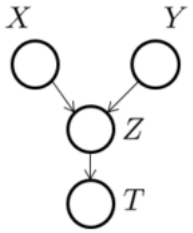
IFT6269-A2017	**Hwk 3 - Solutions**	Name:
Prof: Simon Lacoste-Julien	Due date: Oct 24, 2017	Student id:

As usual, please hand in on paper form your derivations and answers to the questions. You can use any programming language for your source code (submitted on Studium as per the website instructions). All the requested figures should be printed on paper with clear titles that indicate what the figures represent.

1. **DGM (5 points)**
Consider the directed graphical model G on the right. Write down the implied factorization for any joint distribution $p \in \mathcal{L}(G)$. Is it true that $X \perp\!\!\!\perp Y \mid T$ for any $p \in \mathcal{L}(G)$? Prove or disprove.



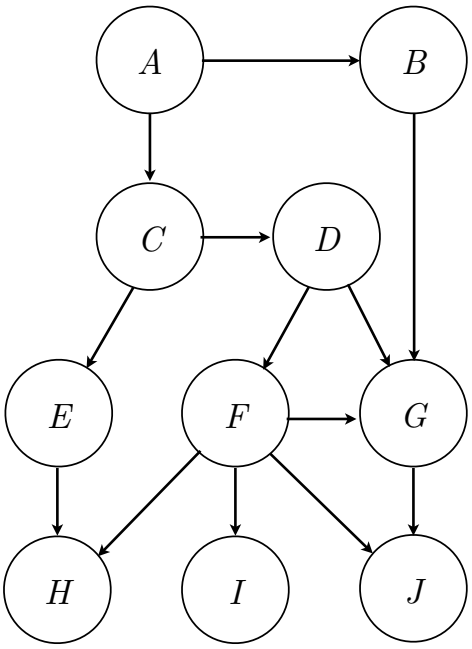
Solution:

For $p \in \mathcal{L}(G)$, the factorization is: $p(x, y, z, t) = p(t|z)p(z|x, y)p(x)p(y)$. The answer is no, X and Y have in general no reason to be independent given T: take X and Y i.i.d., $Z = 1$ if $X < Y$, $Z = 0$ else, and set $T = Z$. Then clearly X and Y are dependent given T. Now, even if T is not deterministic given Z the same problem persists: as a concrete example consider the case of binary variables with $Z = 1$ if and only if $X = Y$ and $p(Z = 1|T = t) = \pi(t)$. Then

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

We therefore have $\mathbb{P}(X = 1, Y = 1|T = t) = \mathbb{P}(X = 0, Y = 0|T = t) = \pi(t)$ and $\mathbb{P}(X = 0, Y = 1|T = t) = \mathbb{P}(X = 1, Y = 0|T = t) = 1 - \pi(t)$. This conditional distribution of (X, Y) can written as a two-by-two table, and conditional independence would mean that this two-by-two table viewed as a matrix is of rank 1, which entails that its determinant is 0. But this is only true if $\pi(t) = 0.5$ which would force T to be independent from Z.

2. **d-separation in DGM (5 points)**
Indicate (yes or no) which conditional independence statements are true?



(a) $C \perp B$?

(b) $C \perp B \mid A$?

(c) $C \perp B \mid A, J$?

(d) $C \perp B \mid A, J, D$?

(e) $C \perp G$?

(f) $C \perp G \mid B$?

(g) $C \perp G \mid B, D$?

(h) $C \perp G \mid B, D, H$?

(i) $C \perp G \mid B, D, H, E$?

(j) $B \perp I \mid J$?

Solution:

(a) No (b) Yes (c) No (d) Yes (e) No (f) No (g) Yes (h) No (i) Yes (j) No

3. **Positive interactions in-V-structure (10 points)**
Let X, Y, Z be binary random variables with a joint distribution parametrized according to the graph: $X \to Z \leftarrow Y$. We define the following:

$$a := P(X = 1), \quad b := P(X = 1 \mid Z = 1), \quad c := P(X = 1 \mid Z = 1, Y = 1)$$

(a) For all the following cases, provide examples of conditional probability tables (and calculate the quantities a, b, c), that render the statements true:

(i) $a > c$

(ii) $a < c < b$

(iii) $b < a < c$.

(b) Think of X and Y as causes and Z as a common effect, for all the above cases (i, ii, et iii), summarize in a sentence or two why the declarations are true for your examples.

Solution (solution reproduced with permission from Dong-Hyun Lee):

We can factorize $P(X, Y, Z)$ using v-Structure $X \to Z \leftarrow Y$.

$$P(X, Y, Z) = P(Z|X, Y)P(X)P(Y)$$

(i) (a)

X	$P(X)$
1	0.5
0	0.5

Y	$P(Y)$
1	0.5
0	0.5

X	Y	$P(Z = 1\|X, Y)$
1	1	0
0	1	0
1	0	1
0	0	1

$$a = P(X) = 0.5$$

$$c = \frac{P(X = 1, Y = 1, Z = 1)}{\sum_X P(X, Y = 1, Z = 1)} = \frac{P(X = 1)P(Y = 1)P(Z = 1|X = 1, Y = 1)}{\sum_X P(X)P(Y = 1)P(Z = 1|X, Y = 1)} = 0$$

$$a = 0.5 > c = 0.0$$

(b) When both causes X, Y occur together, a common effect Z never occur according to the CPT. If we observe Z, Y, X must not occurred. So possibility of X disappears, while X is probable without any observation.

(ii) (a)

X	$P(X)$
1	0.5
0	0.5

Y	$P(Y)$
1	0.5
0	0.5

X	Y	$P(Z = 1\|X, Y)$
1	1	0.9
1	0	1
0	1	0.1
0	0	0

$$a = P(X) = 0.5$$

$$b = \frac{\sum_Y P(X = 1, Y, Z = 1)}{\sum_{XY} P(X, Y, Z = 1)} = \frac{\sum_Y P(X = 1)P(Y)P(Z = 1|X = 1, Y)}{\sum_{XY} P(X)P(Y)P(Z = 1|X, Y)} = 0.95$$

$$c = \frac{P(X=1,Y=1,Z=1)}{\sum_X P(X,Y=1,Z=1)} = \frac{P(X=1)P(Y=1)P(Z=1|X=1,Y=1)}{\sum_X P(X)P(Y=1)P(Z=1|X,Y=1)} = 0.9$$
$$a = 0.5 < c = 0.9 < b = 0.95$$

(b) When we observe Z, X is more probable because Z is observed with high probability when X occurred regardless of Y ($0.9 + 1 > 0.1 + 0$ in the CPT). But if we also observe Y, X is less probable than it ($0.9 < 1$ in the CPT).

(iii) (c)

X	$P(X)$
1	0.5
0	0.5

Y	$P(Y)$
1	0.5
0	0.5

X	Y	$P(Z=1\|X,Y)$
1	1	0.9
1	0	0
0	1	0.1
0	0	1

$$a = P(X) = 0.5$$
$$b = \frac{\sum_Y P(X=1,Y,Z=1)}{\sum_{XY} P(X,Y,Z=1)} = \frac{\sum_Y P(X=1)P(Y)P(Z=1|X=1,Y)}{\sum_{XY} P(X)P(Y)P(Z=1|X,Y)} = 0.45$$
$$c = \frac{P(X=1,Y=1,Z=1)}{\sum_X P(X,Y=1,Z=1)} = \frac{P(X=1)P(Y=1)P(Z=1|X=1,Y=1)}{\sum_X P(X)P(Y=1)P(Z=1|X,Y=1)} = 0.9$$
$$b = 0.45 < a = 0.5 < c = 0.9$$

(b) When we observe Z, X is less probable because Z is observed with less probability when X occurred regardless of Y ($0.9 + 0 < 0.1 + 1$ in the CPT). But if we also observe Y, X is more probable ($0.9 > 0.1$ in the CPT) than without observation.

4. **Flipping a covered edge in a DGM (10 points)**
Let $G = (V, E)$ be a DAG. We say that a directed edge $(i, j) \in E$ is a *covered edge* if and only if $\pi_j = \pi_i \cup \{i\}$. let $G' = (V, E')$, with $E' = (E \backslash \{(i, j)\}) \cup \{(j, i)\}$. Prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

Solution:

At first, we will prove G' is also a DAG by contradiction. Let's assume that G' has a cycle while G has no cycle. The cycle must have nodes i and j because it is created by flipping the edge $i \to j$ into $i \leftarrow j$. This implies that there is a path $i \to ... \to k \in \pi_j$ to form a cycle because j takes incoming arrows only from $k \in \pi_j$ such as $i \to ... \to k \in \pi_j \to j \to i$. However, $k \in \pi_i \cup \{i\}$ from a covered edge $\pi_j = \pi_i \cup \{i\}$. So the above path implies that there is a cycle $i \to ... \to k \in \pi_i \to i$ or $i \to ... \to i$ regardless of j even before the flipping. It's contradiction, so G' is also a DAG.

Let $p \in \mathcal{L}(G)$. We thus have $p(x) = \prod_{k=1}^n p(x_k \mid x_{\pi_k})$, where π_k denotes the parents of k in G. Consider any x_i, x_j, x_{π_i} such that $p(x_i, x_j, x_{\pi_i}) \neq 0$. Then by the chain rule (valid for any distribution), we have

$$p(x_i \mid x_{\pi_i})p(x_j \mid x_i, x_{\pi_i}) = p(x_i, x_j \mid x_{\pi_i}) = p(x_j \mid x_{\pi_i})p(x_i \mid x_j, x_{\pi_i}). \quad (1)$$

As (i, j) is a covered edge, we have $\pi_j = \pi_i \cup \{j\}$. Moreover, by definition of E', we have $\pi'_j = \pi_i$ and $\pi'_i = \pi_j \cup \{j\}$ with π'_i the parents of i in G'. So note that equation (1) can be interpreted as:

$$p(x_i \mid x_{\pi_i})p(x_j \mid x_{\pi_j}) = p(x_j \mid x_{\pi'_j})p(x_i \mid x_{\pi'_i}).$$

As $\pi'_k = \pi_k$ for any $k \neq i, j$, we can simply swap the two terms for i and j in the product factorization of p:

$$p(x) = p(x_i \mid x_{\pi_i}) p(x_j \mid x_{\pi_j}) \prod_{k \neq i,j} p(x_k \mid x_{\pi_k}) = p(x_j \mid x_{\pi'_j}) p(x_i \mid x_{\pi'_i}) \prod_{k \neq i,j} p(x_k \mid x_{\pi'_k}).$$

If $p(x_i, x_j, x_{\pi_i}) = 0$, then both the LHS and RHS above are equal to zero and so are still equal. We thus have $p \in \mathcal{L}(G')$. By symmetry, we can reverse the argument, and thus $\mathcal{L}(G) = \mathcal{L}(G')$.

5. **Equivalence of directed tree DGM with undirected tree UGM (10 points)**
Let G be a directed tree and G' its corresponding undirected tree (where the orientation of edges is ignored). Recall that by the definition of a directed tree, G does not contain any v-structure. Prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

Solution:

If $p \in \mathcal{L}(G)$, then $p(x) = \prod_{j=1}^n p(x_j | x_{\pi_j})$ where $|\pi_j| \leq 1$ as G is a directed tree (has no v-structure). Thus denoting $\psi_j(x_j, x_{\pi_j}) = p(x_j | x_{\pi_j})$, p may be written as the Gibbs model $p(x) = \prod_{j=1}^n \psi_j(x_j, x_{\pi_j})$ and thus $p \in \mathcal{L}(G')$.

For the other direction, we show the result by induction on the size of undirected trees. That is, our induction hypothesis is that for any undirected tree $G' = (V, E')$ with $|V| \leq n$, then $p \in \mathcal{L}(G') \implies p \in \mathcal{L}(G)$ for any directed tree G which is an orientation of G'.

The case $n = 1$ is trivial ($\mathcal{L}(G')$ = all distributions on one node = $\mathcal{L}(G)$).

So now consider an undirected tree $G' = (V, E')$ with $n > 1$ nodes, and $G = (V, E)$ some directed tree version of G'. Let's index the nodes of V from 1 to n so that node n is a leaf which is not the root of the directed tree G and its unique parent is the node $n-1$. For $n > 1$, there exists such a leaf distinct from the root, and for this leaf, we have $(n-1, n) \in E$. Let $p \in \mathcal{L}(G')$, and so we have $p(x) = \frac{1}{Z} \prod_{\{i,j\} \in E'} \psi_{ij}(x_i, x_j)$.

Let $\tilde{p}$ be the marginal of p on $x_{1:(n-1)}$. Then we have:

$$\tilde{p}(x_{1:(n-1)}) = \frac{1}{Z} \tilde{\psi}(x_{n-1}) \prod_{\{i,j\} \in E' \setminus \{n-1,n\}} \psi_{ij}(x_i, x_j) \quad \text{where} \quad \tilde{\psi}(x_{n-1}) := \sum_{x_n} \psi(x_{n-1}, x_n).$$

Let $\tilde{G}$ be the subtree of size $n-1$ obtained from G by removing the leaf n, and $\tilde{G}'$ its undirected version. From the form above, we see that $\tilde{p} \in \mathcal{L}(\tilde{G}')$. Thus by the induction hypothesis, $\tilde{p} \in \mathcal{L}(\tilde{G})$ and so factorizes as: $\tilde{p}(x_1, \ldots, x_{n-1}) = \prod_{i=1}^{n-1} \tilde{p}(x_i | x_{\pi_i})$. Note that in G, $\pi_n = \{n-1\}$; we thus define $f(x_n, x_{\pi_n})$ through

$$f_n(x_n, x_{\pi_n}) := \begin{cases} \psi_{n-1,n}(x_{n-1}, x_n) / \tilde{\psi}(x_{n-1}) & \text{if } \tilde{\psi}(x_{n-1}) \neq 0 \\ 1/K_n & \text{otherwise} \end{cases}$$

with K_n the number of possible values for X_n. We then have, valid for all x:

$$p(x) = \tilde{p}(x_1, \ldots, x_{n-1}) f_n(x_n, x_{\pi_n}) = f_n(x_n, x_{\pi_n}) \prod_{i=1}^{n-1} \tilde{p}(x_i | x_{\pi_i}).$$

Now since $\sum_{x_n} f_n(x_n, x_{\pi_n}) = 1$, we have that p satisfies the conditions in the definition of $\mathcal{L}(G)$, and thus $p \in \mathcal{L}(G)$, completing the induction step and the proof.

We have just shown that oriented and non-oriented trees are *Markov-equivalent*.

6. **Hammersley-Clifford Counter example (10 points)**
In class, I mentioned that the strict positivity of the joint distribution was crucial in the Hammersley-Clifford theorem. Here is a counter-example that shows the problems when we have zero probabilities (it is example 4.4 in Koller & Friedman). Consider a joint distribution p over four binary random variables: X_1, X_2, X_3 and X_4 which gives probability $\frac{1}{8}$ to each of the following eight configurations, and probability zero to all others:

$$\begin{array}{cccc} (0,0,0,0) & (1,0,0,0) & (1,1,0,0) & (1,1,1,0) \\ (0,0,0,1) & (0,0,1,1) & (0,1,1,1) & (1,1,1,1). \end{array}$$

Let G be the usual four nodes undirected graph X_1—X_2—X_3—X_4—X_1. One can show that p satisfies the global Markov property with respect to this graph G because of trivial deterministic relationships. For example, if we condition on $X_2 = 0$ and $X_4 = 0$, then the only value of X_3 with non-zero probability is $X_3 = 0$, and thus $X_3|X_2 = 0, X_4 = 0$ being a deterministic random variable, it is trivially conditionally independent to X_1. By (painfully) going through all other possibilities, we get similar situations (for example $X_2 = 0$ and $X_4 = 1$ forces $X_1 = 0$, etc.). Prove that the distribution p *cannot* factorize according to G (and thus $p \notin \mathcal{L}(G)$). *Hint: argue by contradiction.*

Solution (solution reproduced with permission from Dong-Hyun Lee):

(adopted from page 6 of https://www.cs.helsinki.fi/group/cosco/Teaching/Probability/2010/lecture4_MRF.4pg.pdf)

If p admit a factorized representation in the undirected graph $X_1 - X_2 - X_3 - X_4 - X_1$,

$$P(X_1, X_2, X_3, X_4) = \frac{1}{Z}\phi_{12}(X_1, X_2)\phi_{23}(X_2, X_3)\phi_{34}(X_3, X_4)\phi_{41}(X_4, X_1)$$

We can have following equations :

$$\begin{aligned} 1/8 = P(0,0,0,0) &= \phi_{12}(0,0)\phi_{23}(0,0)\phi_{34}(0,0)\phi_{41}(0,0) \quad ... \quad (a) \\ 0 = P(0,0,1,0) &= \phi_{12}(0,0)\phi_{23}(0,1)\phi_{34}(1,0)\phi_{41}(0,0) \quad ... \quad (b) \\ 1/8 = P(0,0,1,1) &= \phi_{12}(0,0)\phi_{23}(0,1)\phi_{34}(1,1)\phi_{41}(1,0) \quad ... \quad (c) \end{aligned}$$

From (a) and (c),

$$\phi_{12}(0,0) \neq 0, \ \phi_{41}(0,0) \neq 0, \ \phi_{23}(0,1) \neq 0$$

But there is a factor in (b) which is zero. so

$$\phi_{34}(1,0) = 0$$

But

$$1/8 = P(1,1,1,0) = \phi_{12}(1,1)\phi_{23}(1,1)\phi_{34}(1,0)\phi_{41}(0,1)$$

It must be zero because $\phi_{34}(1,0) = 0$, but non-zero, so it's a contradiction, so p cannot be factorized according to G and $p \notin \mathcal{L}(G)$.

7. **[BONUS]: bizarre conditional independence properties (10 bonus points)**
Let (X, Y, Z) be a random vector with a finite sample space. Consider the following statement:

"If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Y$ then ($X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$)."

(a) Is this true if one assumes that Z is a binary variable? Prove or disprove.

(b) Is the statement true in general? Prove or disprove.

Solution:

(a) If Z is binary, the statement is true. Let's prove it. If Y is a constant r.v. (i.e. $\exists y_0$ s.t. $\mathbb{P}(Y = y_0) = 1$), then Y is trivially independent with any r.v. (verify it!), and so $Y \perp\!\!\!\perp Z$. So we now assume that Y takes at least two distinct values with non-zero probability. For any y such that $p(y) \neq 0$, we have

$$p(x) \overset{X \perp\!\!\!\perp Y}{=} \frac{p(x,y)}{p(y)} = \frac{1}{p(y)} \sum_z p(x, y|z)p(z)$$
$$\overset{X \perp\!\!\!\perp Y|Z}{=} \frac{1}{p(y)} \sum_z p(x|z)p(y|z)p(z) = \sum_z p(x|z)p(z|y).$$

Since Z is binary, we thus have for any j such that $\mathbb{P}(Y = j) \neq 0$,

$$\mathbb{P}(X = i) = \mathbb{P}(X = i|Z = 1)\mathbb{P}(Z = 1|Y = j) + \mathbb{P}(X = i|Z = 0)\mathbb{P}(Z = 0|Y = j).$$

Let $u^{(k)}$ be the vector such that $u_i^{(k)} = \mathbb{P}(X = i|Z = k)$ and $v^{(k)}$ be the vector such that $v_j^{(k)} = \mathbb{P}(Z = k|Y = j)$ then

$$A = u^{(0)} {v^{(0)}}^\top + u^{(1)} {v^{(1)}}^\top$$

is the matrix such that $A_{ij} = \mathbb{P}(X = i)$. The columns of A are thus all equal, which means that $u^{(0)}v_j^{(0)} + u^{(1)}v_j^{(1)} = u^{(0)}v_{j'}^{(0)} + u^{(1)}v_{j'}^{(1)}$ for any j, j' such that $\mathbb{P}(Y = j) \neq 0$ and $\mathbb{P}(Y = j') \neq 0$. Since we assume that Y must take at least two different values with non-zero probability, we have that

$$u^{(0)}(v_j^{(0)} - v_{j'}^{(0)}) + u^{(1)}(v_j^{(1)} - v_{j'}^{(1)}) = 0,$$

and so either $u^{(0)}$ and $u^{(1)}$ are collinear or we have both $v_j^{(0)} = v_{j'}^{(0)}$ and $v_j^{(1)} = v_{j'}^{(1)}$.

- In the first case $u^{(0)} = \gamma u^{(1)}$, but we must have $\gamma = 1$ because the entries in $u^{(k)}$ must sum to 1 (it is a probability distribution). So $\mathbb{P}(X|Z = 0) = \mathbb{P}(X|Z = 1)$, implying that $X \perp\!\!\!\perp Z$ (fill in the last details!).
- In the second case, $v_j^{(0)} = v_{j'}^{(0)}$ and $v_j^{(1)} = v_{j'}^{(1)}$ for all pairs (j, j') such that $\mathbb{P}(Y = j) \neq 0$ and $\mathbb{P}(Y = j') \neq 0$. But this means that $\mathbb{P}(Z = 1|Y = j)$ and $\mathbb{P}(Z = 0|Y = j)$ do not depend on j for any j (note in particular that if $\mathbb{P}(Y = j) = 0$ we can set $\mathbb{P}(Z = 1|Y = j) = \mathbb{P}(Z = 1)$ and $\mathbb{P}(Z = 0|Y = j) = \mathbb{P}(Z = 0)$ because on an event of probability 0 the conditional probability can be defined arbitrarily), which means that $Y \perp\!\!\!\perp Z$.

(b) The statement is not true in general. Take (X, Z_1) with dependent components and (Y, Z_2) with also dependent components, such that $(X, Z_1) \perp\!\!\!\perp (Y, Z_2)$. By decomposition, we have $X \perp\!\!\!\perp Y$ (and also $Z_1 \perp\!\!\!\perp Z_2$). Now define $Z = (Z_1, Z_2)$; we also have $X \perp\!\!\!\perp Y \mid Z$ by some kind of decomposition; but let's verify it formally. Note that $p(x, z) = p(x, z_1)p(z_2)$ and that $p(z) = p(z_1)p(z_2)$ so that $p(x|z) = p(x|z_1)$. Symmetrically $p(y|z) = p(y|z_2)$. Thus

$$p(x, y|z) = \frac{p(x, y, z_1, z_2)}{p(z_1, z_2)} = \frac{p(x, z_1)p(y, z_2)}{p(z_1)p(z_2)} = p(x|z_1)p(y|z_2) = p(x|z)p(y|z),$$

so that $X \perp\!\!\!\perp Y \mid Z$, which verifies our claim.

Now by our constructed dependence structure, we cannot have $X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$ (assuming that X, Y, Z_1 and Z_2 each are random variables with at least two possible values).

Note that a particular instance of the situation above is the case where $Z_1 = X$ and $Z_2 = Y$, in which case $Z = (X, Y)$, which provides a simple counterexample, because, conditionally on Z, then X and Y are determined and thus independent. Also note here that Z takes more than 2 values, explaining why this is not contradicting what we prove in part (a).

8. **Implementation: EM and Gaussian mixtures (30 points)**
The file `EMGaussian.train` contains samples of data x_i where $x_i \in \mathbb{R}^2$ (one datapoint per row). The goal of this exercise is to implement the EM algorithm for some mixtures of K Gaussians in $\mathbb{R}^d$ (here $d = 2$ and $K = 4$), for i.i.d. data. (NB: in this exercise, no need to prove any of the formula used in the algorithms except for question (b)).

(a) Implement the K-means algorithm. Represent graphically the training data, the cluster centers, as well as the different clusters (use 4 colors). Try several random initializations and compare results (centers and the actual K-means objective values).

(b) Consider a Gaussian mixture model in which the covariance matrices are proportional to the identity. Derive the form of the M-step updates for this model and implement the corresponding EM algorithm (using an initialization with K-means).

Represent graphically the training data, the centers, as well as the covariance matrices (an elegant way is to represent the ellipse that contains a specific percentage, e.g., 90%, of the mass of the Gaussian distribution).

Estimate and represent (e.g. with different colors or different symbols) the most likely latent variables for all data points (with the parameters learned by EM).

(c) Implement the EM algorithm for a Gaussian mixture with general covariance matrices. Represent graphically the training data, the centers, as well as the covariance matrices.

Estimate and represent (e.g. with different colors or different symbols) the most likely latent variables for all data points (with the parameters learned by EM).

(d) Comment the different results obtained in earlier questions. In particular, compare the normalized log-likelihoods of the two mixture models on the training data, as well as on test data (in `EMGaussian.test`). (Here normalize the log-likelihood by the number of observations (rows) – it makes the number more manageable for comparison and puts it on the right scale).

Solution:

(a) When initializing the centroids of K-means with K random points from the dataset, we obtain in general different results. Most of them are close to the minimum, but some of them may be quite far (see histogram).

(b) The result is close to K-means since we do not take into accounts correlations between variables. The isotropic covariance matrix estimator is (and following the course notations)

$$\Sigma_i^{(t+1)} = \frac{1}{d} \frac{\sum_n \tau_n^{i(t)} \|x_n - \mu_i^{(t+1)}\|^2}{\sum_n \tau_n^{i(t)}}$$

(NB: don't forget to divide by d). The other parameters estimate ($\mu_i^{(t+1)}$ and $\pi_i^{(t+1)}$) during the M-step are the same as seen in class.
A reasonable estimate for the value of the latent variable for each n can be made by maximizing the a posteriori probability $p(z_n|x_n)$, i.e., through $\arg\max_{1 \leq i \leq K} \tau_n^i$.
For a standard multivariate Gaussian, i.e., so that $\mu = 0$ et $\Sigma = I_d$, the disk corresponding to 90% of the mass is centered at zero and has radius R so that $P(r^2 \leqslant R^2) = .9$, r^2 being the sum of the d squares of independent standard univariate Gaussians. This is by definition a variable with a χ^2-distribution with d degrees of freedom. In the general case, the ellipse is obtained through an affine transformation (see code).

(c) The covariance matrix estimator is (and following the course notations)

$$\Sigma_i^{(t+1)} = \frac{\sum_n \tau_n^{i(t)} (x_n - \mu_i^{(t+1)})(x_n - \mu_i^{(t+1)})^\top}{\sum_n \tau_n^{i(t)}}$$

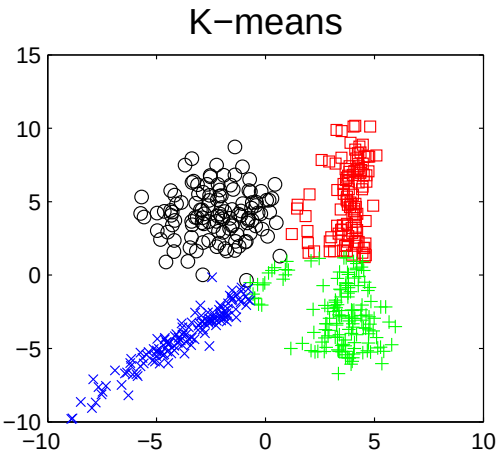
(d) We show below the log-likelihood divided by N_{train} and N_{test} respectively (we normalize to obtain values which remain small when the number of data points increases and to be able to compare "test" and "train"):

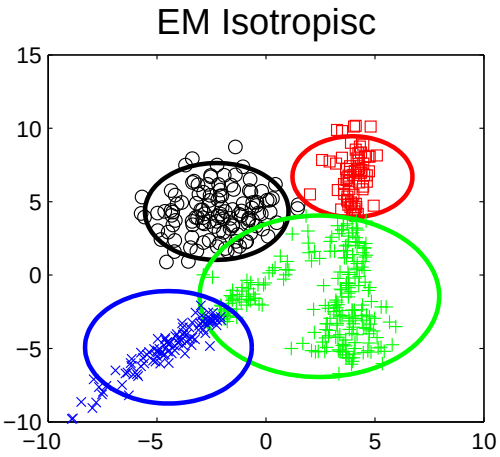
	Train	Test
Isotropic	-5.2910	-5.3882
General	-4.6554	-4.8180

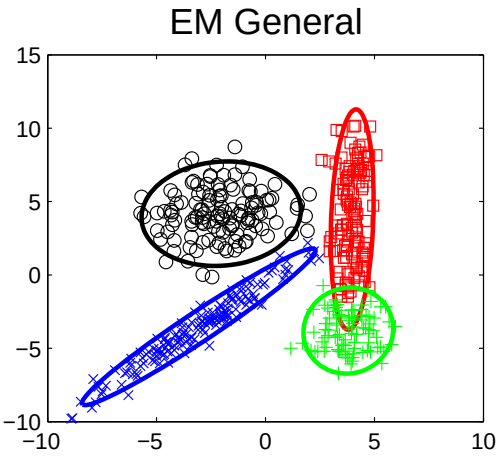
Unnormalized log-likelihoods:

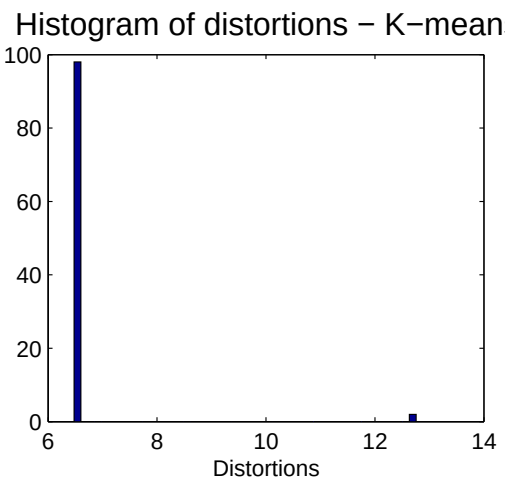
	Train	Test
Isotropic	-2.6455 $\times 10^3$	-2.6941 $\times 10^3$
General	-2.3277 $\times 10^3$	-2.4090 $\times 10^3$

The training log-likelihoods are always greater for more flexible models (the situation may be different for the testing log-likelihoods as the model may be too flexible and we have overfitting). The test log-likelihoods are on average lower than the train ones.


K−means
15
10
5
0
−5
−10
−10
−5
0
5
10


EM Isotropisc
15
10
5
0
−5
−10
−10
−5
0
5
10


EM General
15
10
5
0
−5
−10
−10
−5
0
5
10


Histogram of distortions − K−means
100
80
60
40
20
0
6
8
10
12
14
Distortions