Graphical models - HWK2

Abdessamad Ed-dahmouni

November 2018

1 Conditional independence and factorizations

1.1 In general, we have:

$$p(x) \in \mathcal{L}(G) \Leftrightarrow X_i \perp\!\!\!\perp X_{\mathrm{nd}(i)\setminus\pi_i} | X_{\pi_i}$$

This yields:

$$p(x) \in \mathcal{L}(G) \Leftrightarrow \begin{cases} X \perp\!\!\!\perp Y \\ T \perp\!\!\!\perp (X,Y)|Z \end{cases}$$

$X \perp\!\!\!\perp Y|T$ is not true in general:
We consider X and Y two independent variables with values in $\{-1,1\}$ with equal probability. We define $T \coloneqq Z \coloneqq XY$, we see that once T is known, X and Y are perfectly correlated:
$Y = TX$. and so $X \not\perp\!\!\!\perp Y|T$.

1.2.(a) If Z is binary, we can express $p(x,y)$ as:

$$\begin{aligned} p(x,y) &= \sum_{z\in\{0,1\}} p(x|z)p(y|z)p(z) \\ &= p(x)p(y) \sum_{z\in\{0,1\}} \frac{p(z|x)p(z|y)}{p(z)} \end{aligned}$$

Which by using $\sum_z p(z|x) = \sum_z p(z|y) = \sum_z p(z) = 1$ provides:

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

And using $p(0|x) - p(0) = p(1) - p(1|x)$ (similar formula for y), we get $(p(0|x)-p(0))(p(0|y)-p(0)) = 0$
Which means $X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$ (because Z is binary, and if A and B are two independent events, $\overline{A}$ and B are independent as well).

1.2.(b) We'll construct a counter-example. We fix three finite sets A, B, C where X,Y and Z take values respectively. We also define $n_A \coloneqq |A|$, $n_B \coloneqq |B|$ and $n \coloneqq |C|$. We consider the case $n_A, n_B > 1$, and $n > (2^{n_A}2^{n_B} + n_A - 1)/(n_A - 1)$
We can choose Y s.t. $Y \not\perp\!\!\!\perp Z$, and X such that $X \perp\!\!\!\perp Y|Z$. We then look for X satisfying $X \perp\!\!\!\perp Y$, and $X \not\perp\!\!\!\perp Z$:

$$\begin{aligned} X \perp\!\!\!\perp Y \Leftrightarrow & \forall G \subset A, H \subset B, \quad \sum_z (p(z)p(H|z)) \sum_{x\in G} p(x|z) = p(H) \sum_z p(z) \sum_{x \in G} p(x|z) \\ \Leftrightarrow & \exists M \in \mathbb{R}_+^{n_A \times n} \begin{cases} \forall G \subset A, H \subset B, \quad \sum_z p(z)(p(H|z) - p(H)) \sum_{x\in G} M_{x,z} = 0 \\ \forall z \in C, \quad \sum_x M_{x,z} = 1 \\ \forall x, z, \quad p(X = x|Z = z) = M_{x,z} \end{cases} \end{aligned}$$

This is a system of $2^{n_A}2^{n_B} + n$ linear equations. We know that this system is consistent (take M from $U \coloneqq \{M_{x,z} = f(x) > 0$, with $\sum_x f(x) = 1\}$, which corresponds to $X \perp\!\!\!\perp Z$), we choose an element $M_1 \in U$.
We also know that the dimension of the set of solutions of the corresponding homogeneous system is at least $dim(S) \geq n_A n - 2^{n_A}2^{n_B} - n > 0$ and by removing the cases $V = S \cap \{M_{x,z} = f(x)\}$ we remove $n_A - 1$ dimensions. Which leaves us with $n_A n - 2^{n_A}2^{n_B} - n - n_A + 1 > 0$. This means that we can find a solution $M_2 \neq 0, M_2 \notin V$. We construct a positive solution by choosing $M = M_1 + \epsilon M_2$, with $\epsilon \neq 0$ small.
M corresponds to a distribution of $X|Z$, where X satisfies $X \perp\!\!\!\perp Y|Z$, $X \perp\!\!\!\perp Y$, and $X \not\perp\!\!\!\perp Z$.

2 Distributions factorizing in a graph

2.1 This is a direct application of Bayes formula:

$$\begin{aligned}
p(x) \in \mathcal{L}(G) \Leftrightarrow & \forall x, \quad p(x) = \prod_{k=1}^{n} p(x_k | x_{\pi_k}) \\
\Leftrightarrow & \forall x, \quad p(x) = p(x_i | x_{\pi_i}) p(x_j | x_{\pi_i}, x_i) \prod_{k \in \{1,n\} \backslash \{i,j\}} p(x_k | x_{\pi_k})
\end{aligned}$$

And $p(x_i|x_{\pi_i})p(x_j|x_{\pi_i}, x_i) = p(x_j|x_{\pi_i})p(x_i|x_{\pi_i}, x_j) = p(x_i, x_j|x_{\pi_i})$ (by applying Bayes formula to $\widetilde{p}(z) = p(z|x_{\pi_i})$)
This proves:

$$\begin{aligned}
p(x) \in \mathcal{L}(G) \Leftrightarrow & \forall x, \quad p(x) = p(x_j | x_{\pi_i}) p(x_i | x_{\pi_i}, x_j) \prod_{k \in \{1,n\} \backslash \{i,j\}} p(x_k | x_{\pi_k}) \\
\Leftrightarrow & \forall x, \quad p(x) = \prod_{k=1}^{n} p(x_k | x_{\pi_k(G')}) \\
\Leftrightarrow & p(x) \in \mathcal{L}(G')
\end{aligned}$$

2.2 A directed tree has no v-structure, which means that no node has more than one parent.
In G', there are no cliques with more than 2 nodes: we consider the case where a node v has two neighbors or more, and we select two of them. They are either:
1. children of v in G, in which case they are not neighbors in G because it leads to a v-structure;
2. or one is a parent of v and one is a child, in which case they are not neighbors in G because it would create a v-structure or a 3-node cycle.
This proves that the cliques of G' are sets of two elements (parent-child pairs in G') or one element. We consider $1 \to n$ a topological order on G. This way the root of the tree is the node 1. and we can write:

$$\begin{aligned}
p(x) \in \mathcal{L}(G') \Leftrightarrow & \forall x, \quad p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \\
\Leftrightarrow & \forall x, \quad p(x) = \frac{1}{Z} \prod_{i=2}^{n} \psi_{i,\pi_i}(x_i, x_{\pi_i}) \prod_{i=1}^{n} \psi_i(x_i) \\
\Leftrightarrow & \forall x, \quad p(x) = \frac{1}{Z} \prod_{i=1}^{n} \phi_i(x_i, x_{\pi_i}) \qquad (*)
\end{aligned}$$

Where $\phi_i(x_i, x_{\pi_i}) = \psi_{i,\pi_i}(x_i, x_{\pi_i})\psi_i(x_i)$ for $i \in \{2, \dots, n\}$ and $\phi_1(x_1, x_{\pi_1}) = \psi_1(x_1)$
To meet the conditions $\forall x_{\pi_i}, \quad \sum_{x_i} \phi_i(x_i, x_{\pi_i}) = 1$, we define:

$$G_0 \coloneqq G, \quad \phi_i^0 = \phi_i, \quad h_i^0(x_{\pi_i}) \coloneqq \sum_{x_i} \phi_i^0(x_i, x_{\pi_i})$$

We proceed recursively starting with $G_0 = G$. At step k, we do:
• For each leaf i of G_k, we update ϕ_i^k to $\phi_i^{k+1}(x_i, x_{\pi_i}) = \frac{\phi_i^k(x_i,x_{\pi_i})}{h_i^k(x_{\pi_i})}$, where $h_i^k(x_{\pi_i}) \coloneqq \sum_{x_i} \phi_i^k(x_i, x_{\pi_i})$
• For each interior node of G_k, we update ϕ_i^k to $\phi_i^{k+1}(x_i, x_{\pi_i}) = \phi_i^k(x_i, x_{\pi_i}) \prod_{c \in leaves(G_k), \pi_c = i} h_c^k(x_i)$
• For each node i in G but not in G_k, $\phi_i^{k+1} = \phi_i^k$.
• We repeat for $G_{k+1} = (V_{k+1}, E_{k+1}) = (V_k \backslash leaves(V_k), E_k \backslash \{(i, \pi_i) \text{ for } i \in leaves(G_k)\})$ till we reach $G_k = (\{1\}, \emptyset)$ included (with 1 considered as a leaf and not an interior point at the end, and $h_1^S = \sum_{x_1} \phi_1^S(x_1)$)
At each step, we have:

$$\forall i \in G_{k+1}, \quad \pi_i(G_{k+1}) = \pi_i(G_k)$$
$$\forall i \in G_k \backslash G_{k+1}, \quad \sum_{x_i} \phi_i^{k+1}(x_i, x_{\pi_i(G_k)}) = 1$$

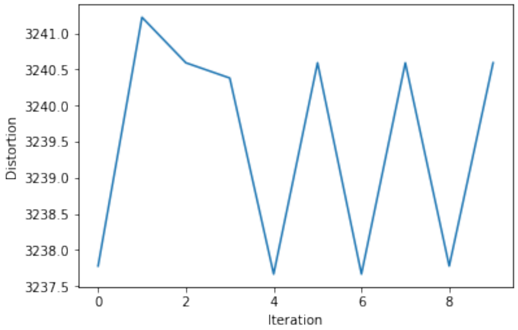
By the end of this algorithm (which necessarily terminates since at each step $|V_{k+1}| < |V_k|$), say at step S, we have:

$$p(x) = \frac{h_1^S}{Z} \prod_{i=1}^{n} \phi_i^S(x_i, x_{\pi_i})$$

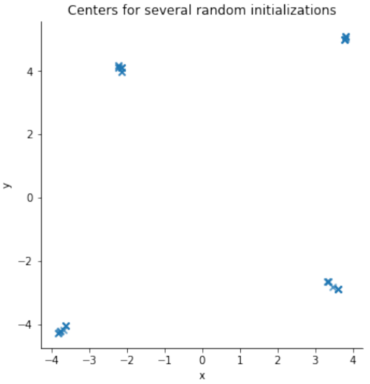
With $\forall x_{\pi_i}, \quad \sum_{x_i} \phi_i^S(x_i, x_{\pi_i}) = 1$. $\sum_x p(x) = 1$ leads to $h_1^S = Z$ and this proves $\mathcal{L}(G') \subset \mathcal{L}(G)$. $\mathcal{L}(G) \subset \mathcal{L}(G')$ is easy to see with the equivalences mentioned in (*).

3 Implementation - Gaussian mixtures

3.(a) Centers: (8 iterations for convergence): $\hat{\mu}_1 = \begin{bmatrix} -3.80 \\ -4.25 \end{bmatrix}, \quad \hat{\mu}_2 = \begin{bmatrix} 3.36 \\ -2.66 \end{bmatrix}, \quad \hat{\mu}_3 = \begin{bmatrix} -3.80 \\ 5.10 \end{bmatrix}, \quad \hat{\mu}_4 = \begin{bmatrix} -2.24 \\ 4.13 \end{bmatrix}$



(a) Distortion for different random initializations



(b) Centers for different random initializations

1.(b) By maximizing the expectation of the log likelihood, we find for $j \in \{1, \dots, K\}$:

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \tau_i^j$$

$$\hat{\mu}_j = \frac{1}{\sum_{i=1}^{n} \tau_i^j} \sum_{i=1}^{n} \tau_i^j x_i$$

$$\hat{\sigma}_j^2 = \frac{1}{2 \sum_{i=1}^{n} \tau_i^j} \sum_{i=1}^{n} \tau_i^j ||x_i - \hat{\mu}_j||^2$$

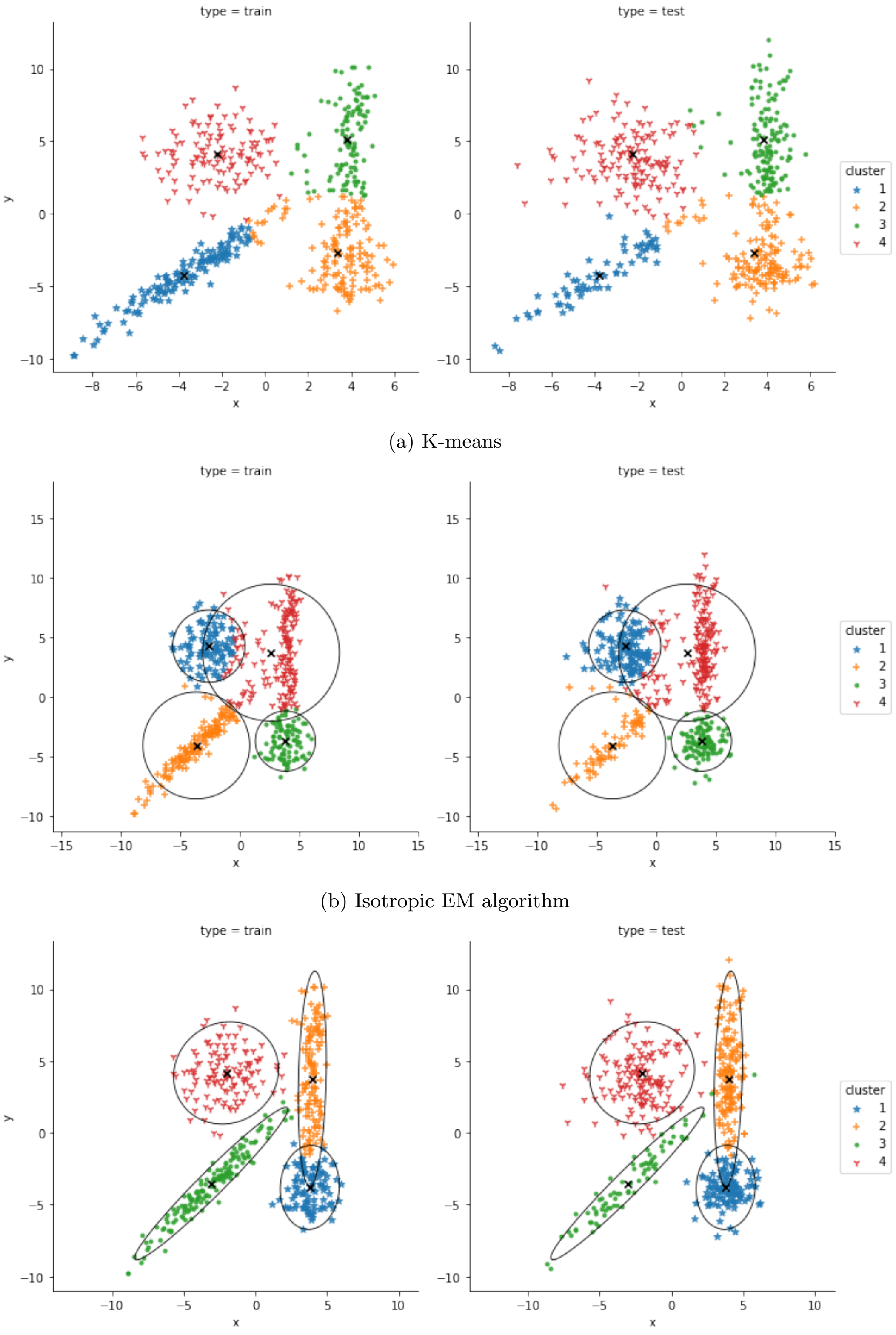
Estimates for the isotropic EM:
$\hat{\pi}_1 = 0.17, \quad \hat{\pi}_2 = 0.27, \quad \hat{\pi}_3 = 0.20, \quad \hat{\pi}_4 = 0.37$
$\hat{\mu}_1 = \begin{bmatrix} -2.61 \\ 4.25 \end{bmatrix}, \quad \hat{\mu}_2 = \begin{bmatrix} -3.66 \\ -4.08 \end{bmatrix}, \quad \hat{\mu}_3 = \begin{bmatrix} 3.82 \\ -3.72 \end{bmatrix}, \quad \hat{\mu}_4 = \begin{bmatrix} 2.61 \\ 3.70 \end{bmatrix}$
$\hat{\sigma}_1^2 = 2.00, \quad \hat{\sigma}_2^2 = 4.36, \quad \hat{\sigma}_3^2 = 1.39, \quad \hat{\sigma}_4^2 = 7.17$

3.(c) Estimates for the general EM:
$\hat{\pi}_1 = 0.18, \quad \hat{\pi}_2 = 0.26, \quad \hat{\pi}_3 = 0.31, \quad \hat{\pi}_4 = 0.25$
$\hat{\mu}_1 = \begin{bmatrix} 3.80 \\ -3.80 \end{bmatrix}, \quad \hat{\mu}_2 = \begin{bmatrix} 3.98 \\ 3.77 \end{bmatrix}, \quad \hat{\mu}_3 = \begin{bmatrix} -3.06 \\ -3.53 \end{bmatrix}, \quad \hat{\mu}_4 = \begin{bmatrix} -2.03 \\ 4.17 \end{bmatrix}$
$\hat{\Sigma}_1 = \begin{bmatrix} 0.92 & 0.06 \\ 0.06 & 1.87 \end{bmatrix}, \quad \hat{\Sigma}_2 = \begin{bmatrix} 0.21 & 0.29 \\ 0.29 & 12.24 \end{bmatrix}, \quad \hat{\Sigma}_3 = \begin{bmatrix} 6.24 & 6.05 \\ 6.05 & 6.18 \end{bmatrix}, \quad \hat{\Sigma}_4 = \begin{bmatrix} 2.90 & 0.21 \\ 0.21 & 2.76 \end{bmatrix}$

3.(d) We notice that the 4th cluster (red) fitted by isotropic Gaussian distributions to the data crosses to other clusters. Also, some clusters are skewed in one direction, which should be a sign for adopting a general GM model instead of an isotropic one.
The values of log likelihoods confirm this:

Method	Training set	Test set
Isotropic EM	-2639.569	-2614.603
General EM	-2327.716	-2408.978

Overall, with general EM we get a higher likelihood on both the training and test set (I noticed that the log likelihood decreases for general EM on the test set, but increases for isotropic EM, this could be a sign of overfitting).



(a) K-means

(b) Isotropic EM algorithm

(c) General EM algorithm

Figure 2: Clusters of training and test sets and regions of 90% of density for each cluster