Graphical models homework 2

Léo Alberge, Romain Petit

January 9, 2019

Exercise 1.1

$$\boxed{\text{For any } p \in \mathcal{L}(G),\ p(x,y,z,t) = p(x)\,p(y)\,p(z|x,y)\,p(t|z)}$$

Let X and Y be two independant random variables taking values in $\{0,1\}$. Let $Z = XY$ and $T = Z$. Let us show that p, the joint distribution of (X,Y,Z,T), is in $\mathcal{L}(G)$. We have that $p(x,y,z,t) = p(t\,|\,x,y,z)\,p(x,y,z)$. Since $X \perp Y$, $p(x,y,z) = p(z\,|\,x,y)\,p(x,y) = p(z\,|\,x,y)\,p(x)\,p(y)$. Since $T = Z$, $p(t\,|\,x,y,z) = \delta_{z,t} = p(t\,|\,z)$. We therefore have $p(x,y,z,t) = p(x)\,p(y)\,p(z\,|\,x,y)\,p(t\,|\,z)$ which means $p \in \mathcal{L}(G)$.

$\mathbb{P}(X=0\,|\,Y=1,\,T=0) = \mathbb{P}(X=0\,|\,Y=1,\,XY=0) = 1$
$\mathbb{P}(X=0\,|\,T=0) = \mathbb{P}(X=0\,|\,XY=0) \neq 1$, which shows that :

$$\boxed{\text{We do not have } X \perp Y\,|\,T \text{ for all } p \in \mathcal{L}(G)}$$

Exercise 1.2

(a) Since $X \perp Y$, $p(x,y) = p(x)p(y)$. We therefore have :

$$p(x,y) = \big[p(z=0)\,p(x\,|\,z=0) + p(z=1)\,p(x\,|\,z=1)\big]\big[p(z=0)\,p(y\,|\,z=0) + p(z=1)\,p(y\,|\,z=1)\big] \quad (1)$$

Since $X \perp Y\,|\,Z$, we also have :

$$p(x,y) = p(z=0)\,p(x,y\,|\,z=0) + p(z=1)\,p(x,y\,|\,z=1) \quad (2)$$
$$p(x,y) = p(z=0)\,p(x\,|\,z=0)\,p(y\,|\,z=0) + p(z=1)\,p(x\,|\,z=1)\,p(y\,|\,z=1) \quad (3)$$

Developping (1) gives :

$$\begin{aligned} p(x,y) = p(z=0)^2\,p(x\,|\,z=0)\,p(y\,|\,z=0) + p(z=1)^2\,p(x\,|\,z=1)\,p(y\,|\,z=1) \\ +p(z=0)\,p(z=1)\Big[p(x\,|\,z=0)\,p(y\,|\,z=1) + p(x\,|\,z=1)\,p(y\,|\,z=0)\Big] \end{aligned} \quad (4)$$

Combining (3) and (4) gives :

$$p(z=0)\,p(z=1)\big[p(y\,|\,z=1) - p(y\,|\,z=0)\big]\big[p(x\,|\,z=0) - p(x\,|\,z=1)\big] = 0 \quad (5)$$

If Z is not almost surely constant, we therefore have :

$p(x\,|\,z=0) = p(x\,|\,z=1) = p(x)$ or $p(y\,|\,z=0) = p(y\,|\,z=1) = p(y)$, which means $\boxed{X \perp Z \text{ or } Y \perp Z}$

(b) $\boxed{\text{In the general case, we do not have that } X \perp Y\,|\,Z \text{ and } X \perp Y \text{ implies } X \perp Z \text{ or } Y \perp Z}$
The proof of this result is given in the appendix.

Exercise 2.1

Let $G = (V, E)$ be a DAG, and $i \to j \in E$ a covered edge.

Let us first show that $G' = (V, E')$ with $E' = E \setminus \{i \to j\} \cup \{j \to i\}$ is a DAG. If there exists a cycle in G', since G is a DAG, it necessarily includes $j \to i$ (otherwise it would be a cycle in G). There therefore exists an elementary cycle (thanks to König's lemma) of the form $(a_1, ..., a_k, j, i, ..., a_1)$ with $a_k \in \pi'_j$, where π'_j denotes the set of j's parents in G'. Since $i \to j$ is a covered edge in G, we have that $\pi'_i = \pi'_j \cup \{j\}$. We therefore have that $a_k \in \pi'_i$, and that $(a_1, ..., a_k, i, .., a_1)$ is a cycle in G' that does not include $j \to i$, which means it is a cycle in G. Since G is a DAG, this is impossible and there does not exist any cycle in G'.

$$\boxed{G' \text{ is therefore a DAG and } \mathcal{L}(G') \text{ is well defined.}}$$

We will now show the following preliminary result, that holds for any distribution p over a product set $\mathcal{X}_1, ..., \mathcal{X}_n$ ($n = |G| = |G'|$), not necessarily factorizing in G or G' :

$$\boxed{p(x_i \,|\, x_{\pi_i})\, p(x_j \,|\, x_{\pi_j}) = p(x_i \,|\, x_{\pi'_i})\, p(x_j \,|\, x_{\pi'_j})}$$

Since $\pi_j = \pi_i \cup \{i\}$, $p(x_i \,|\, x_{\pi_i})\, p(x_j \,|\, x_{\pi_j}) = p(x_i \,|\, x_{\pi_i})\, p(x_j \,|\, x_{\pi_i}, x_i) = \frac{p(x_i, x_{\pi_i})}{p(x_{\pi_i})} \frac{p(x_i, x_j, x_{\pi_i})}{p(x_i, x_{\pi_i})} = p(x_i, x_j \,|\, x_{\pi_i})$
Since $\pi_i = \pi'_j$, $p(x_i, x_j \,|\, x_{\pi_i}) = p(x_i, x_j \,|\, x_{\pi'_j}) = \frac{p(x_i, x_j, x_{\pi'_j})}{p(x_{\pi'_j})} = \frac{p(x_i, x_j, x_{\pi'_j})}{p(x_j, x_{\pi'_j})} \frac{p(x_j, x_{\pi'_j})}{p(x_{\pi'_j})} = p(x_i \,|\, x_j, x_{\pi'_j})\, p(x_j \,|\, x_{\pi'_j})$
We also have $\pi'_i = \pi'_j \cup \{j\}$, which shows $p(x_i \,|\, x_{\pi_i})\, p(x_j \,|\, x_{\pi_j}) = p(x_i, x_j \,|\, x_{\pi_i}) = p(x_i \,|\, x_{\pi'_i})\, p(x_j \,|\, x_{\pi'_j})$

Since $\pi_k = \pi'_k$ for all $k \neq i, j$, we finally have $\prod_{i=1}^{n} p(x_i, x_{\pi_i}) = \prod_{i=1}^{n} p(x_i, x_{\pi'_i})$ and therefore that $\boxed{\mathcal{L}(G) = \mathcal{L}(G')}$

Exercise 2.2

Let us show this result by induction. If G has a single node, then $\mathcal{L}(G) = \mathcal{L}(G')$ is the set of probability distributions over a single variable.

Let us assume that, for a certain $n \geq 2$, we have the property for all directed trees with $n-1$ nodes. Let G be a directed tree with n nodes and G' its corresponding undirected tree. Let us consider a leaf of G. We will denote X_n the random variable associated to that leaf. Let $p \in \mathcal{L}(G)$. We have that $p(x_{V\setminus\{n\}})$ factorizes in the induced graph on $V\setminus\{n\}$, which is of size $n-1$. The induction assumption gives that it also factorizes in its corresponding undirected tree, and therefore that there exist non negative functions $(f_i)_{i=1,...,n}$ and $(\psi_i)_{i=1,...,n-1}$ such that :

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} f_i(x_i, x_{\pi_i}) = f_n(x_n, x_{\pi_n}) \prod_{i=1}^{n-1} f_i(x_i, x_{\pi_i}) = f_n(x_n, x_{\pi_n}) \frac{\prod_{C \in \mathcal{C}_{n-1}} \psi_C(x_C)}{\sum_{x'_1, ..., x'_{n-1}} \prod_{C \in \mathcal{C}_{n-1}} \psi_C(x'_C)}$$

where $\mathcal{C}_{n-1}$ is the set of maximal cliques of the undirected tree coresponding to the graph induced by $V\setminus\{n\}$. Since G is a tree, we have that $\mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n, \pi_n\}\}$ where π_n denotes the unique parent of n. If we set $\psi_{\{n,\pi_n\}}(x_n, x_{\pi_n}) = \frac{f_n(x_n, x_{\pi_n})}{\sum_{x'_1, ..., x'_{n-1}} \prod_{C \in \mathcal{C}_{n-1}} \psi_C(x'_C)}$, since $\sum_{x_1, ..., x_n} \prod_{C \in \mathcal{C}_n} \psi_C(x_C) = \sum_{x_1, ..., x_n} p(x_1, ..., x_n) = 1$, we have $p(x_1, ..., x_n) = \frac{\prod_{C \in \mathcal{C}_n} \psi_C(x_C)}{\sum_{x'_1, ..., x'_n} \prod_{C \in \mathcal{C}_n} \psi_C(x'_C)}$, which shows $\mathcal{L}(G) \subset \mathcal{L}(G')$. Conversely, if $p \in \mathcal{L}(G')$, setting $f_n(x_n, x_{\pi_n}) = \frac{\sum_{x'_1, ..., x'_{n-1}} \prod_{C \in \mathcal{C}_{n-1}} \psi_C(x'_C)}{\sum_{x'_1, ..., x'_n} \prod_{C \in \mathcal{C}_n} \psi_C(x'_C)} \psi_{\{n,\pi_n\}}(x_n, x_{\pi_n})$ shows $p \in \mathcal{L}(G)$ and therefore that $\mathcal{L}(G') \subset \mathcal{L}(G)$.

$$\boxed{\text{By induction, we have } \mathcal{L}(G) = \mathcal{L}(G') \text{ for any directed tree.}}$$

Exercise 3.a

To measure the initialization's impact on the results, we decided to run the algorithm several times (ten times in the results presented here) with random initializations. We then computed the mean and standard deviation of the distortion measure after convergence, as well as a transportation distance between cluster centers.

For four clusters, we get a mean distortion measure of 3239, with a standard deviation of 1.44. The maximum (over all couples of runs) transportation distance between centers is 0.23. Overall these results show that the algorithm's results are really robust to changes of initial conditons.

However, things are completely different for six clusters. In this case, we get a mean distortion measure of 1984, with a standard deviation of 257, and a maximum transportation distance between centers of 80. This is a drastic change with respect to the previous setting, which highlights the following fact :

K-means is highly sensitive to the choice of its main hyper-parameter : the number of clusters.

Exercise 3.b : M-step updates (isotropic case)

Derivations are detailed in the appendix. The M-step consists in maximizing the following quantity :

$$l_t(\pi, \mu, \Sigma) = \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j log(\pi_{j,t}) + \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j \left[log(\frac{1}{(2\pi)^{\frac{d}{2}}}) + log(\frac{1}{|\Sigma_{j,t}|^{\frac{1}{2}}}) - \frac{1}{2}(x_i - \mu_{j,t})^T \Sigma_{j,t}^{-1}(x_i - \mu_{j,t}) \right]$$

In the isotropic case, we have $\Sigma_{j,t} = \sigma_{j,t}^2 I$. Optimal solutions are given by :

$$\boxed{\hat{\pi}_{j,t} = \frac{1}{n}\sum_{i=1}^{n} \tau_i^j \qquad \hat{\mu}_{j,t} = \frac{\sum\limits_{i=1}^{n} \tau_i^j x_i}{\sum\limits_{i=1}^{n} \tau_i^j} \qquad \hat{\sigma}_{j,t}^2 = \frac{\sum_{i=1}^{n} \tau_i^j ||x_i - \hat{\mu}_{j,t}||_2^2}{d \sum_{i=1}^{n} \tau_i^j}}$$

Exercise 3.c : covariance estimator (general case)

If we do not restrict ourselves to the case where $\Sigma_{j,t} = \sigma_{j,t}^2 I$, and consider general covariance matrices $\Sigma_{j,t}$, the optimal solution is :

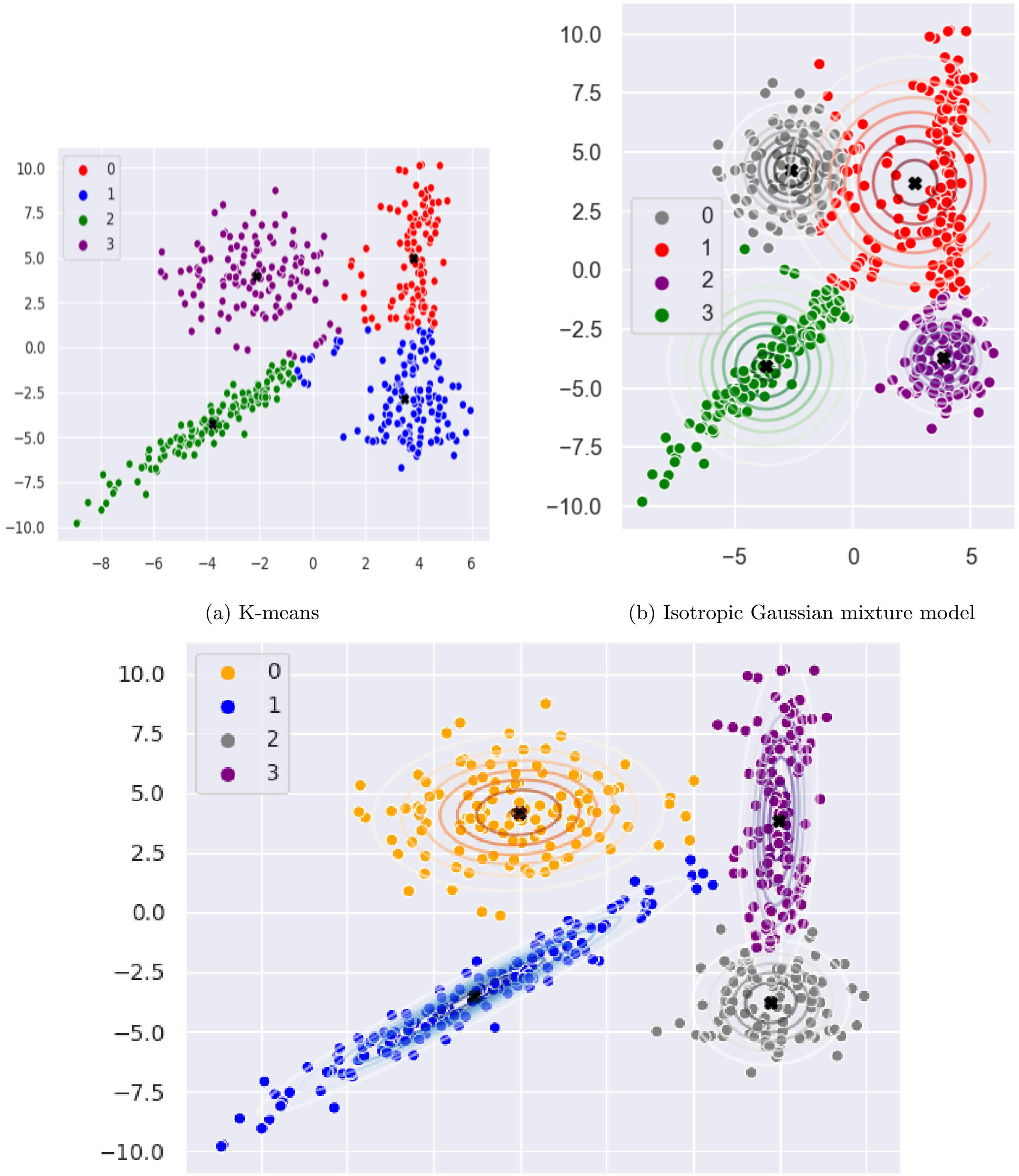
$$\boxed{\hat{\Sigma}_{j,t} = \frac{\sum_{i=1}^{n} \tau_i^j (x_i - \hat{\mu}_{j,t})(x_i - \hat{\mu}_{j,t})^T}{\sum_{i=1}^{n} \tau_i^j}}$$

Exercise 3.d

The following table sums up the results for both models :

Model	log likelihood (training set)	log likelihood (test set)
Isotropic GMM	$-2.7\ 10^3$	$-7.4\ 10^3$
General GMM	$-2.4\ 10^3$	$-19.0\ 10^3$

We see that the general Gaussian mixture model has a higher training log likelihood than the isotropic model, but also has a smaller test log likelihood. This result is consistent with the fact that the former is more expressive, which means it can fit better to some training data, but is also more likely to overfit.



(a) K-means

(b) Isotropic Gaussian mixture model

(c) General Gaussian mixture model

Appendix

Exercise 1.2.(b)

Let us consider three random variables X, Y, Z on finite spaces, such that their joint distribution p satisfies $p(x, y, z) = p(x \mid z)\, p(y \mid z)\, p(z)$. We therefore have that $X \perp Y \mid Z$. We also have that $X \perp Y$ if and only if $p(x, y) = p(x)\, p(y)$.

Let us now assume that the support of Z is of cardinal 3 (without loss of generality, we can assume it is exactly {0, 1, 2}). Adapting the derivations of question 1.2.(a) yields $X \perp Y$ if and only if :

$$\begin{aligned}
&p(z=0)\, p(z=1)\big(p(x \mid z=0) - p(x \mid z=1)\big)\big(p(y \mid z=0) - p(y \mid z=1)\big) \\
&+ p(z=0)\, p(z=2)\big(p(x \mid z=0) - p(x \mid z=2)\big)\big(p(y \mid z=0) - p(y \mid z=2)\big) \\
&+ p(z=1)\, p(z=2)\big(p(x \mid z=1) - p(x \mid z=2)\big)\big(p(y \mid z=1) - p(y \mid z=2)\big) = 0
\end{aligned}$$

This is satisfied, for example, if $Z \sim \mathcal{U}(\{0,1,2\})$, $p(x \mid z=0) \neq p(x \mid z=1) = p(x \mid z=2)$ and :

$$p(y \mid z=0) = \frac{1}{2}\big(p(y \mid z=1) + p(y \mid z=2)\big)$$

Let us now construct a concrete counter-example. Suppose we have a fair three face dice D, and four different coins C_1, C_2, C_3, C_4. We take Z as the result of rolling D. If we observe $Z = 0$, X is the result of tossing C_1. Otherwise, X is the result of a tossing C_2. If we observe $Z = 1$ (resp. $Z = 2$), Y is the result of tossing C_3 (resp. C_4). If we observe $Z = 0$, C_3 and C_4 are tossed with probability one half each. The experiment's design ensures $X \perp Y \mid Z$. Since C_1 and C_2 are different, we therefore have $p(x \mid z=0) \neq p(x \mid z=1) = p(x \mid z=2)$. We also have $p(y \mid z=0) = \frac{1}{2}\big(p(y \mid z=1) + p(y \mid z=2)\big)$. From the derivations above we therefore get that $X \perp Y$. However since $p(x \mid z=0) \neq p(x \mid z=1)$ and $p(y \mid z=1) \neq p(y \mid z=2)$ we neither have $X \perp Z$ nor $Y \perp Z$.

Exercise 3.b

Let us derive the form of the M-step in the case of covariance matrices proportional to the identy. The M-step consists in maximizing the following quantity :

$$l_t(\pi, \mu, \Sigma) = \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j log(\pi_{j,t}) + \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j \left[log(\frac{1}{(2\pi)^{\frac{d}{2}}}) + log(\frac{1}{|\Sigma_{j,t}|^{\frac{1}{2}}}) - \frac{1}{2}(x_i - \mu_{j,t})^T \Sigma_{j,t}^{-1}(x_i - \mu_{j,t}) \right] \quad (6)$$

where $\Sigma_{j,t} = \sigma_{j,t}^2 I$. The maximization with respect to $\pi_{j,t}$ and $\mu_{j,t}$ is the same as in the general case and we denote $\hat{\pi}_{j,t}$, $\hat{\mu}_{j,t}$ the optimal solutions. The maximization of (6) with respect to $\sigma = (\sigma_{j,t})_{j=1,\ldots,k}$ writes :

$$\max_{\sigma > 0} \sum_{j=1}^{k} d\ log(\frac{1}{\sigma_{j,t}}) \left[\sum_{i=1}^{n} \tau_i^j \right] - \frac{1}{2\sigma_{j,t}^2} \left[\sum_{i=1}^{n} \tau_i^j ||x_i - \widehat{\mu_{j,t}}||_2^2 \right] \quad (7)$$

Performing a change of variable $\lambda_{j,t} = \frac{1}{\sigma_{j,t}}$ in (7) yields :

$$\max_{\lambda > 0} \sum_{j=1}^{k} d\ log(\lambda_{j,t}) \left[\sum_{i=1}^{n} \tau_i^j \right] - \frac{\lambda_{j,t}^2}{2} \left[\sum_{i=1}^{n} \tau_i^j ||x_i - \hat{\mu}_{j,t}||_2^2 \right] \quad (8)$$

The objective in (8), that we will denote f, is now concave in λ, as a sum of positive weighted concave functions.

If, for a given $j \in \{1, ..., k\}$, we have $\sum\limits_{i=1}^{n} \tau_i^j = 0$, then for all i, $\tau_i^j = 0$, and the value of the objective does not depend on $\lambda_{j,t}$. We can therefore restrict ourselves to the case where, for all j, there exist at least one i such that $\tau_i^j = 1$.

In that case, f goes to $+\infty$ as any $\lambda_{j,t}$ goes to 0, and its domain is therefore $(\mathbb{R}_+^*)^k$. We are therefore maximizing a concave differentiable function over a convex domain.

$$\frac{\partial\, l_t}{\partial \lambda_{j,t}} = \sum_{i=1}^{n} \tau_i^j \left[\frac{d}{\lambda_{j,t}} - \lambda_{j,t} ||x_i - \hat{\mu}_{j,t}||_2^2 \right]$$

f therefore has unique stationary point $\hat{\lambda}_{j,t} = \sqrt{\frac{d \sum_{i=1}^{n} \tau_i^j}{\sum_{i=1}^{n} \tau_i^j ||x_i - \hat{\mu}_{j,t}||_2^2}}$, which is therefore optimal.

We finally get :

$$\boxed{\hat{\sigma}_{j,t}^2 = \frac{\sum_{i=1}^{n} \tau_i^j ||x_i - \hat{\mu_{j,t}}||_2^2}{d \sum_{i=1}^{n} \tau_i^j}}$$

Let us stress that this expression is valid only if $\sum\limits_{i=1}^{n} \tau_i^j \neq 0$. Otherwise, the value of the objective does not depend on $\sigma_{j,t}$.