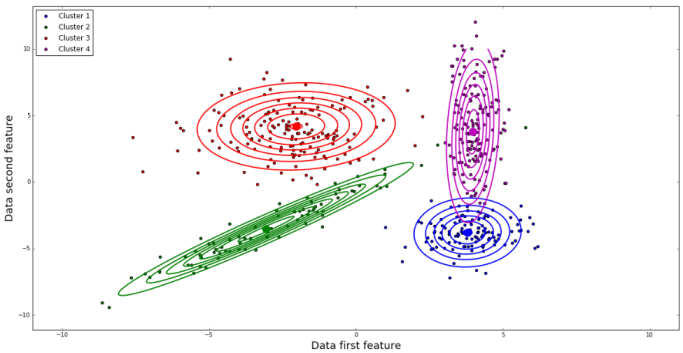

FIGURE 7 – Clusters in test set for a general GMM

Appendix

Exerice 1- Question 2

We will prove here the equality :

$$\forall (x,y) \qquad (P(y|z=0) - P(y))(P(x|z=0) - P(x)) = 0$$

We note here $p = P(z=0)$, $p > 0$ because if it is not the case, the result is trivial.
With the hypothesis $X \perp\!\!\!\perp Y$ and $(X,Y) \perp\!\!\!\perp Z$, we have :

$$\forall (x,y), \quad P(x,y) = P(x)P(y) = pP(x|z=0)P(y|z=0) + (1-p)P(x|z=1)P(y|z=1)$$

. We note the following equation (1)

$$P(x)P(y) = pP(x|z=0)P(y|z=0) + (1-p)P(x|z=1)P(y|z=1)$$

But we also have : $\forall x, \;\; P(x) = P(x|z=0)p + P(x|z=1)(1-p)$ and $\forall y, \;\; P(y) = P(y|z=0)p + P(y|z=1)(1-p)$ we then obtain , that : $\forall x, \;\; P(x|z=1) = \frac{P(x) - pP(x|z=0)}{1-p}$ and the same hold for Y, for all y. Using those results in (1), we have that :

$$P(x)P(y) = pP(x|z=0)P(y|z=0) + (1-p)(P(x) - pP(x|z=0)(P(y) - pP(y|z=0)\frac{1}{(1-p)^2}$$

Pursuing the calculus, we have that :

$$(1-p)P(x)P(y) = p(1-p)P(x|z=0)P(y|z=0) + (P(x) - pP(x|z=0))(P(y) - pP(y|z=0))$$

Then

$$P(x)[(1-p)P(y) - P(y) + pP(y|z=0)] = P(x|z=0)[p(1-p)P(y|z=0) - pP(y) + p^2P(y|z=0)$$

Hence

$$P(x)(P(y|z=0) - P(y)) = P(x|z=0)(P(y|z=0) - P(y))$$

and finally

$$(P(y|z=0) - P(y))(P(x|z=0) - P(x)) = 0$$

This equality gives us that $P(x|z=0) = P(x)$ for any x or that $P(y|z=0) = P(y)$ for any y. But as we have the relation $P(x|z=1) = \frac{P(x) - pP(x|z=0)}{1-p}$ and $P(x|z=1) = \frac{P(y) - pP(y|z=0)}{1-p}$, we directly conclude that $P(x|z=1) = P(x)$ for any x or that $P(y|z=1) = P(y)$ for any y. Hence we can conclude our proof.

Exercise 3 - EM derivations

Consider an isotropic Gaussian mixtures model (X, z) such that

$$p(z) = \prod_{k=1}^{K} \alpha_k^{z_k} \qquad p(x|z;(\mu_k,\Sigma_k)_k) = \sum_{k=1}^{K} z_k \mathcal{N}(x;\mu_k,\Sigma_k)$$

where $z = (z_1, ..., z_K)^T \in \{0,1\}^K$ follows a multinomial model and each covariance matrix of the K clusters is isotropic $\Sigma_k = \sigma_k^2 \mathcal{I}_d$ ($\mathcal{I}_d$ being the $\mathbb{R}^{dxd}$ identity matrix).

The Expectation step of our algorithm computes $q_i^{(t)}(z^{(i)})$ (with the course's notations), which is a multinomial distribution defined by

$$q_i^{(t)} = p(z^{(i)}|x^{(i)};\theta^{(t-1)})$$

where $\theta^{(t)} = \{\alpha^{(t)}, \mu^{(t)}, (\sigma^2)^{(t)}\}$ is our set of parameters. The Maximization step for our algorithm finds the estimators converging towards optimal parameter values by maximizing the expected complete log-likelihood as seen in the corresponding lecture :

$$\mathbb{E}_{q^{(t)}}[\widetilde{l}(\theta)] \quad = \quad \sum_{i,k} log\mathcal{N}(x;\mu_k,\sigma_k^2) + \sum_{i,k} q_{ik}^{(t)} log(\alpha_k)$$

This yields by derivation of the likelihood the following updates for each estimator at the t^{th} iteration :

$$\mu_k^{(t)} = \frac{\sum_{i=1}^{n} x_i q_{ik}^{(t)}}{\sum_{i=1}^{n} q_{ik}^{(t)}} \qquad \alpha_k^{(t)} = \frac{\sum_{i=1}^{n} q_{ik}^{(t)}}{\sum_{i=1}^{n}\sum_{k'=1}^{K} q_{ik'}^{(t)}}$$

By assigning the likelihood's derivative with respect to $sigma_k^2$ to 0, we find the last estimator in the isotropic case (using $|\Sigma_k| = det(\Sigma_k) = \sigma^{2d}$) :

$$\frac{\partial \mathbb{E}_{q^{(t)}}[\widetilde{l}(\theta)]}{\partial \sigma_k^2} = \sum_{i=1}^{n} q_{ik}^{(t)} \left[-\frac{d}{2\sigma^2} + \frac{1}{2}\frac{1}{\sigma_k^4}\|x_i - \mu_k\|^2 \right] = 0$$

Therefore, the estimators for our isotropic covariance matrices are :

$$\forall k \in \{0, ..., K\}, \qquad (\sigma_k^2)^{(t)} = \frac{\sum_{i=1}^{n} q_{ik}^{(t)} \|x_i - \mu_k^{(t)}\|^2}{d \sum_{i=1}^{n} q_{ik}^{(t)}}$$

In the general case, the GMM covariance matrices' estimator is computed at each iteration t by the formula :

$$\forall k \in \{0, ..., K\}, \qquad \Sigma_k^{(t)} = \frac{\sum_{i=1}^{n} q_{ik}^{(t)} (x_i - \mu_k^{(t)})(x_i - \mu_k^{(t)})^T}{\sum_{i=1}^{n} q_{ik}^{(t)}}$$