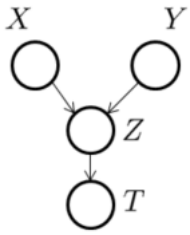
Master M2 MVA 2018/2019 - Graphical models - HWK 2

— These exercises are due on or before November 7th 2018.
— They should be submitted on the Moodle.
— **They can be done in groups of two students**.
— The write-up can be in English or in French, can be typed or scanned (except for the figures).
— **Please follow precisely the formatting described in Section 4.**
— Please submit your answers as a pdf file that you will name `MVA_DM2_<your_name>.pdf` if you worked alone or `MVA_DM2_<name1>_<name2>.pdf` with both of your names if you work as a group of two. Indicate your name(s) as well in the documents. Please submit your code as a separate zipped folder and name it `MVA_DM2_<your_name>.zip` if you worked alone or `MVA_DM2_<name1>_<name2>.zip` with both of your names if you worked as a group of two. Note that your files should weight no more than 16Mb.

1 Conditional independence and factorizations

1. Consider the directed graphical model G on the right. Write down the implied factorization for any joint distribution $p \in \mathcal{L}(G)$. Is it true that $X \perp\!\!\!\perp Y \mid T$ for any $p \in \mathcal{L}(G)$? Prove or disprove.



2. Let (X, Y, Z) be a r.v. on a finite space. Consider the following statement :

   "If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Y$ then ($X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$)."

   (a) Is this true if one assumes that Z is a binary variable ? Prove or disprove.

   (b) Is the statement true in general ? (*harder*) Prove or disprove.

2 Distributions factorizing in a graph

1. Let $G = (V, E)$ be a DAG. We say that an edge $i \to j \in E$ is a *covered edge* if and only if $\pi_j = \pi_i \cup \{i\}$; let $G' = (V, E')$, with $E' = (E \backslash \{i \to j\}) \cup \{j \to i\}$. Prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

2. Let G be a directed tree and G' its corresponding undirected tree (where the orientation of edges is ignored). Recall that by the definition of a directed tree, G does not contain any v-structure. Prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

3 Implementation - Gaussian mixtures

The file "EMGaussian.data" contains sample of data x_n where $x_n \in \mathbb{R}^2$. The goal of this exercise is to implement the EM algorithm for certain mixtures of K Gaussians in $\mathbb{R}^d$ (here $d = 2$ and $K = 4$), for i.i.d. data. (NB : in this exercise, no need to prove any of the formulas used in the algorithms except for question (b)).

The choice of the programming language is yours (we however recommend Python, Matlab, Scilab, Octave, or R). The source code should be handed in along with results. However all the requested figures should be printed on paper or part of a pdf file which is turned in (as requested in Section 4), with clear titles that indicate what the figures represent. The discussions may of course be handwritten.

(a) Implement the K-means algorithm. Represent graphically the training data, the cluster centers, as well as the different clusters. Try several random initializations and compare results (centers and distortion measures).

(b) Consider a Gaussian mixture model in which the covariance matrices are proportional to the identity. Derive the form of the M-step updates for this model and implement the corresponding EM algorithm (using an initialization with K-means).

Represent graphically the training data, the centers, as well as the covariance matrices (an elegant way is to represent the ellipse that contains a certain percentage, e.g., 90%, of the mass of the Gaussian distribution).

Estimate and represent (e.g., with different colors or different symbols) the latent variables for all data points (with the parameters learned by EM).

(c) Implement the EM algorithm for a Gaussian mixture with general covariance matrices. Represent graphically the training data, the centers, as well as the covariance matrices.

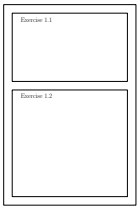
Estimate and represent (e.g., with different colors or different symbols) the latent variables for all data points (with the parameters learned by EM).

(d) Comment the different results obtained in earlier questions. In particular, compare the log-likelihoods of the two mixture models on the training data, as well as on test data (in "EMGaussian.test").

4 Formatting

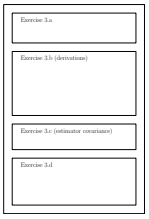
In order to reduce the correction time of these homeworks, please follow the following formatting precisely :

— Page 1



— Page 2

— Page 3



— Page 4 : Figures