1 Conditional independence and factorizations

1.1

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

It is very similar to the "Explaining away" graph in 3 dimensions that we saw in class. For this graph we saw that $X \perp Y$ but $X \perp Y|Z$ is wrong. Here we can prove the same thing. Let's take T = 2 * Z , X and Y being 2 independent dices throw and Z the resulting sum. We still have $X \perp Y$ but $X \perp Y|T$ is wrong because if we know T then we know Z and if we know Z then Y can't be independent of X because we have X = Z - Y.

1.2

Soit X, Y, Z telles que $X \perp\!\!\!\perp Y \mid Z$ et $X \perp\!\!\!\perp Y$

Pour tout y s.t. $p(y) \neq 0$ on a:

$$p(x) = \frac{p(x, y)}{p(y)} = \frac{1}{p(y)} \sum_z p(x|z)p(y|z)p(z) = \sum_z p(x|z)p(z|y)$$

a) Si Z is binaire:

$$P(X = x) = P(X = x|Z = 0)P(Z = 0|Y = y) + P(X = x|Z = 1)P(Z = 1|Y = y)$$

Sachant que X et Y sont définies sur des ensemple finis, notons $x_1, ..., x_n$ les valeurs possible de X tq, pour tout $i \in [1, n], P(X = x_i) > 0$

et, $y_1, ..., y_m$ les valeurs de Y tq, pour tout $j \in [1, m], P(Y = y_j) > 0$

Posons $A_0, A_1 \in R^n$ tq

$$[A_k]_i = P(X = x_i|Z = k)$$

Et $B_0, B_1 \in R^m$ tq

$$[B_k]_j = P(Z = k|Y = y_j)$$

Alors en posant la matrice M_X de taille $n \times m$ tq

$$[M_X]_{ij} = P(X = x_i)$$

c'est à dire la répétition m fois de la même colonne, on a:

$$M_X = A_0 B_0^T + A_1 B_1^T$$

où $A_k B_k^T$ sont aussi des matrices de taille $n \times m$.

Sachant que toutes les colonnes sont égales, alors en prenant les colonnes j et j' avec $j \neq j'$:

$$A_0[B_0]_j + A_1[B_1]_j = A_0[B_0]_{j'} + A_1[B_1]_{j'}$$

$$A_0([B_0]_j - [B_0]_{j'}) + A_1([B_1]_j - [B_1]_{j'}) = 0$$

qui est une égalité vectorielle.

Ainsi, soit (1) les deux vecteurs sont nuls, soit (2) les deux vecteurs sont colinéaires et opposés.

(1) Donc pour tout $j \neq j$, $[B_0]_j = [B_0]_{j'}$ et $[B_1]_j = [B_1]_{j'}$ donc pour $k \in \{0, 1\} : P(Z = k|Y = y_j) = P(Z = k|Y = y_{j'})$

donc $Y \perp\!\!\!\perp Z$

(2) $A_0 = \alpha A_1$ or les deux vecteurs ont pour contrainte d'être de poids 1 car $\sum_i P(X = x_i|Z = k) = 1$ donc

$$A_0 = A_1$$

i.e. pour tout $i \in [1, n]$:

$$P(X = x_i|Z = 0) = P(X = x_i|Z = 1) = P(X = x_i)$$

donc $X \perp\!\!\!\perp Z$

b) Si Z n'est pas binaire, alors on ontient une equation du type :

$$M_X = \sum_k A_k B_k^T$$

et on voit bien que l'argument de la colinéarité ne va pas fonctionner. Dur à montrer formellement cependant.

Une autre méthode est de prendre X, Y tq $X \perp\!\!\!\perp Y$, et $Z = (X, Y)$. Alors on a bien aussi $X \perp\!\!\!\perp Y \mid Z$ car si l'on connait Z alors X et Y sont déterminés. De plus, on a évidemment pas $X \perp\!\!\!\perp Z$ ni $Y \perp\!\!\!\perp Z$. C'est donc un contre-example.

2 Distribution factorizing in graphs

2.1

If $p(x) \in \mathcal{L}(G)$ then:

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

Let note π_k the parents of k in G and $\pi_k^{'}$ the parents of k in $G^{'}$

$$p(x) = (\prod_{k\neq i, k\neq j} p(x_k|x_{\pi_k^{'}}))p(x_i|x_{\pi_i})p(x_j|x_{\pi_j})$$

and

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

$$= p(x_i|x_{\pi_i}, x_j)p(x_j|x_{\pi_i}) = p(x_i|x_{\pi_i^{'}})p(x_j|x_{\pi_j^{'}})$$

Thus

$$p(x) = \prod_k p(x_k|x_{\pi_k^{'}})$$

So $p(x) \in \mathcal{L}(G^{'})$

3 Implementation - Gaussian mixtures

a. We note that for each iteration of the algorithm we get a different result. The distortion measures are roughly the same (around 1100) but there are still some differences. The same remark goes for the centers.

b. As seen in the class notes, for the M step we need to maximize

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

here we make the assumption that $\Sigma_j = \sigma_j^2 I$ We can inject it in the equation above, which gives :

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

$\Leftrightarrow$

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\tau_i^j log(\pi_{j,t}) + \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_i^j \left[log(\frac{1}{(2\pi)^{1/k}}) - d * log(\sigma_{j,t}) - \frac{1}{2\sigma_{j,t}^2}||x_i - \mu_{j,t}||_2^2\right]$$

We note that besides $\Sigma_{j,t}$, the rest of the parameters aren't impacted. Their value are thus :

$$\pi_{j,t+1} = \frac{1}{n}\sum_{i=1}^{n}\tau_i^j$$

$$\mu_{j,t+1} = \frac{\sum_{i=1}^{n}\tau_i^j x_i}{\sum_{i=1}^{n}\tau_i^j}$$

As for $\sigma_{j,t+1}$, we want to maximize the previous equation with respect to $\sigma_{j,t}$ which leads to the system :

$$\sum_{i=1}^{n}\tau_i^j\left(-\frac{d}{\sigma_{j,t+1}} + \frac{1}{\sigma_{j,t+1}^3}||x_i - \mu_{j,t+1}||_2^2\right) = 0$$

$\Leftrightarrow$

$$\sigma_{j,t+1} = \frac{1}{d}\frac{\sum_{i=1}^{n}\tau_i^j||x_i - \mu_{j,t+1}||_2^2}{\sum_{i=1}^{n}\tau_i^j}$$

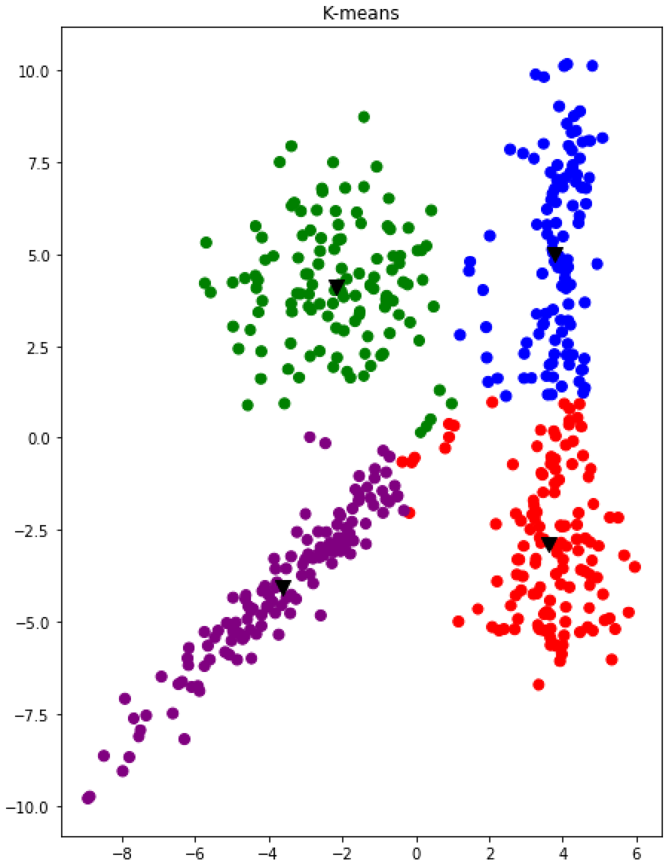
c. The formula for Σ is the same as the general one we have in the class' notes, i.e.

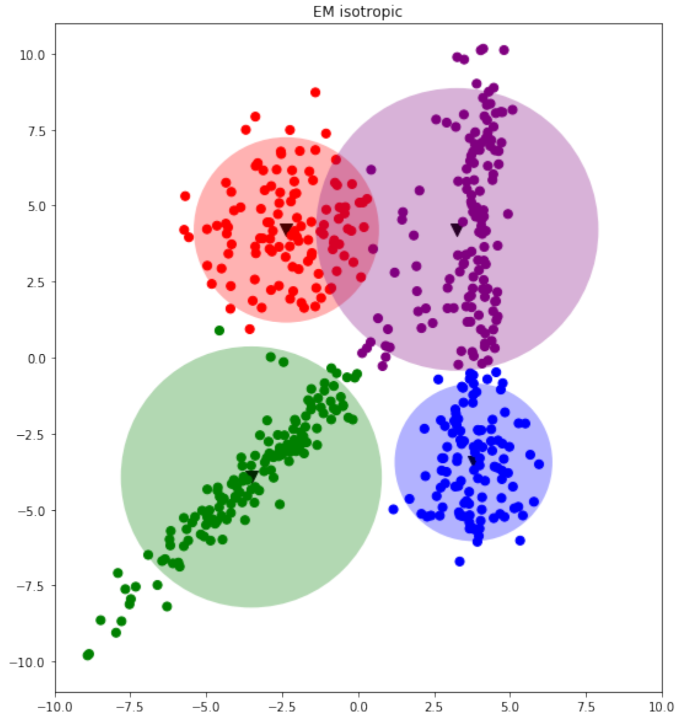
$$\Sigma_{j,t+1} = \frac{\sum_{i=1}^{n}\tau_i^j(x_i - \mu_{j,t+1})(x_i - \mu_{j,t+1})^T}{\sum_{i=1}^{n}\tau_i^j}$$

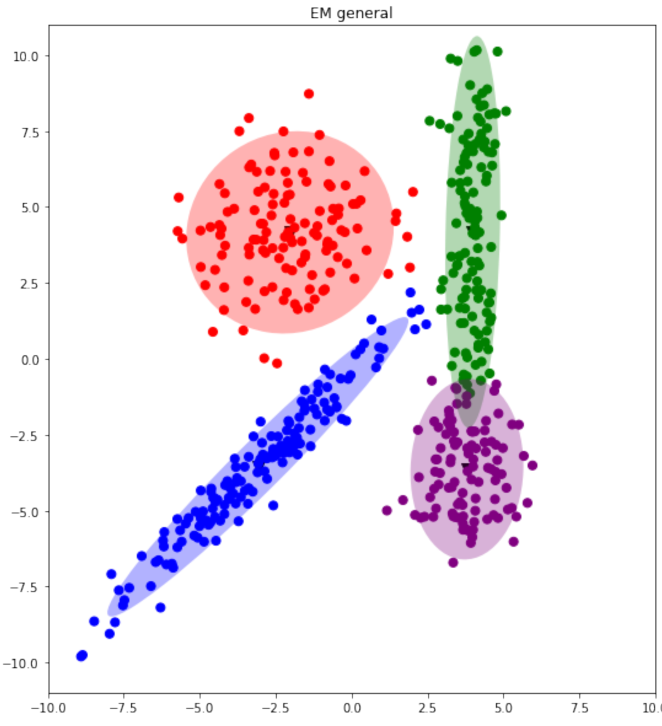
d. See code for comments on the results

Table 1: Table of log-likelihoods (average)

Data Type	Isotropic	General
train	- 5.35	- 4.69
test	- 5.30	- 4.85


K-means
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−8
−6
−4
−2
0
2
4
6


EM isotropic
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


EM general
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
-7.5
-10.0
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
10.0