**Master M2 MVA 2013/2014 - Graphical models**
**Take Home Exam**

Due on Wednesday December 18th, 2013.

You are requested to work on this take home exam alone, without exchanging information with other students.

If you intend to turn in this exam as a pdf file, please name the file

`MVA_DM3_<your_name>.pdf`

# 1 Inference in Gaussian graphical models

We consider an unoriented Gaussian graphical model with canonical parameters denoted $\eta \in \mathbb{R}^n$ for the loading vector and $\Lambda \in \mathbb{R}^{n \times n}$ for the precision matrix. We consider the model in which each scalar component of the Gaussian model is associated with a different node and we assume that the graphical model is a tree $G = (V, E)$. The goal of this exercise is to derive the form of a belief propagation algorithm to compute the marginal distributions on each of the individual nodes. As seen in class, the sum-product algorithm requires to exchange messages between nodes which take the form of potential-functions. In the case of discrete graphical models these messages are essentially vectors but in the case of Gaussian variables these potentials are functions from $\mathbb{R}$ to $\mathbb{R}$.

1. Show that, although, at an abstract level, the sum-product algorithm consists in exchanging messages $M_{s \to t}$ that are exponentials of quadratic functions, in practice, it is sufficient for node $s$ to send to another node $t$ a message which consists of two scalars $\lambda^{s \to t}$ and $\eta^{s \to t}$. Give the form of the recursion which allows to compute the messages $\lambda^{s \to t}$ and $\eta^{s \to t}$ sent from node $s$ to node $t$ given the messages received by $s$ from other nodes and the canonical parameters $\eta_s$, $\Lambda_{ss}$ and $\Lambda_{st}$.
2. Show that upon termination of the belief propagation algorithm, it is immediate to compute from the exchanged messages the quantities $\mu = \mathbb{E}[X]$ and $(\mathbb{E}[X_s X_t])_{\{s,t\} \in E}$.
3. What is the complexity of the belief propagation algorithm here ? What would be, for a general multivariate Gaussian distribution, the complexity of the computation of $\mu$ from $\Lambda$ and $\eta$ as a function of $n$ ? What is in fact the complexity of the computation of $\mu$ from $\Lambda$ and $\eta$ when the graph is a tree ? Comment on this result.
4. Let $A \in \mathbb{R}^{n \times n}$ a positive definite matrix such that for all $s \neq t$ and $(s, t) \notin E$ we have $A_{st} = 0$ ; let $b \in \mathbb{R}^n$. Deduce from the previous results an efficient algorithm to solve the linear system $Ax = b$. What is its complexity ?

## 2 Learning the structure of a tree graphical model

The goal of this exercise is to derive an algorithm to learn from i.i.d. data the structure of a graphical model constrained to be a tree.

1. Let $X$ be a discrete random variable over the finite set $\mathcal{X}$.
   Denote by $\eta(x) = p(X = x)$ the vector of parameters. Given an i.i.d. sample from the data $(x^n)_{n=1,\dots,N}$, let $\hat{p}(x)$ be the empirical distribution of the data defined by $\hat{p}(x) = \frac{1}{N}\sum_{n=1}^{N} \delta(x^n = x)$. Define the empirical entropy of $X$ based on the sample as
   $$\widehat{H}(X) = -\sum_{x\in\mathcal{X}} \hat{p}(x)\log\hat{p}(x).$$
   Show that the maximal value of the log-likelihood $\max_\eta \sum_{n=1}^{N} \log p(x^n|\eta)$ (attained for the maximum likelihood estimator) can be expressed as a function of $\widehat{H}(X)$ and $N$.
2. For two discrete random variables $X$ and $Y$ taking values respectively in $\mathcal{X}$ and $\mathcal{Y}$, we denote the joint entropy and the conditional entropy respectively by
   $$H(X,Y) = -\sum_{(x,y)} p(x,y)\log p(x,y) \quad \text{and} \quad H(X|Y) = -\sum_{(x,y)} p(x,y)\log p(x|y).$$
   Show that $H(X) + H(Y|X) = H(X,Y)$.
3. Let $(x^n, y^n)$, $n = 1,\dots,N$ be a sample of size $N$ for this pair of variables. Given the maximum likelihood estimator of the joint distribution $\hat{p}(x,y)$ we define the estimates of the joint entropy and the conditional entropy respectively as
   $$\widehat{H}(X,Y) = -\sum_{(x,y)} \hat{p}(x,y)\log\hat{p}(x,y) \quad \text{and} \quad \widehat{H}(X|Y) = -\sum_{(x,y)} \hat{p}(x,y)\log\hat{p}(x|y).$$
   Express the maximal value of the conditional log-likelihood $\max_\eta \sum_{n=1}^{N} \log p(y^n|x^n,\eta)$ as a function of $N$, $\widehat{H}(X,Y)$, $\widehat{H}(X)$ and $\widehat{H}(Y)$.
4. We now consider $P$ discrete random variables $X_1,\dots,X_P$ taking values respectively in the finite sets $\mathcal{X}_1,\dots,\mathcal{X}_P$. Given an i.i.d. sample of this vector of variables, $(x_p^n)$, $p = 1,\dots,P$, $n = 1,\dots,N$, we denote by $\hat{p}(x_1,\dots,x_P)$ the empirical distribution defined by
   $$\hat{p}(x_1,\dots,x_P) = \frac{1}{N}\sum_{n=1}^{N} \delta(x_1^n = x_1)\cdots\delta(x_P^n = x_P).$$
   This joint distribution induces the distributions $\hat{p}(x_p,x_q)$ et $\hat{p}(x_q)$ via marginalization.

Consider a spanning tree of the complete graph on $P$ vertices, and given an orientation of the tree (without v-structure), consider the corresponding graphical model, i.e. consisting of distributions that factorize as a product of conditionals according to the directed tree. Give the exact form of these distributions. What are the parameters ? For a node $p$, we denote its parent $\pi_p$, if it exists. Show that, once maximized with respect to its parameters, the log-likelihood for a given tree $\ell(T)$ can be expressed as a function of $N$ and of all the marginal empirical entropies $\widehat{H}(X_p)$ and $\widehat{H}(X_p, X_{\pi_p})$.

5. For all pairs $(p, q)$, the empirical mutual information is the quantity $\widehat{I}(X_p, X_q) = -\widehat{H}(X_p, X_q) + \widehat{H}(X_p) + \widehat{H}(X_q)$. Express it as a Kullback-Leibler divergence and show that it is non-negative.
6. Express $\ell(T)$ as a function of the entropies $\widehat{H}(X_p)$ and of the empirical mutual informations only.
7. Based on this expression of $\ell(T)$, we now consider the problem of maximizing $\ell(T)$ with respect to the choice of the tree $T$. Recognize that this maximization problem corresponds to a classical problem in graph theory and describe an algorithm to learn the structure of the tree having maximal likelihood.
8. Assuming that $|\mathcal{X}_1| = \ldots = |\mathcal{X}_P| = K$ what is the complexity of the algorithm as a function of $K$, $P$ and $N$ ?

# 3 HMM - Implementation

We consider the same training data as in the previous homework, provided as the "EMGaussienne.dat" file (and we will test on the corresponding testing data from the "EMGaussienne.test" file), but this time we use an HMM model to account for the possible temporal structure of the data. The data are of the form $u_t = (x_t, y_t)$ where $u_t = (x_t, y_t) \in \mathbb{R}^2$, for $t = 1, \ldots, T$. The goal of this exercise is to implement the probabilistic inference algorithm and the EM algorithm to learn parameters as well as the Viterbi algorithm. It is recommended to make use of the code of the previous homework.

We consider the following HMM model : the chain $(q_t)$ has $K = 4$ possible states, with an initial probability distribution $\pi \in \mathbb{R}^4$ and a probability transition matrix $A \in \mathbb{R}^{4\times 4}$, and conditionally on the current states we have observations obtained from Gaussian emission probabilities $u_t | q_t = i \sim \mathcal{N}(\mu_i, \Sigma_i)$.

1. Implement the recursions $\alpha$ et $\beta$ seen in class (and that can be found in the polycopié as well) to compute $p(q_t | u_1, \ldots, u_T)$ and $p(q_t, q_{t+1} | u_1, \ldots, u_T)$.
2. Using the same parameters for the means and covariance matrix of the 4 Gaussians as the ones obtained in the previous homework, taking a uniform initial

probability distribution $\pi$, and setting $A$ to be the matrix with diagonal coefficients $A_{ii} = \frac{1}{2}$ and off-diagonal coefficients $A_{ij} = \frac{1}{6}$ for all $(i, j) \in \{1, \ldots, 4\}^2$, compute $\alpha_t$ and $\beta_t$ for all t on the test data ("EMGaussienne.test" file) and compute $p(q_t|u_1, \ldots, u_T)$. Finally, represent $p(q_t|u_1, \ldots, u_T)$ for each of the 4 states as a function of time for the 100 first datapoints in the file. Note that only the 100 first points should be plotted by that filtering should be done with all the data (i.e. $T = 500$). This will be the same for the subsequent questions. (In Matlab the command subplot might be handy to make long horizontal plots.)

3. Derive the estimation equations of the EM algorithm.
4. Implement the EM algorithm to learn the parameters of the model ($\pi, A, \mu_k, \Sigma_k, k = 1 \ldots, 4$). The means and covariances could be initialized with the ones obtained in the previous homework. Learn the model from the training data in"EMGaussienne.dat".
5. Plot the log-likelihood on the train data "EMGaussienne.dat" and on the test data "EMGaussienne.test" as a function of the iterations of the algorithm. Comment.
6. Return in a table the values of the log-likelihoods of the Gaussian mixture models and of the HMM on the train and on the test data. Compare these values. Does it make sense to make this comparison ? Conclude. Compare these log-likelihoods as well with the log-likelihoods obtained for the different models in the previous homework.
7. Implement Viterbi decoding (aka MAP inference) to estimate the most likely sequence of states, i.e. $\arg\max_q p(q_1, \ldots, q_T|y_1, \ldots, y_T)$.

   For the set of parameters learned with the EM algorithm, compute the most likely sequence of states with the Viterbi algorithm and represent the data in 2D with the cluster centers and with markers of different colors for the datapoints belonging to different classes.
8. For the datapoints in the test file "EMGaussienne.test", compute the marginal probability $p(q_t|u_1, \ldots, u_T)$ for each point to be in state $\{1, 2, 3, 4\}$ for the parameters learned on the training set. For each state plot the probability of being in that state as a function of time for the 100 first points (i.e., as a function of the datapoint index in the file).
9. For each of these same 100 points, compute their most likely state according to the marginal probability computed in the previous question. Make a plot representing the most likely state in $\{1, 2, 3, 4\}$ as function of time for these 100 points.
10. Run Viterbi on the test data. Compare the most likely sequence of states obtained for the 100 first data points with the sequence of states obtained in the previous question. Make a similar plot. Comment.
11. In this problem the number of states was known. How would you choose the number of states if you did not know it ?