Rhoulam Adil

PGM HOMEWORK2 REPORT

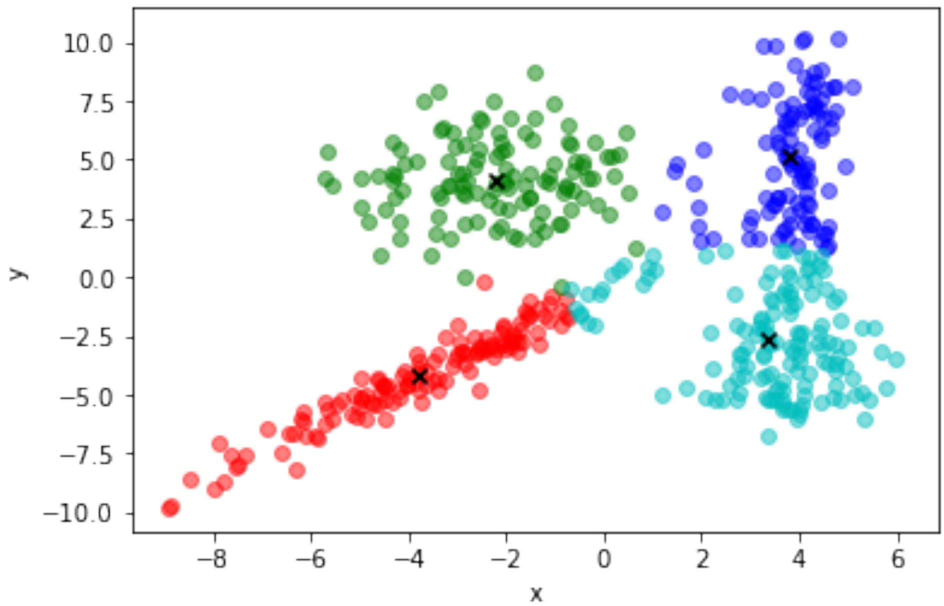
Please, see the scans in the end of the report for theoretical questions.

4 Implementation - Gaussian mixtures :

a) For the implementation of the first question: *K_means(data,k)* which is initialized randomly by k different points from data and returns a list containing in the first element the clusters found and in the second their centers. The function *distortion(kmeans)* takes the k_means results as arguments and return the distortion of the classes found.

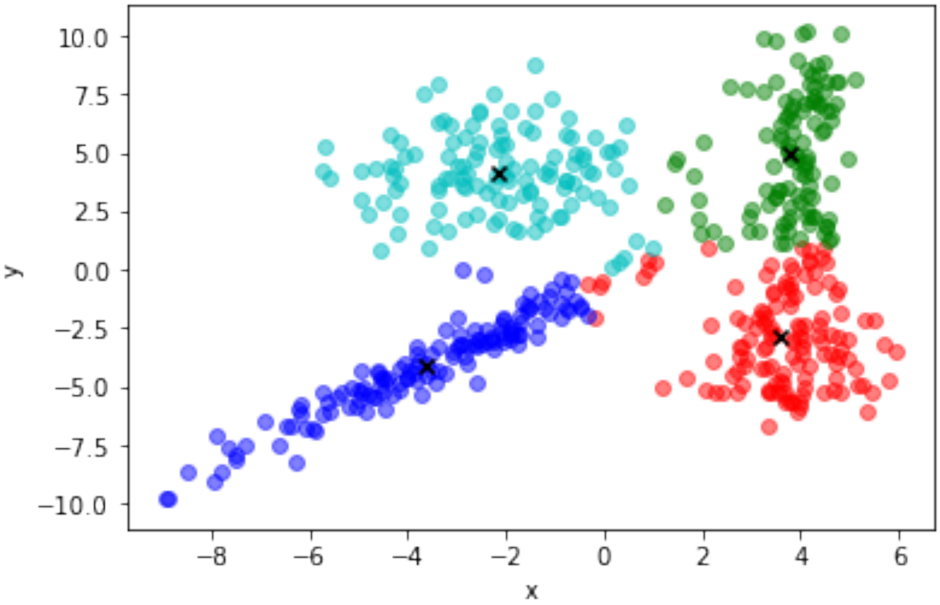
To plot the clusters and theirs centers, I implemented the function *plt_kmeans(cluster,center,color)* which *plot* the cluster with color color and its center in black.

Figure 1: K-means algorithm Result for a random intialization
Distortion = 1108.46




```
Cluster centers are: [[-3.78 -4.22]
                      [-2.24  4.16]
                      [ 3.8   5.1 ]
                      [ 3.36 -2.66]]
```

Figure 2: K-means algorithm Result for a second random intialization
Distortion = 1102.55




```
Cluster centers are: [[ 3.6  -2.89]
                      [-2.16  4.11]
                      [-3.64 -4.05]
                      [ 3.79  5.  ]]
```

Table 1: Comparison between different results of k-means algorithms on train data

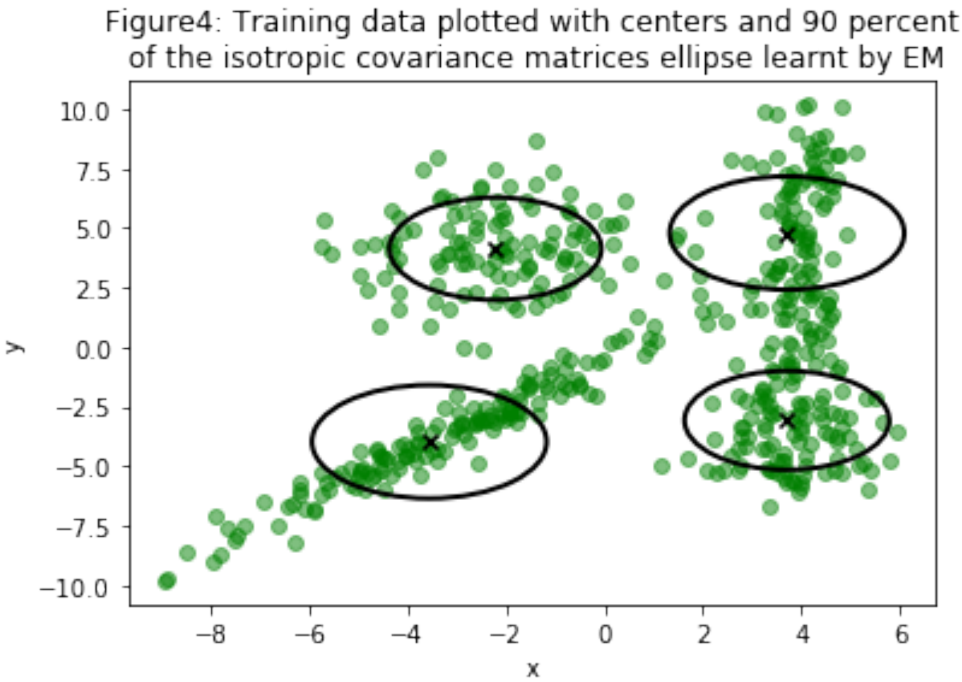
Clusters' centers	Distortion	Clusters' centers	Distortion
[[3.6 -2.89] [-2.16 4.11] [-3.64 -4.05] [3.79 5.]]	1102.54	[[-2.14 3.97] [-3.72 -4.18] [3.79 5.] [3.57 -2.88]]	1103.92
[[-3.78 -4.22] [-2.24 4.16] [3.8 5.1] [3.36 -2.66]]	1108.46	[[-2.24 4.13] [-3.82 -4.27] [3.8 5.1] [3.34 -2.64]]	1109.42
[[3.8 5.1] [-3.66 -4.07] [3.48 -2.7] [-2.24 4.24]]	1105.84	[[-3.78 -4.22] [3.36 -2.71] [-2.24 4.16] [3.8 5.03]]	1107.88

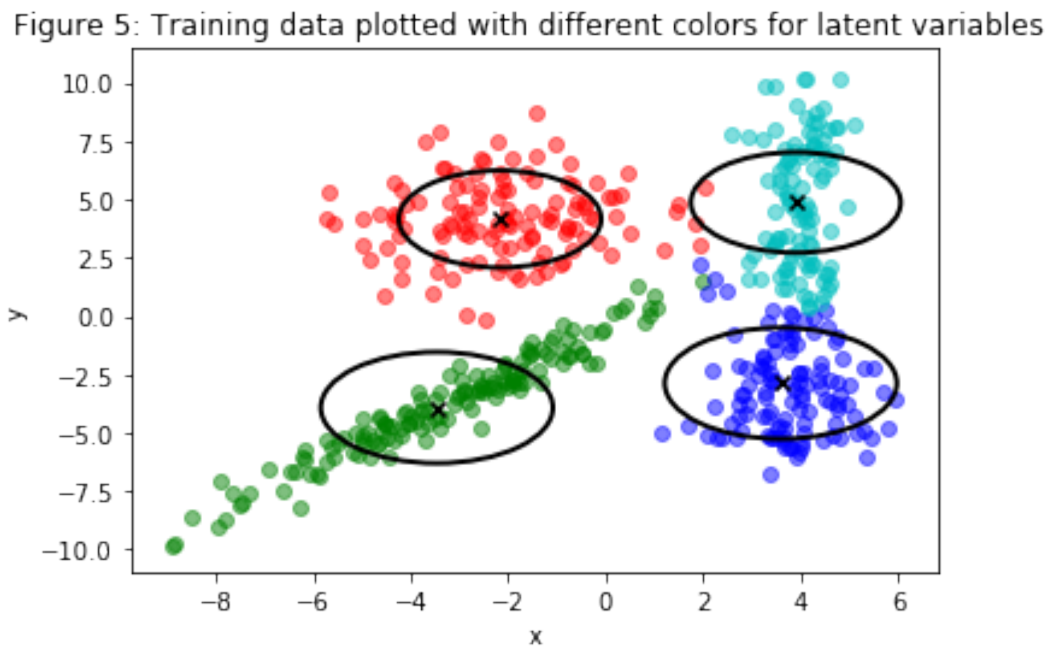
From the different results, we could conclude that k-means algorithm get trapped in local minimas and we have to run it several time with random initialization and preserve the best clustering which has the minimal distortion. However, we can notice that for the results in the table 1 the distortion value is in the range [1102,1110] which gives an error of 8/1106 = 0.7% which is not a large precision error. We can notice also that the centers are approximately the same (differ by the first value after comma).

b) For the theoretical question, please see the scans in the end of the end of the report.
For the implementation of the EM algorithm for Gaussian Mixture when the covariance matrices are isotropic (proportional to the identity matrix), I used the functions: *initializationStep(data,k):* Which returns initialized means of gaussians and weights with K_means algorithm and coefficients of covariances with a strict positive random value, *stepE(data,means,cov,weights,k)* which returns the responsibilities (posterior probabilities) and the current log-likelihood, *stepM(data,resp,k)* which returns the updated means, variances and weights given the responsabilities. Then implemented the EM algorithm with function *algorithmEM(data,k)* which returns the means, covariances, weights and responsibilities when the log-likelihood converges.

The graphical Representation of the training data, the centers, as well as 90% percentage of covariance ellipses learnt by EM algorithm is shown in Figure 4.

For the representation of the latent variables for all data points with the parameters learned by EM, I implemented the function *clusters(data,k)* which assigns each point to the cluster with the highest value of responsibility. The Figure 5 shows the clusters with different colors.

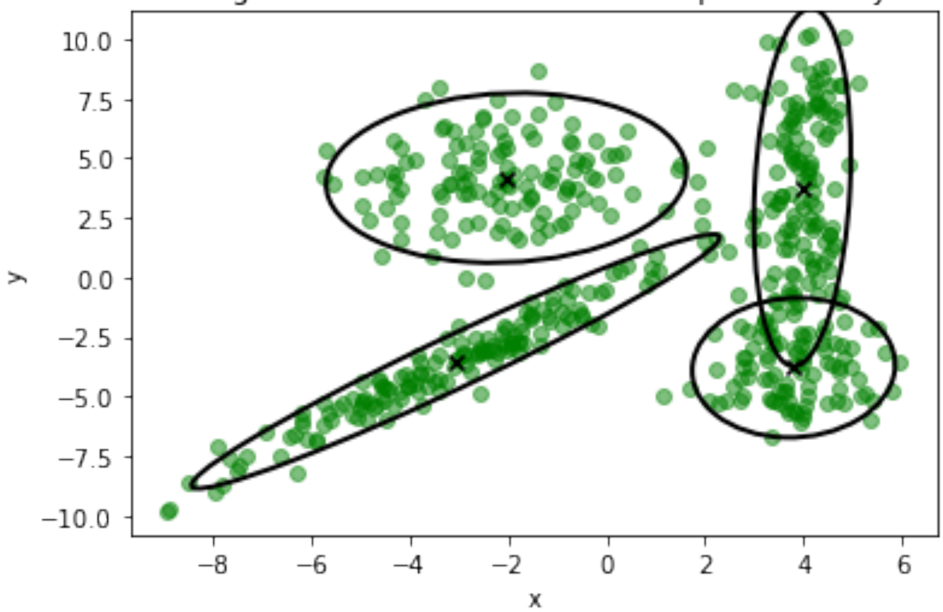




c). For the implementation of the EM algorithm for Gaussian Mixture when the covariance matrices are in general form, I used the functions: *intializationStepGeneralCase(data,k)* Which returns initialized means of gaussians, weights and covariances with K_means algorithm. *stepE(data,means,cov,weights,k)* the same one as for the isotropic case and *stepMGeneralCase(data,resp,k)* which returns the updated means, covariance matrices and weights given the responsabilities. Then implemented the EM algorithm with function *algorithmEMgeneralCase(data,k)* which returns the means, covariances, weights and responsibilities when the log-likelihood converges.

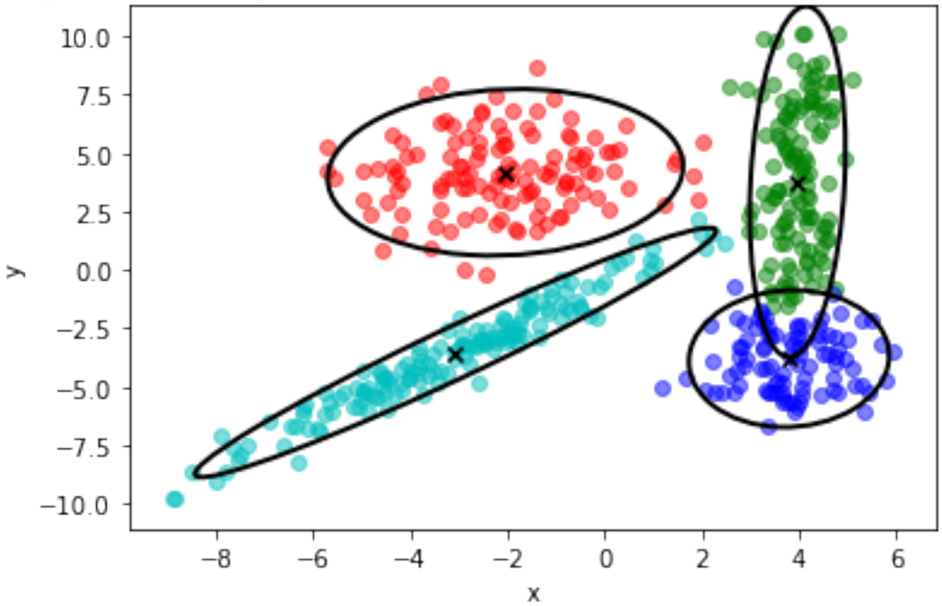
The graphical Representation of the training data, the centers, as well as 90% percentage of covariance ellipses learnt by EM algorithm is shown in Figure 6.

Figure 6: Training data plotted with centers and 90 percent of the general covariance matrices ellipse learnt by EM



For the representation of the latent variables for all data points with the parameters learned by EM, I implemented the function *clustersEMgeneralCase(data,k)* which assigns each point to the cluster with the highest value of responsibility. The Figure 7 shows the clusters with different colors.

Figure 7: Training data plotted with different colors for latent variables



d) The comparison of maximum log-likelihood between different models and data (train and test) is show in the following table:

Table 2: Comparison of Maximum Log-likelihood of Train data and Test data

	Maximum Log-Likelihood on Train data	**Maximum Log-Likelihood on Test data**
Isotropic covariance matrices	-4469.33	-4341.51
General covariance matrices	-3358.18	-3404.92

We can notice that the maximum log-likelihood when using general covariance matrices is larger than the one using isotropic covariance matrices, but if we compare the elapsed times while convergence of the two models (34.39 seconds on test-data with isotropic model and 48.83 seconds on test-data with general model), we find that the second one converges rapidly than the first, this was expected because the isotropic covariances has only 1 parameter to learn while in the general case we have d*(d+1)/2 with *d: the dimension of data.* Moreover, by analyzing the figures 4, 5, 6 and 7, we can say that the model of isotropic covariances isn't suitable for our data.

RHOULAM
ADIL

HOMEWORK 2 - PGM

1) Conditional independence and factorizations:

1) if $X \perp\!\!\!\perp Y | Z$ then $p(x,y|z) = p(x|z)\,p(y|z)$

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

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

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

if $p(x|y,z) = p(x|z)$:

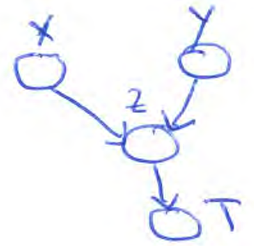
$$p(x,y|z) = \frac{p(x,y,z)}{p(z)} = \frac{p(x|y,z)\,p(y,z)}{p(z)}$$

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

then $p(x,y|z) = p(x|z)\,p(y|z)$

2) - let $p \in \mathcal{L}(G)$

then $p = p(x)\,p(y)\,p(z|x,y)\,p(T|z)$



- $X \perp\!\!\!\perp Y | T$?

Using d-separation:

- if C is the chain from X, Y

we have $Z \in C$ and (X, Z, Y) is a V-structure

and $Z \notin S$ ($= S = \{T\}$ set that is observed)

but T is a descendant of Z and it's observed

then $X \not\perp\!\!\!\perp Y | T$

3)- a) let $X \perp Y \mid Z$ ①; $X \perp\!\!\!\perp Y$ ② and $Z \in \{0,1\}$

such that $P(Z=1) = q \in [0,1]$

let $j \in \{0,1\}$:

$$P(X,Y,Z=j) = P(Z=j)P(X,Y \mid Z=j) \overset{①}{=} P(Z=j)P(X|Z=j)\,P(Y|Z=j)$$

$$= P(Z=j) \times \frac{P(X)P(Z=j|X)}{P(Z=j)} \times \frac{P(Y)P(Z=j|Y)}{P(Z=j)} \quad ③$$

From another side:

$$P(X,Y,Z=j) = P(X,Y)\,P(Z=j|X,Y) \overset{②}{=} P(X)P(Y)P(Z=j/X,Y) \quad ④$$

From ③ and ④, we deduce:

$$P(Z=j)P(Z=j/X,Y) = P(Z=j/X)P(Z=j/Y)$$

if we denote $x = P(Z=1|X)$, $y = P(Z=1|Y)$

then:

$$\begin{cases} q\,P(Z=1|X,Y) = xy \\ (1-q)\,P(Z=0|X,Y) = (1-x)(1-y) \end{cases}$$

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

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

$$= (1-q)q - q(1-x)(1-y)$$

then

$$(1-q)xy = (1-q)q - q(1-x)(1-y)$$

$$\Leftrightarrow xy - qxy = q - q^2 - q(1-x-y+xy)$$

$$\Leftrightarrow xy - qxy - q + q^2 + q - qx - qy + qxy = 0$$

$$\Leftrightarrow xy - qx - qy + q^2 = 0$$

$$\Leftrightarrow x(y-q) = q(y-q)$$

$$\Leftrightarrow (x-q)(y-q) = 0 \Leftrightarrow x-q=0 \text{ or } y-q=0$$

$$\Leftrightarrow \text{or}\begin{cases} P(Z=1/X) = P(Z=1) \\ P(Z=1/Y) = P(Z=1) \end{cases}$$

$$\Leftrightarrow \text{or}\begin{cases} P(Z=1/X) = P(Z=1) \text{ and } P(Z=0|X) = P(Z=0) \\ P(Z=1/Y) = P(Z=1) \text{ and } P(Z=0|Y) = P(Z=0) \end{cases}$$

$$\Leftrightarrow \text{or}\begin{cases} X \perp\!\!\!\perp Z \\ Y \perp\!\!\!\perp Z \end{cases}$$

b) Let (X, Y, Z) random variables in a finite space we can enumerate $Z(\Omega)$ such that:

$$Z(\Omega) = \{z_1, \dots, z_n\}$$

Let $A \subset Z(\Omega)$ $A \cup A^c = Z(\Omega)$ and $A \cap A^c = \emptyset$

(A, A^c) a partition of $Z(\Omega)$: $q_A = P(Z \in A)$ and

$1 - q_A = q_{A^c} = P(Z \in A^c)$ $\left(\begin{array}{l} z' = 0 \Leftrightarrow Z \notin A \\ z' = 1 \Leftrightarrow Z \in A \end{array}\right)$

z' is a binary variable, we will have the same result as the question before: let reprove the result.

$$P(X, Y, Z \in A) = P(X) P(Y) P(Z \in A \mid X, Y)$$

and $P(X, Y, Z \in A) = P(Z \in A) \dfrac{P(Z \in A \mid X) P(X)}{P(Z \in A)} \dfrac{P(Z \in A \mid Y) P(Y)}{P(Z \in A)}$

then $P(Z \in A \mid X, Y) P(Z \in A) = \underbrace{P(Z \in A \mid X)}_{x_A} \underbrace{P(Z \in A \mid Y)}_{y_A}$

$$\begin{cases} q_A P(Z \in A \mid X, Y) = x_A y_A \\ (1 - q_A) P(Z \in A^c \mid X, Y) = (1 - x_A)(1 - y_A) \end{cases}$$

we do calculation, we find:

$$(x_A - q_A)(y_A - q_A) = 0 \quad \forall A \in Z(\Omega)$$

then $\begin{cases} x_A = q_A \\ \text{or} \\ y_A = q_A \end{cases}$ $\forall A \in Z(\Omega)$

$\Leftrightarrow \begin{cases} P(Z \in A \mid X) = P(Z \in A) \\ \text{or} \\ P(Z \in A \mid Y) = P(Z \in A) \end{cases}$ $\forall A \in Z(\Omega)$

$\Rightarrow \begin{cases} P(Z = z_i \mid X) = P(Z = z_i) \\ \text{or} \\ P(Z = z_i \mid Y) = P(Z = z_i) \end{cases}$ $A = \{z_i\}$ $\forall i$ $\forall i \in [\![1, n]\!]$

Let $J = \{ i \in [1, m] \mid P(Z = z_i \mid X) = P(Z = z_i) \}$

Then: $\forall i \notin J$: $P(Z = z_i \mid Y) = P(Z = z_i)$

We should show that $J = \emptyset$ or $J = [1, m]$

$$\sum_{i \in J} P(Z = z_i) + \sum_{i \in J^c} P(Z = z_i) = \sum_{i \in J} P(Z = z_i \mid X) + \sum_{i \notin J} P(Z = z_i \mid Y))$$

$$1 = \sum_{i \in J} P(Z = z_i \mid X) + \sum_{i \in J^c} P(Z = z_i \mid Y))$$

$$\Rightarrow \sum_{i \in J} (P(Z = z_i \mid X) - P(Z = z_i \mid Y)) = 0$$

$$\Rightarrow \sum_{i \in J} \frac{P(Z = z_i \cap X)}{P(X)} - \frac{P(Z = z_i \cap Y))}{P(Y)} = 0$$

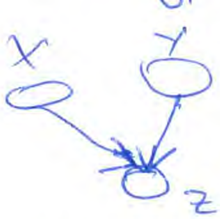
$$\Rightarrow \sum_{i \in J} P(Y) P(Z = z_j \cap X) - P(Z = z_j \cap Y)) P(X) = 0$$

$$\sum_{i \in J} P(Z = z_i \mid X) = \sum_{i \in J} P(Z = z_i \mid Y))$$

I didn't have time to think more about it but maybe by using $|A| = 2$, $|A| = 3$ ... we could find the independance, because it seems logical that $X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$. If we suppose that $X \not\perp\!\!\!\perp Z$ and $Y \not\perp\!\!\!\perp Z$ then

G

The graphs G and G' are equivalente since $L(G) = L(G')$ but they don't have the same independencies since a result shows that two graphs are Markov equivalent iff they have the same links (edges without orientations) and the same set of v structures



G'

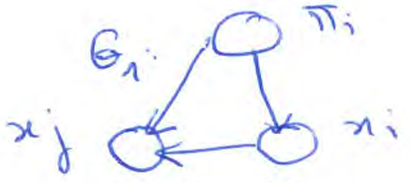
The demo is given by Pearl and Verma but
I didn't find it in the web.

2 - Distributions factorizing in a graph:

1) $G = (V, E)$ a DAG ; $i \to j \in E$ is a covered edge (i.e $\pi_j = \pi_i \cup \{i\}$).

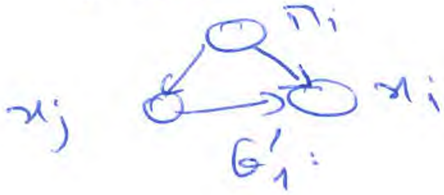
$G' = (V, E')$ with $E' = E \setminus \{i \to j\} \cup \{j \to i\}$

let show $L(G) = L(G')$.



first $\pi_j = \pi_i \cup \{i\} \Rightarrow$

$E' = E \setminus \{i \to j\} \cup \{j \to i\} \Rightarrow$



the first graph is in G

and the second in G'.

to show that $L(G) = L(G')$

it's sufficient to show that $L(G_1) = L(G_1')$

since the edge that relates i and j is the only one that changed orientation

let $p \in L(G_1)$. $p = p(\pi_i)\, p(x_j | \pi_i, x_i)\, p(x_i | \pi_i)$

$$p(x_j | \pi_j) = p(x_j | x_i, \pi_i) = \frac{p(x_j, x_i | \pi_i)}{p(x_i | \pi_i)}$$

$$= \frac{p(x_i | x_j, \pi_i)\, p(x_j | \pi_i)}{p(x_i | \pi_i)}$$

then $p(x_j | x_i, \pi_i)\, p(x_i | \pi_i) = p(x_i | x_j, \pi_i)\, p(x_j | \pi_i)$

and: $p = p(\pi_i)\, p(x_i | x_j, \pi_i)\, p(x_j | \pi_i) \in L(G_1')$

By equivalence $L(G_1) = L(G_1')$

then $L(G) = L(G')$.

2) Let G be a directed tree and G' it's corresponding undirected tree

First let show that the maximal cliques in G' are composed of only two nodes if $|V| \geq 2$

Let suppose that there exists a clique of 3 nodes A, B, C in G' then it takes the form

then since G' is the undirected tree corresponding to G, there exists an orientation of this clique that is in G.

We can see that all the $3^2 = 9$ possible orientations gives a V-structure; but the directed tree G doesn't contains V-structures then it's an absurd. then: $|Cliques_{max}(G')| \leq 2$

- if $G' = (\{root\}, \phi)$ then $cliques(G') = 1$.

and $L(G') = L(G)$ obviously.

- if $|V| \geq 2$ then there exists a node that have a parent then $|cliques_{max}(G')| \geq 2$.

finally $|cliques_{max}(G')| = 2$

without loss of generality let $(x_1 \rightarrow x_n)$ be the nodes of G in a chronological order:

$p \in L(G) \Leftrightarrow p = \prod_{i=1}^{n} p(x_i | \pi_i)$

$\pi_i \in \{x_1 \rightarrow x_{i-1}\}$ and $|\pi_i| = 1$ (Cardinal of the sets of parents of x_i is equal to one by the definition of a tree).

- for G' the only cliques that exists are (π_i, x_i) $\forall i \in [\![1, n]\!]$

so $p \in L(G') \iff p = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$

$$= \frac{1}{Z} \prod_{i=1}^{n} \psi_{C_i}(x_{C_i})$$

with: $C_i = \{\{i\} \cup \{\pi_i\}\}$

by identifying $\psi_{C_i}(x_{C_i}) = p(x_i | \pi_i)$

$$Z = \sum_{x_1 \dots x_n} \prod_{i=1}^{n} \psi_{C_i}(x_{C_i})$$

$$= \sum_{x_1 \dots x_n} \prod_{i=1}^{n} p(x_i | \pi_i)$$

$$= \prod_{i=1}^{n} \sum_{x_i} p(x_i | \pi_i) = 1.$$

we find that: $p = \prod_{i=1}^{n} p(x_i | \pi_i) \in L(G)$

then $L(G) = L(G')$ by equivalence.

3- Entropy and Mutual Information:

1) X a discrete RV with $|X| = k$

the entropy: $H(X) = -\sum_{x \in X} P_X(x) \log P_X(x)$

with $P_X(x) = \mathbb{P}(X = x)$

$$D(p\|q) = \begin{cases} +\infty & \text{if } \exists x \in \mathcal{X}, \text{ st } q(x) = 0 \text{ and } p(x) \neq 0 \\ \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} & \text{otherwise,} \end{cases}$$

a) let consider $g : [0, +\infty[\longrightarrow \mathbb{R}$

$x \longmapsto -x \log(x)$ $[0^+ \log 0^+ = 0]$

$g'(x) = -\log(x) + 1 \Rightarrow g''(x) = -\frac{1}{x} < 0 \quad \forall x \in]0, +\infty[$

then g is concave.

$H(X) = -\sum_{x \in X} P_X(x) \log P_X(x) = +\sum_{x \in X} g(P_X(x))$

$\geq \sum_{x \in X} P_X(x) g(1) = 0$ because $g(1) = 0$

Jensen inequality

$H(X) = 0 \Rightarrow -\sum_{x \in X} P_X(x) \log P_X(x) = 0$

$P_X(x) \log P_X(x) \geq 0 \quad \forall x \in X$ because $P_X(x) \in [0, 1]$

then $P_X(x) \log P_X(x) = 0 \quad \forall x \in X$

$\Rightarrow P_X(x) = 0$ or $P_X(x) = 1 \quad \forall x \in \mathcal{X}$

and because $\sum_{x \in X} P_X(x) = 1$

then $\exists! x_0 \in \mathcal{X}$ st $P_X(x_0) = 1$ and $\forall x \neq x_0$ $P_X(x) = 0$

$\Leftarrow$ Reciprocally; if $\exists! x_0 \in X$ st $P_X(x_0) = 1$ and $P_X(x) = 0 \quad \forall x \neq x_0$

then $H(X) = -\sum_{x \in X} P_X(x) \log P_X(x) = 0$

finally, the equality holds if and only if X is a constant with probability 1.

b) let $p_X \sim X$ and $q \sim U([\![1,k]\!])$

$$D(p_X \| q) = \sum_{x \in \mathcal{X}} p_X(x) \log\left(\frac{p_X(x)}{1/k}\right)$$
$$= -H(X) + \sum_{x \in \mathcal{X}} p_X(x) \log(k)$$
$$= -H(X) + \log(k)$$

$$\boxed{D(p_X \| q) = -H(X) + \log(k)}$$

c) Using the fact that $D(p_X \| q) \geq 0$, we deduce that $H(X) \leq \log(k)$

2) (X_1, X_2) two RVs defined over the finite set $\mathcal{X}_1 \times \mathcal{X}_2$. $P_{1,2}$, P_1 and P_2 denote respectively the joint distribution of (X_1, X_2), the marginals of X_1, X_2.

$$I(X_1, X_2) = \sum_{(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} P_{1,2}(x_1,x_2) \log\left(\frac{P_{1,2}(x_1,x_2)}{P_1(x_1) P_2(x_2)}\right)$$

a) $$I(X_1, X_2) = -\sum_{(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} P_{1,2}(x_1,x_2) \log \frac{P_1(x_1) P_2(x_2)}{P_{1,2}(x_1,x_2)}$$
$$= +\sum_{(x_1,x_2) \in \mathcal{X}_1 \times \mathcal{X}_2} P_{1,2}(x_1,x_2)\, f\left(\frac{P_1(x_1) P_2(x_2)}{P_{1,2}(x_1,x_2)}\right)$$

with $f : \mathbb{R}_*^{+} \longrightarrow \mathbb{R}$, $x \longmapsto -\log(x)$; $f''(x) = \frac{1}{x^2} > 0$

then f is convex and using Jensen Inequality

$$I(X_1, X_2) \geq f\left(\sum_{(x_1,x_2) \mathcal{X}_1 \times \mathcal{X}_2} P_{1,2}(x_1,x_2) \frac{P_1(x_1) P_2(x_2)}{P_{1,2}(x_1,x_2)}\right)$$
$$= f\left(\sum_{x_1 \in \mathcal{X}_1} P_1(x_1) \cdot \sum_{x_2 \in \mathcal{X}_2} P_2(x_2)\right) = f(1 \times 1) = 0$$

then $I(X_1, X_2) \geq 0$

b)

$$I(X_1,X_2) = \sum_{(x_1,x_2)\in X_1\times X_2} P_{1,2}(x_1,x_2)\log(P_{1,2}(x_1,x_2)) - \sum_{(x_1,x_2)\in X_1\times X_2} P_{1,2}(x_1,x_2)\log(P_1(x_1)) - \sum_{(x_1,x_2)\in X_1\times X_2} P_{1,2}(x_1,x_2)\log(P_2(x_2))$$

$$\sum_{(x_1,x_2)\in X_1\times X_2} P_{1,2}(x_1,x_2)\log(P_1(x_1)) = \sum_{x_1\in X_1}\left(\sum_{x_2\in X_2} P_{1,2}(x_1,x_2)\right)\log(P_1(x_1))$$
$$= \sum_{x_1\in X_1} P_1(x_1)\log(P_1(x_1)) = -H(X_1)$$

the same for the last term

then $\boxed{I(X_1,X_2) = -H(X_1,X_2) + H(X_1) + H(X_2)}$

c) we have $I(X_1,X_2) \geq 0$ then

$$H(X_1,X_2) \leq H(X_1) + H(X_2)$$

$H(X_1,X_2)$ is maximal when $I(X_1,X_2) = 0$ which means $H(X_1,X_2) = H(X_1) + H(X_2)$

$$I(X_1,X_2) = 0 \iff \sum_{(x_1,x_2)\in X_1\times X_2} P_{1,2}(x_1,x_2) f\left(\frac{P_1(x_1)P_2(x_2)}{P_{1,2}(x_1,x_2)}\right) =$$
$$\sum_{(x_1,x_2)\in X_1\times X_2} f\left(P_{1,2}(x_1,x_2)\frac{P_1(x_1)P_2(x_2)}{P_{1,2}(x_1,x_2)}\right)$$ (Equality holds in the Jensen inequality)

we have:

$I(X_1,X_2) = D(P_{1,2} \| P_1P_2)$

with D is the Kullback-Leibler over $X_1\times X_2$

$D(P_{1,2} \| P_1P_2) = 0 \iff P_{1,2} = P_1P_2$

then the joint distribution of maximal entropy is given by $P_{1,2} = P_1P_2$ which means X_1 and X_2 are independent

4) Gaussian Mixtures:

b) Let $D = (x_1 \dots x_N)$ be N iid observations of x that have the distribution $p(x|z_k) = N(x|\mu_k, \Sigma_k)$ with $p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$ ($\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0 \ \forall k \in [1, K]$)

$$p(x) = \sum_{k=1}^{K} p(x|z_k=1)\, p(z_k) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

$$\log p(D|\pi,\Sigma,\mu) = \log \prod_{i=1}^{N} p(x_i|\pi,\Sigma,\mu)$$

$$\boxed{\log p(D|\pi,\Sigma,\mu) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k) \right)}$$

If we denote $\gamma(z_{ik}) = p(z_k=1|x_i)$ then by bayes formula we deduce that:

$$\gamma(z_{ik}) = \frac{\pi_k N(x_i|\mu_k, \Sigma_k)}{\sum_{p=1}^{K} \pi_p N(x_i|\mu_p, \Sigma_p)}$$

and considering that the covariance matrices are diagonal: $\Sigma_k = \sigma_k^2 I$; $\sigma_k^2 > 0 \ \forall k \in \{1 \dots K\}$

Calculating the M-step:

- Maximum likelihood:

  •) For μ_k:

$$\frac{\partial \log(p(D))}{\partial \mu_k} = 0 \Rightarrow \sum_{i=1}^{N} \frac{\pi_k N(x_i|\mu_k, \Sigma_k)}{\sum_{p=1}^{K} \pi_p N(x_i|\mu_p, \Sigma_p)} \sigma_k^2 (x_i - \mu_k) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \gamma(z_{ik}) \sigma_k^2 (x_i - \mu_k) = 0$$

$$\Rightarrow \boxed{\hat{\mu}_k = \frac{\sum_{i=1}^{N} \gamma(z_{ik}) x_i}{\sum_{i=1}^{N} \gamma(z_{ik})}}$$

  •) For $\sigma_k^2 = v_k$:

$$\frac{\partial \log p(D)}{\partial v_k} = \sum_{i=1}^{N} \frac{\pi_k}{\sum_{p=1}^{K} \pi_p N(x_i|\hat{\mu}_p, v_p)} \frac{\partial N(x_i|\hat{\mu}_p, v_p)}{\partial v_k} = 0$$

$$\mathcal{N}(x_i \mid \mu_k, \nu_k) = \frac{1}{(2\pi)^{D/2} \det(\nu_k I)^{1/2}} \exp\left(-\frac{1}{2\nu_k}(x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k)\right)$$

$$\det(\nu_k I) = \nu_k^D$$

then: $$\frac{\partial \mathcal{N}(x_i)}{\partial \nu_k} = -\frac{D}{2} \nu_k^{-\frac{D}{2}-1} \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2\nu_k}(x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k)\right)$$

$$+ \frac{1}{2\nu_k^2} (x_i - \hat{\mu}_k)^T (x_i - \mu_k) \, \mathcal{N}(x_i \mid \mu_k, \nu_k)$$

$$= -\frac{D}{2} \nu_k^{-1} \mathcal{N}(x_i \mid \mu_k, \nu_k) + \frac{1}{2\nu_k^2} (x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k) \, \mathcal{N}(x_i \mid \mu_k, \nu_k)$$

$$\frac{\partial \log P(D)}{\partial \nu_k} = 0 \Rightarrow \sum_{i=1}^{N} \gamma(z_{ik}) \left(-\frac{D}{2\nu_k} + \frac{1}{2\nu_k^2} (x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k)\right) = 0$$

$$\Rightarrow \boxed{\nu_k = \frac{\sum_{i=1}^{N} \gamma(z_{ik}) (x_i - \hat{\mu}_k)^T (x_i - \hat{\mu}_k)}{D \sum_{i=1}^{N} \gamma(z_{ik})}}$$

- For π_k: we should takes into account the condition that $\sum_{k=1}^{K} \pi_k = 1$

We have to maximize:

$$\log p(D \mid \mu, \pi, \nu) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1\right)$$

taking the gradient of this quantity by π_k we get:

$$\sum_{n=1}^{N} \frac{\mathcal{N}(x_i \mid \mu_k, \nu_k)}{\sum_{p=1}^{K} \pi_p \mathcal{N}(x_n \mid \mu_p, \nu_p)} + \lambda = 0$$

then multiplying by π_k and taking the sum over k, we get: $N + \lambda = 0 \Rightarrow \lambda = -N$

then $$\sum_{i=1}^{N} \frac{1}{\pi_k} \gamma(z_{ik}) - N = 0$$

$$\Rightarrow \boxed{\pi_k = \frac{\sum_{i=1}^{N} \gamma(z_{ik})}{N}}$$

π_k, μ_k have the same form as in the case of general covariance matrix.