Probabilistic Graphical Models

Assignment 2

Bastien PONCHON
bastien.ponchon@ens-paris-saclay.fr

1 Conditional independence and factorizations

1.1

We want to prove that

$$X \perp\!\!\!\perp Y|Z \Leftrightarrow \forall (y,z)|p(y,z) > 0,\ p(x|y,z) = p(x|z)$$

Let us assume that $X \perp\!\!\!\perp Y|Z \Leftrightarrow$. Be (x,y,z) such that $p(y,z) > 0$, this implies that $p(y|z)p(z) > 0$, and therefore $p(y|z) > 0$ and $p(z) > 0$. We then have:

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

Let us now assume that $\forall (y,z)|p(y,z) > 0,\ p(x|y,z) = p(x|z)$. We then have, for all (x,y,z) such that $p(y,z) > 0$:

$$\begin{aligned} p(x|y,z) &= p(x|y,z)p(y|z) \\ &= p(x|z)p(y|z) \end{aligned}$$

An hence, $X \perp\!\!\!\perp Y|Z$.

1.2

$$p \in \mathcal{L}(G) \Leftrightarrow p(x,y,z,t) = p(x)p(y)p(z|x,y)p(t|z)$$

And, for any $p \in \mathcal{L}(G)$ we do not necessarily have $X \perp\!\!\!\perp Y|T$. Indeed, X and Y are not seperated by T as (X,Z,Y) is a V-structure and T is a descendant of Z.

1.3

2 Distributions factorizing in a graph

2.1

We have:

$$
\begin{aligned}
p(x_j|x_{\pi_j})p(x_i|x_{\pi_i}) =& p(x_j|x_{\{i\}\cup\pi_i})p(x_i|x_{\pi_i}) \\
&= p(x_j|x_i, x_{\pi_i})p(x_i|x_{\pi_i}) \\
&= \frac{p(x_i|x_j, x_{\pi_i})p(x_j|x_{\pi_i})}{p(x_i|x_{\pi_i})}p(x_i|x_{\pi_i}) \\
&= p(x_i|x_{\{j\}\cup\pi_i})p(x_j|x_{\pi_j\setminus\{i\}}) \\
&= p(x_i|x_{\pi'_i})p(x_j|x_{\pi'_j})
\end{aligned}
$$

Therefore, as the other edges (and thus parents) remain unchanged between the two graphs, we can say that any p that factorizes in one, factorizes in the other. Both graphs are thus Markov equivalent.

2.2

Let $p \in \mathcal{L}(G)$, then $p(x) = \prod_i p(x_i|x_{\pi_i})$. We assume the vertices are in topological order, we will also arbitrarily order the set $\mathcal{C}$ of cliques of the undirected graph G'. We then construct the potential of the cliques of G' in an iterative way, by taking them in the order, and defining the potential of the clique C_k by:

$$
\phi_{C_k}(x_c) = \prod_{x_i \in C_k, x_i \notin C_{(j<k)}} p(x_i|x_{\pi_i})
$$

We can then easily verify that $p(x) = \prod_{C\in\mathcal{C}} \phi_C(x_C)$, that $\phi_C \geq 0, \forall C \in \mathcal{C}$ (as a product of positive function). We thus have $p \in \mathcal{L}(G')$.

Now let us assume that $p \in \mathcal{L}(G')$. We can thus write $p(x) = \frac{1}{Z}\prod_{C\in\mathcal{C}} \phi_C(x_C)$ with $\phi_C \geq 0, \forall C \in \mathcal{C}$ and $Z = \sum_x \prod_{C\in\mathcal{C}} \phi_C(x_C)$.

We assumed G to be a directed tree. Each vertex can therefore have at most one parent. As all cliques in G' are sets of two vertices (otherwise there would have been a V-structure), there is as many cliques as there are edges. Each clique is thus associated to a potential function depending only of a vertex and its parent. Thus by normalizing properly, we can define conditional probabilities for all children given their parent. Therefore $p \in \mathcal{L}(G)$.

3 Entropy and Mutual Information

3.1

3.1.1

We can easily show that the entropy is greater than or equal to 0, as $\forall x \in \mathcal{X}, 1 \geq p(x) \geq 0$. This implies that $\forall x \in \mathcal{X}, -p(x)log(p(x)) \geq 0$, and thus the

entropy is positive, as a sum of positive elements.

Let us now assume that X is constant. We then have $p(x) = 1$ for a unique $x \in \mathcal{X}$ and $p(x') = 0, \forall x' \in \mathcal{X}, x' \neq x$. We then have $H(X) = 0$ because either $p(x) = 0$ or $log(p(x)) = 0$ for all $x \in \mathcal{X}$.

Reciprocally, let us assume than $H(X) = 0$. Then, as it is a sum of positive elements, we have $\forall x \in \mathcal{X}, p(x)log(p(x)) = 0$ which means that $\forall x \in \mathcal{X}, p(x) = 0$ *or* $p(x) = 1$, which implies that X is constant, because p sums to one.

3.1.2

q is the uniform distribution on $\mathcal{X}$. Therefore, $\forall x \in \mathcal{X}, q(x) = \frac{1}{|\mathcal{X}|}$.

We thus have:

$$\begin{aligned} D(p_X \parallel p) &= \sum_x p(x) log(\frac{p(x)}{q(x)}) \\ &= \sum_x p(x) log(p(x)) - \sum_x p(x) log(q(x)) \\ &= -H(x) + log(\mid \mathcal{X} \mid) \\ &\geq 0 \end{aligned}$$

He therefore have $H(X) \leq log(\mid \mathcal{X} \mid) = log(K)$.

3.2

3.2.1

$$I(X_1, X_2) = D(p1, 2 \parallel p_1 p_2) \geq 0$$

3.2.2

$$\begin{aligned} I(X_1, X_2) &= \sum_{(x_1, x_2)} p_{1,2}(x_1, x_2) log(\frac{p_{1,2}(x_1, x_2)}{p_1(x_1) p_2(x_2)} \\ &= \sum_{(x_1, x_2)} p_{1,2}(x_1, x_2) log(p_{1,2}(x_1, x_2)) - \sum_{(x_1, x_2)} p_{1,2}(x_1, x_2) log(p_1(x_1)) - \sum_{(x_1, x_2)} p_{1,2}(x_1, x_2) log(p_2(x_2)) \\ &= -H(X_1, X_2) - \sum_{x_1} log(p_1(x_1)) \sum_{x_2} p_{1,2}(x_1, x_2) - \sum_{x_2} log(p_2(x_2)) \sum_{x_1} p_{1,2}(x_1, x_2) \\ &= H(X_1) + H(X_2) - H(X_1, X_2) \end{aligned}$$

3.2.3

We have, for a given marginals p_1 and p_2:

$$I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2) \geq 0$$

So the $H(X_1) + H(X_2)$ is an upper bound of the entropy of the joint probability $p_{1,2}$ probability of maximum entropy. This upper bound is reached if and only if $I(X_1, X_2) = 0$, and thus if and only if $X_1 \perp\!\!\!\perp X_2$.

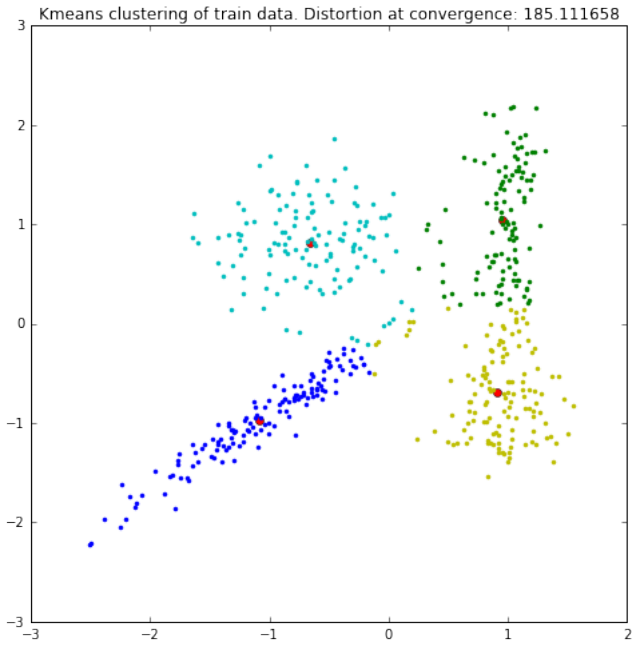
The joint probability of maximum entropy is thus defined by:

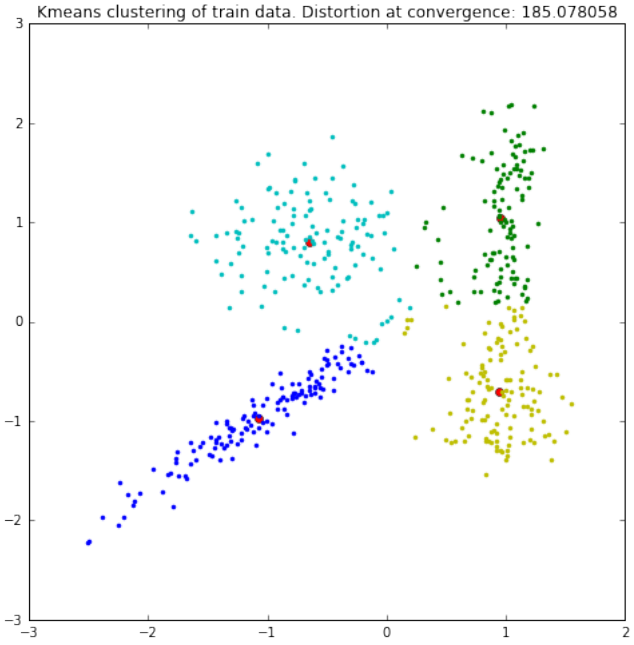
$$\forall (x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2, p_{1,2}(x_1, x_2) = p_1(x_1) p_2(x_2)$$

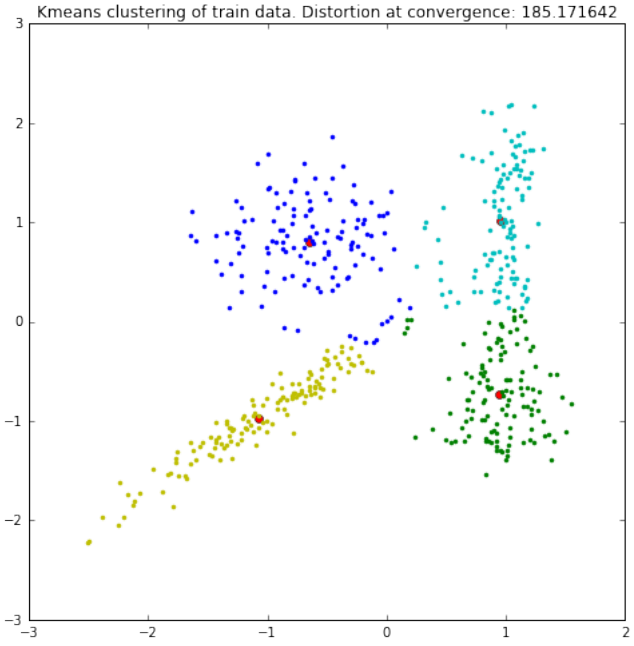
4 Implementation - Gaussian mixtures

4.1 Kmeans algorithm

We implemented the Kmeans algorithm, with random initialization of cluster centers (randomly chosen among the training points). We got the following results. Note that we do not get exactly the same distortion at convergence and points assignment each time we retrain our model.




Kmeans clustering of train data. Distortion at convergence: 185.078058
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2


Kmeans clustering of train data. Distortion at convergence: 185.171642
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2

4.2 EM algorithm on mixture of isotropic Gaussians

For mixture of isotropic gaussians ($\forall j \in [1,k], \Sigma_j = \sigma_j^2 Id$, the E-step of the algorithm remains unchanged (but the Gausian are isotropic). Let us derive again the M-step update for this model for the covariacne parameters (the others remain unchanged).

We want to maximize the complete likelihood expectation $E_{Z|X}\ (l_{c,t})$ with regards to the parameters of our gaussian mixture: π_j, μ_j and $\sigma_j, \forall j \in [1,k]$.

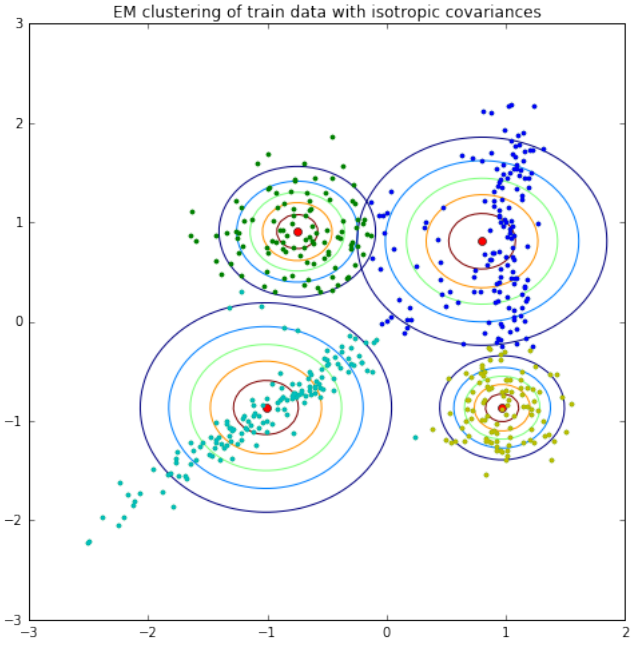
$$f(\theta_t) = E_{Z|X}\ (l_{c,t}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j log(\pi_{j,t}) + \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_i^j \left(log(\frac{1}{(2\pi)^{\frac{d}{2}}}) - 2log(\sigma_{j,t}) - \frac{1}{2}\frac{(x_i - \mu_{j,t})^T(x_i - \mu_{j,t})}{\sigma_{j,t}^2} \right)$$

$$\frac{\partial f(\theta_{t+1})}{\partial \sigma_{j,t+1}} = 0$$

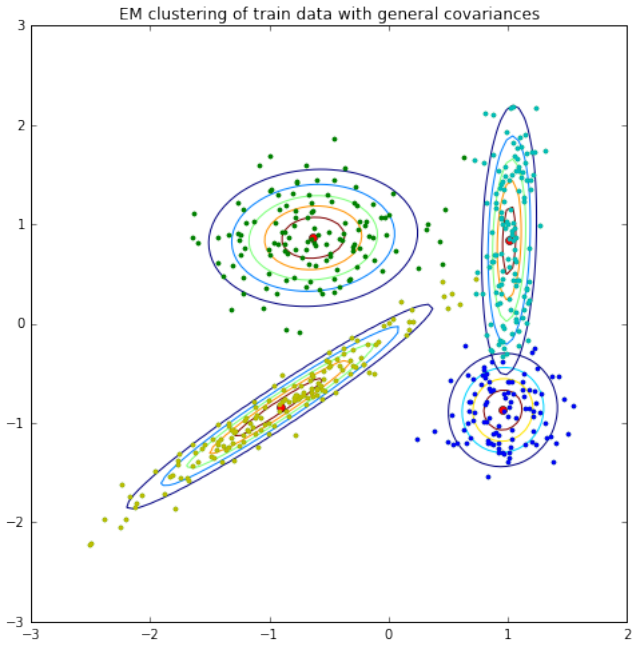
$$\Leftrightarrow \sum_i \tau_i^j \left(-\frac{2}{\sigma_j} + \frac{1}{\sigma_j^3}(x_i - \mu_{j,t})^T(x_i - \mu_{j,t}) \right) = 0$$

$$\Leftrightarrow \frac{1}{\sigma_j^2}\sum_i \tau_i^j (x_i - \mu_{j,t})^T(x_i - \mu_{j,t}) = 2\sum_i \tau_i^j$$

$$\Leftrightarrow \sigma_{j,t+1}^2 = \frac{\sum_i \tau_i^j (x_i - \mu_{j,t})^T(x_i - \mu_{j,t})}{2\sum_i \tau_i^j}$$


EM clustering of train data with isotropic covariances
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2

4.3 EM algorithm on mixture of general Gaussians



4.4 Comparing

The log-likelihood of our data given the mixutre models is:

$$\begin{aligned}
&\sum_{i=1}^{n} log\left(\sum_{j=1}^{K} p(x_i, z_i = j|\theta)\right) \\
=&\sum_{i=1}^{n} log\left(\sum_{j=1}^{K} p(x_i|z_i = j, \theta)p(z_i = j|\theta)\right) \\
=&\sum_{i=1}^{n} log\left(\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)\right)
\end{aligned}$$

Let us now compare the log-likelihood of each data set given the model and its parameters:

Model	Train Set	Test Set
Isotropic Mixture of Gaussian	-1615.98	-1707.76
General Mixture of Gaussian	-462.913	-761.764

As we could have expected, we can see that both train and test sets are more likely to have been generated by the General Mixture Model than by the Isotropic one. We can also notice than the train set is, logically enough, more likely to have been generated by the model trained on it than the test set.