Exam (2h)

April 3, 2018

Exercises are independent and have roughly the same importance in the final grade. Please try to allocate half of your time to each part.

Exercise 1

We consider a random graph model with n vertices. Each vertex i is first associated with a latent vector Z_i assumed to be sampled from a multinomial :

$$Z_i \sim \mathcal{M}(1, \pi),$$

where $\pi^\intercal = (\pi_1, \ldots, \pi_K)$, $0 < \pi_k < 1, \forall k \in \{1, \ldots, K\}$ and $\sum_{k=1}^K \pi_k = 1$. By definition, $Z_i \in \{0,1\}^K$ such that $\sum_{k=1}^K Z_{ik} = 1, \forall i \in \{1, \ldots, n\}$. Moreover, all vectors Z_i are *iid*. Then, given Z_i and Z_j, an edge is generated with probability λ if $Z_i = Z_k$, and with probability ϵ if $Z_i \neq Z_j$. Thus, considering a directed graph without self loops, characterised by the adjacency matrix $X = (X_{ij})_{ij}$ $(\forall i \neq j)$, we have:

$$X_{ij} | Z_i, Z_j \sim \mathcal{B}(\mu(Z_i, Z_j)),$$

where $\mu(Z_i, Z_j) = \lambda$ if $Z_i = Z_j$ and $\mu(Z_i, Z_j) = \epsilon$ otherwise. Finally, given $Z = (Z_i)_i$ the set of all latent vectors, all edges in X are assumed to be independent.

Question 1 Draw the graphical model associated with this random graph model, for the (oriented) pair (i, j) of vertices.

Question 2 Explain the type of clusters of vertices that is modelled here. Give some of its properties.

Question 3 Give a R code to simulate a network according to this random graph model.

Question 4 Explain why the posterior $p(Z|X, \pi, \lambda, \epsilon)$ does not allow the use of the EM algorithm to estimate the model parameters π, λ, and ϵ.

Question 5 Considering the functional $r(Z)$, approximation of $p(Z|X, \pi, \lambda, \epsilon)$, for π, λ, and ϵ, give the variational decomposition of the observed data log-likelihood $\log p(X|\pi, \lambda, \epsilon)$ as a lower bound and a Kullback-Leibler divergence term.

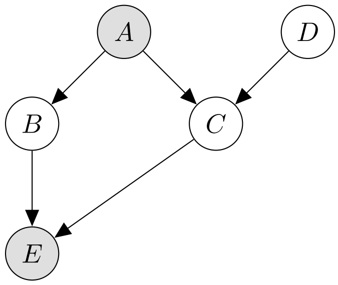
Question 6 Assuming $r(Z) = \prod_{i=1}^n r(Z_i) = \prod_{i=1}^n \prod_{k=1}^K \tau_{ik}^{Z_{ik}}$, give the complete expression of the lower bound.

Question 7 Give the M-step of the variational EM algorithm maximizing the lower bound to estimate π.

Question 8 Give the M-step of the variational EM algorithm maximizing the lower bound to estimate the probabilities λ and ϵ.

Exercise 2

We study the following directed graphical model:



Question 1 Write the factorization condition that has to be fulfilled by any distribution on the random variables A, B, C, D and E that is compatible with the **directed** graphical model.

Question 2 Justify the truth or the falseness of each of the following (conditional) independence properties for distributions compatible with the **directed** graphical model.

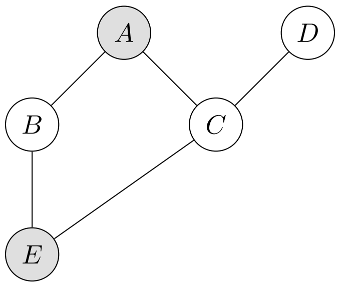
$$A \perp\!\!\!\perp E \mid B \tag{1}$$
$$B \perp\!\!\!\perp D \tag{2}$$

Question 3 Are there groups of variables that are conditionally independent given E for a distribution compatible with the **directed** graphical model?

Question 4 Express the conditional distribution $p(e|a,d)$ for a distribution compatible with the **directed** graphical model as a sum over b and c of products of conditional distributions that naturally appear in compatible distributions.

Question 5 Draw a factor graph that represents distributions compatible with the **directed** graphical model.

We introduce now the following undirected graphical model:



Question 6 Write the factored form of a distribution on A, B, C, D and E that is compatible with the **undirected** graphical model.

We assume now on that the random variables A, B, C, D and E are binary. We denote $q(x,p)$ the likelihood of observation x for a Bernoulli distribution with parameter p (that is $q(x,p) = p^x(1-p)^{1-x}$).

Question 7 Give an example of a simple distribution on A, B, C, D and E that is compatible with the **directed** graphical model but not with the **undirected** graphical model.

Question 8 Add the minimal number of edges needed to the **undirected** graphical model such that any distribution compatible with the **directed** graphical model is also compatible with the modified **undirected** graphical model.

Question 9 Write the factored form of a distribution on A, B, C, D and E that is compatible with the modified **undirected** graphical model.

Question 10 Give an example of two variables that are conditionally independent given a third one for distributions that are compatible with the **directed** graphical model but for which there are distributions compatible with the **modified undirected** graphical model under which that are not independent regardless of the conditioning set. In other words find X_1, X_2 and X_3 (among A, B, C, D and E) such that for any distribution compatible with the **directed** graphical model $X_1 \perp\!\!\!\perp X_2 \mid X_3$ and such that a distribution in which $X_1 \perp\!\!\!\perp X_2 \mid S$ is false (for any S subset of $\{A, B, C, D, E\} \setminus \{X_1, X_2\}$) can be compatible with the **modified undirected** graphical model.

Question 11 Using the **directed** graphical model, we assume that A and E are observed and that B, C, and D are latent. What conditional distribution would be needed if we were to apply the EM algorithm to maximize the likelihood of (a, e) with respect to the parameters of a distribution compatible with the **directed** graphical model.

Question 12 Write $p(a, e)$ using only the conditional distributions that naturally appear in distributions compatible with the **directed** graphical model in a way that is maximally factored (i.e. which corresponds to the most efficient evaluation strategy).