**Master M2 MVA 2014 - Graphical models**

Exercises for November 5, 2014.

SOLUTIONS

# 1 Distributions factorizing in a graph

(a.1) We assume that the $p$ vertices are indexed through a topological order. If, in such an order, we have $j = i + k$, then $(1, \dots, i, j, i+1, \dots, i+k-1, i+k+1, \dots, n)$ is also a topological order since (a) $\pi_j \subset \{1, \dots, i\}$ and (b) if $\forall m > 0$, $\pi_{i+m} \subset \{1, \dots, i+m-1\}$ then $\pi_{i+m} \subset \{1, \dots, i+m-1\} \cup \{j\}$. We may thus assume $j = i + 1$

If $p \in \mathcal{L}(G)$, we thus have $p(x) = \prod_{k=1}^{p} p(x_k \mid x_{\pi_k})$ where $\pi_k$ denotes the parents of $k$ in $G$ and

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

Since $\pi'_j = \pi_i$ et $\pi'_i = \pi_j \cup \{j\}$ with $\pi'_i$ the parents of $i$ in $G'$, we have $p \in \mathcal{L}(G')$. Finally, by symmetry, $\mathcal{L}(G) = \mathcal{L}(G')$.

(a.2) If $p \in \mathcal{L}(G)$ then $p(x) = \prod_{j=1}^{n} p(x_j|x_{\pi_j})$, 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, if $p \in \mathcal{L}(G')$ with $G' = (V, E')$ we show the result by induction on the number of nodes. It is trivial $n = 1$. Then, for $n > 1$, we index the nodes from 1 to $n$ so that node $n$ is a leaf which is not the root of the directed tree $G$ and its unique neighbor is the node $n-1$. For $n > 1$, there exists such a leaf distinct from the root, and for this leaf, we have $(n-1, n) \in E$. We also have $p(x) = \frac{1}{Z} \prod_{\{i,j\} \in E'} \psi_{ij}(x_i, x_j)$.

We define

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

By induction hypothesis, $\tilde{p}$ factorises in the subgraph defined by $G$ in $\{1, \dots, n-1\}$ and we have $\tilde{p}(x_1, \dots, x_{n-1}) = \prod_{i=1}^{n-1} \tilde{p}(x_i|x_{\pi_{n-1}})$, where the order $(1, \dots, n-1)$ is not necessarily topological. Finally, we defined $f(x_n, x_{n-1})$ through

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

with $K_n$ the number of possible values for $X_n$. We then have

$$p(x) = \tilde{p}(x_1, \dots, x_{n-1}) f_n(x_n, x_{n-1}) = f_n(x_n, x_{n-1}) \prod_{i=1}^{n-1} \tilde{p}(x_i|x_{\pi_{n-1}})$$

and, since the edge $(n-1, n)$ is oriented in a way which is coherent with the fact that $\sum_{x_n} f_n(x_n, x_{n-1}) = 1$, by applytion the proposition 4.1 of class 4, we have shown $p \in \mathcal{L}(G)$.

NB : In general, oriented and non-orientrees are Markov-equivalent.

(b) The question above show that for all graph $G$ which is a tree or a union of disjoint trees, if $G'$ is an oriented version of $G$ (without v-structures), then $\mathcal{L}(G) = \mathcal{L}(G')$. If we list all graphs with at most 4 nodes and at most 4 edges, the obtained graphs are trees or union of trees, except :

— the complete graph on 3 nodes $K_3$, for which $\mathcal{L}(G)$ is the set of all distributions over 3 variables.

— the graph obtained from $K_3$ together with an isolated node, for which it is also trivial than all acyclic orentiation may be taken.

— the graph with 4 nodes corresponding to de Gibbs distributions of the form

$$p(x) = \frac{1}{Z}\psi_{123}(x_1, x_2, x_3)\psi_{34}(x_3, x_4)$$

for which it is easy by defining $f_4(x_4, x_3) = \psi_{34}(x_3, x_4)/(\sum_{x_4} \psi_{34}(x_3, x_4))$, to proceed as in the previous exercise and show that these are exactly the distributions of the form $p(x) = p(x_1, x_2, p_3)p(x_4|x_3)$.

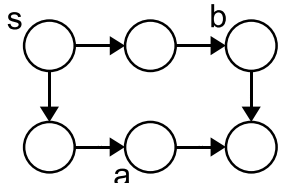
— the cycle on 4 nodes $C_4$ for which we will argue that there is no equivalent DAG. Indeed, if $(1, 2, 3, 4)$ is the cycle in this order, then $p \in \mathcal{L}(G)$ satisfies $X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4)$ and $X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3)$. However, all acyclic orientation of the graph has to create a v-structure Without loss of generality, the v-structure may be taken at node $X_2$ ; then we cannot have $X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4)$. Adding edges does not remove the v-structure and removing edges leads to strictly smaller models.

# 2 d-separation

(a) Let $a$ and $b$ be two nodes in the graph $G$ and $v = (a = v_1, \ldots, v_k = b)$ a chain in $G$ with $a = v_1$ and $v_k = b$. Consider the sequence of nodes $v'_1, \ldots, v'_{k'}$ obtained from the previous chain by removing all nodes where the chain has a v-structure. By definition, there is no v-structure at the extremities of the chain, so that $a = v'_1$ et $v'_{k'} = b$. Let us show that $(v'_1, \ldots, v'_{k'})$ forms a chain in $G_M$. Indeed, for any node $v_i$ of the original chain that has a v-structure, $v_{i-1}$ are $v_{i+1}$ its parents, and thus by moralization, they are connected in $G_M$. Since two consecutive nodes in a chain cannot both exhibit a v-structure, the vertices just before and after a node with a v-structure are preserved in the chain $v'$ and two consecutive nodes of the chain $v'$ are thus connected in $G_M$. But, if the chain $v$ is not blocked by a separator $S$, then none of the nodes that have no v-structure is included in $S$ ; in other words, no node in $v'$ is in $S$. Thus the chain $v'$ is not blocked by $S$.

Finally, if $A$ and $B$ are not d-separated by $S$ in $G$, there exist a chain which is not d-blocked by $S$ in $G$. This allows us to build non-blocked chain in $G'$ (in the usual undirected senses), and thus shows that $A$ and $B$ are not separated in $G'$.

(b) $A$ and $B$ may not be d-separated by $T$, the following graph gives a counter-example if we choose $A = \{a\}$, $B = \{b\}$, $S = \{s\}$ and $T = V \backslash \{a, b, s\}$.




(c) — We have $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid X_3$ : the chains $(1, 8, 4)$ et $(2, 8, 4)$ are the only ones connecting $\{1, 2\}$ and $\{4\}$, and they are blocked at node 8 since it is not in the separator and 3 is not one of its descendants.

— We have $X_{\{1,2\}} \not\perp\!\!\!\perp X_4 \mid X_5$ since $(1, 8, 4)$ and $(2, 8, 4)$ are not blocked in 8 since 5 is a descendant descendant of 8 which is in the separator.

— We have $X_1 \perp\!\!\!\perp X_6 | X_{\{2,4,7\}}$ since the only chain connecting 1 and 6 is $(1, 8, 7, 6)$ which is blocked in 7 since 7 is in the separator and there is no v-structure there.

# 3 Mixtures of Gaussians

(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 latent variable is estimated for each $n$ by maximizing the a posteriori probaility, i.e., through $\arg\max_i \tau_n^i$.

For a standard multivariate Gaussian, i.e., so that $\mu = 0$ et $\Sigma = I_d$, le disk corresponding to 90% of the mass is centerd in 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 $\chi^2$-distribution with $d$ degrees of freedom. In the general case, the ellipse is obtained by affine transform (see code).

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

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

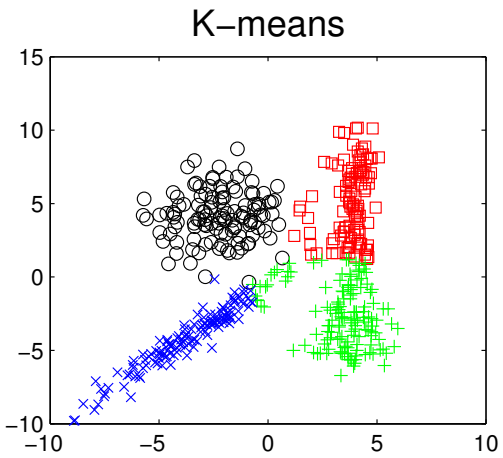
(d) Log-likelihood normalized by $N$ $N$ and $N_{test}$ (we normalize to obtain values which remain small when the number of data points increases and to be able to compare "test" and "train") :

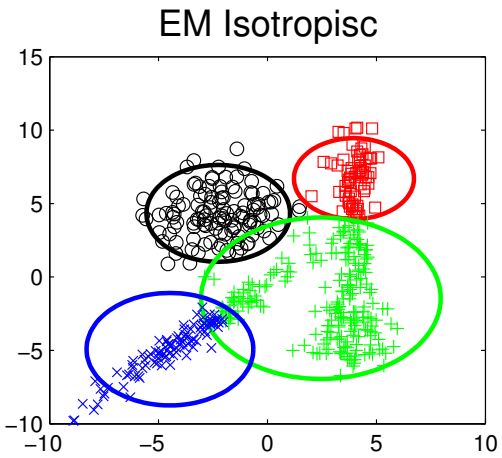
| | Train | Test |
|---|---|---|
| Isotropic | -5.2910 | -5.3882 |
| General | -4.6554 | -4.8180 |

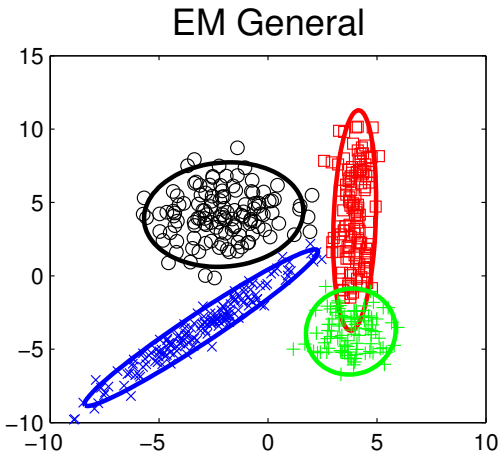
Unnormalized log-likelihoods :

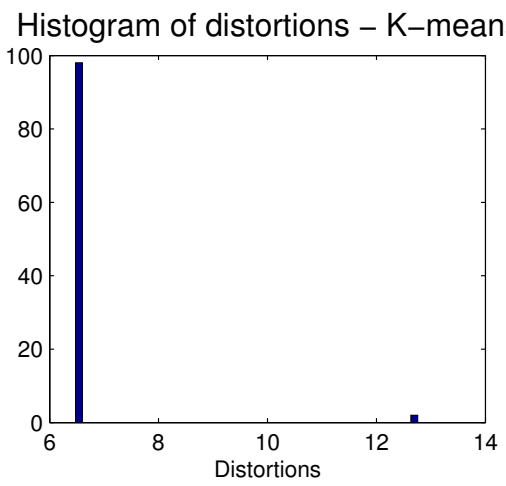
| | Train | Test |
|---|---|---|
| Isotropic | -2.6455 $\times 10^3$ | -2.6941 $\times 10^3$ |
| General | -2.3277 $\times 10^3$ | -2.4090 $\times 10^3$ |

The training log-likelihoods are always greater for more flexible models (the situation may be different for the testing log-likelihoods are 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