Master MVA 2016-2017
Probabilistic graphical models: Final exam

December 14th 2016

The duration of the exam is 3 hours. You may use any printed references including books. The use of any electronic device (computer, tablet, calculator, smartphone) is forbidden.

All questions require a proper mathematical justification or derivation, but most questions can be answered in just a few lines.

You may write your answers either in French or in English.

1 Mixture of Pareto distributions (4 points: 1 + 3)

The Pareto distributions are distributions on $[1, \infty)$ which admit a density on the real line of the form

$$p(x) = (\alpha - 1)\, x^{-\alpha}\, 1_{\{x \geq 1\}}.$$

(a) Show that the family of Pareto distributions form an exponential family and specify its sufficient statistic, its canonical parameter, its log-partition function and its domain (the set of values of the canonical parameter such that the distribution is well defined).

We have

$$p(x) = \exp\big(-\alpha \log(x) + \log(\alpha - 1)\big),$$

We thus have

- $\phi(x) = -\log(x)$
- $\eta = \alpha$
- $\Omega = \{\eta \mid A(\eta) < \infty\} = (1, \infty)$
- $A(\alpha) = -\log(\alpha - 1)$
- $\mu = \mathbb{E}[\phi(X)] = -\mathbb{E}[\log(X)] = A'(\alpha) = -\frac{1}{\alpha - 1}$

(b) Compute its moment parameter.

(c) Consider distributions with densities of the form $p(x) = 1_{x \geq 1} \sum_{k=1}^{K} \pi_k(\alpha_k - 1)x^{-\alpha_k}$, with $\pi_1 + \ldots + \pi_K = 1$. Propose an EM algorithm to estimate the parameters π_k and α_k for $k \in \{1, \ldots, K\}$.

Step E:

$$q_{ik}^t := p_{\theta^{t-1}}(z_{ik} = 1 \mid x_i) = \frac{\pi_k^{t-1}(\alpha_k^{t-1} - 1)x_i^{\alpha_k^{t-1}}}{\sum_{j=1}^{K} \pi_j^{t-1}(\alpha_j^{t-1} - 1)x_i^{\alpha_j^{t-1}}}.$$

Step M: By moment matching on the expected log-likelihood:

$$\mu_k^t = -\frac{\sum_i q_{ik}^t \log(x_i)}{\sum_i q_{ik}^t}, \quad \pi_k^t = \frac{1}{n}\sum_i q_{ik}^t, \quad \alpha_k^t = 1 - \frac{1}{\mu_k^t}.$$

Note that by definition $\mu_k^t < 0$ so that we always have $\alpha_k^t > 1$.

2 Gaussian Markov Chain (4 points: 2 + 2)

Assume that the ε_i are i.i.d. with $\varepsilon_i \sim \mathcal{N}(0,1)$ and let $X_1 = \varepsilon_1$, $X_2 = \rho X_1 + \varepsilon_2$, $X_3 = \rho X_2 + \varepsilon_3$.

(a) What is the precision matrix of the joint distribution of (X_1, X_2, X_3)?

(b) Compute $\mathbb{E}[X_2 \mid X_1, X_3]$ and $\mathrm{Var}(X_2 \mid X_1, X_3)$.

The joint density is

$$\propto \exp -\frac{1}{2}\left[x_1^2 + (x_2 - \rho x_1)^2 + (x_3 - \rho x_2)^2\right],$$

which reveals that

$$\Lambda = \begin{bmatrix} 1+\rho^2 & -\rho & 0 \\ -\rho & 1+\rho^2 & -\rho \\ 0 & -\rho & 1 \end{bmatrix} \quad \text{and} \quad \eta = \mu = 0.$$

So that

$$Var(X_2 \mid X_1, X_3) = \Lambda_{22}^{-1} = \frac{1}{1+\rho^2}$$

and

$$\mathbb{E}[X_2 \mid X_1, X_3] = \eta_2 - \Lambda_{22}^{-1}\Lambda_{2,(1,3)}(X_1, X_3)^\top = \frac{\rho}{1+\rho^2}(X_1 + X_3).$$

The computation of the conditional mean variance can also be obtained from completing the square in the expression of the density:

$$p(x_2|x_1,x_3) \propto \exp -\frac{1}{2}\left[(1+\rho^2)x_2^2 - 2\rho x_1 x_2 - 2\rho x_3 x_2\right] \propto \exp -\frac{1}{2}(1+\rho^2)\big(x_2 - 2\tfrac{\rho}{1+\rho^2}(x_1+x_3)\big)^2.$$

3 Linear regression and Gaussian likelihood (4 points: 1 + 3)

(a) Let $\psi : H \to \Theta$ be a surjective mapping from $H \in \mathbb{R}^p$ to $\Theta \in \mathbb{R}^p$. Consider the two statistical models

$$\mathcal{P} := \{p_{\psi(\eta)} \mid \eta \in H\}, \quad \text{and} \quad \mathcal{P}' := \{p_\theta \mid \theta \in \Theta\},$$

with $p_\theta = p_{\psi(\eta)}$ when $\theta = \psi(\eta)$. Assume that based on a sample $\mathcal{S}$, the maximum likelihood estimator $\hat{\eta}$ for η in $\mathcal{P}$ exists and is unique. Show that the maximum likelihood estimator $\hat{\theta}$ for θ in $\mathcal{P}'$ exists, is unique, and that $\hat{\theta} = \psi(\hat{\eta})$.

Let $\tilde{\theta} = \psi(\hat{\eta})$. The surjectivity of ψ entails that, for all $\theta' \in \Theta$ there exists $\eta' \in H$ such that $\theta' = \psi(\eta')$ and we have

$$p_{\theta'} = p_{\psi(\eta')} \leq p_{\psi(\hat{\eta})} = p_{\tilde{\theta}}.$$

Since this is true for all θ', this shows that $\tilde{\theta}$ maximizes the likelihood. This is moreover the unique maximizer, because if there existed another maximizer $\tilde{\theta}' \neq \tilde{\theta}$, given that ψ is surjective, there would exist $\tilde{\eta}'$ such that $\tilde{\theta}' = \psi(\tilde{\eta}')$ and $\tilde{\eta}'$ would be a second maximizer of the likelihood in $\mathcal{P}$ which would contradict the uniqueness of $\hat{\eta}$. We thus have shown that $\tilde{\theta}$ is the unique maximizer of the likelihood and so $\hat{\theta} = \tilde{\theta}$.

(b) Consider a pair of random variables (X, Y) with $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ and with some joint distribution P. Assume that an i.i.d. sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn from P is available. Assume that P is sufficiently nice and n sufficiently large, so that the maximum likelihood estimator for the parameters of the Gaussian model exists and is unique. Let

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_x \\ \hat{\mu}_y \end{pmatrix} \quad \text{and} \quad \widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_{xx} & \widehat{\Sigma}_{xy} \\ \widehat{\Sigma}_{yx} & \widehat{\sigma}_y^2 \end{pmatrix}$$

be respectively the maximum likelihood estimator of the mean and covariance matrix with $\hat{\mu}_x \in \mathbb{R}^p$, $\hat{\mu}_y \in \mathbb{R}$, $\widehat{\Sigma}_{xx} \in \mathbb{R}^{p \times p}$, $\widehat{\Sigma}_{xy} = \widehat{\Sigma}_{yx}^\top \in \mathbb{R}^p$, and $\widehat{\sigma}_y^2 \in \mathbb{R}_+$. In particular we assume that $\widehat{\Sigma}$ is invertible.

Assume now that still based on the same sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, a linear regression model of the form $y = w^\top x + b$ is estimated using ordinary least-squares regression (where w and b are the parameters for the uncentered and unnormalized data). Let $\hat{w}$ and $\hat{b}$ be obtained estimators and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{w}^\top x_i - \hat{b})^2$ the average squared residuals. Show that $\hat{w}$, $\hat{b}$ and $\hat{\sigma}^2$ can be expressed as a function of $\hat{\mu}_x, \hat{\mu}_y, \widehat{\Sigma}_{xx}, \widehat{\Sigma}_{xy}$ and $\widehat{\sigma}_y^2$. Explain why and provide the formulas.

The linear regression function $z \mapsto \hat{w}^\top z + \hat{b}$ obtained by ordinary least-squares and the average squared residuals $\hat{\sigma}^2$ are actually the maximum likelihood estimates, respectively for the conditional expectation and for the conditional variance of a conditional Gaussian model of Y given X. A joint Gaussian model on (X, Y) can thus be reparameterized by splitting it into a Gaussian marginal distribution of X with mean $\mu_x^{marg} = \mu_x$ and covariance $\Sigma_{xx}^{marg} = \Sigma_{xx}$ and a conditional Gaussian distribution of Y given X with

$$\begin{aligned} \mathbb{E}[Y \mid X] = w^\top X + b &= \Sigma_{yx}\Sigma_{xx}^{-1}(X - \mu_x) + \mu_y \\ \mathrm{Var}(Y \mid X) = \sigma^2 &= \sigma_y^2 - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} \end{aligned}$$

$$\begin{bmatrix} \mu_x^{marg} \\ \Sigma_{xx}^{marg} \\ w \\ b \\ \sigma^2 \end{bmatrix} = \psi(\mu, \Sigma) = \begin{bmatrix} \mu_x \\ \Sigma_{xx} \\ \Sigma_{xx}^{-1}\Sigma_{xy} \\ \mu_y - \Sigma_{yx}\Sigma_{xx}^{-1}\mu_x \\ \sigma_y^2 - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} \end{bmatrix}.$$

Clearly ψ is a bijection since

$$\begin{bmatrix} \Sigma_{yx} \\ \mu_y \\ \sigma_y^2 \end{bmatrix} = \begin{bmatrix} \Sigma_{xx} w \\ b + w^\top \mu_x \\ \sigma^2 + w^\top \Sigma_{xx} w \end{bmatrix}$$

Applying the result of the previous question, if the MLE is unique for μ and Σ then it is also unique for w, b and σ^2 and

$$\begin{bmatrix} \hat{w} \\ \hat{b} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \widehat{\Sigma}_{xx}^{-1}\widehat{\Sigma}_{xy} \\ \mu_y - \widehat{\Sigma}_{yx}\widehat{\Sigma}_{xx}^{-1}\mu_x \\ \widehat{\sigma}_y^2 - \widehat{\Sigma}_{yx}\widehat{\Sigma}_{xx}^{-1}\widehat{\Sigma}_{xy} \end{bmatrix}.$$

4 Message passing for tridiagonal systems

(12 points: 1 + 1 + 2 + 2 + 2 + 2 + 1 + 1)

We consider a positive definite matrix $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. We assume that A is tridiagonal (i.e., $|A_{ij}| = 0$ as soon as $|i-j| > 1$), and we aim to solve the system $A\mu = b$ in the variable $\mu \in \mathbb{R}^n$, using the sum-product algorithm.

(a) We consider the Gaussian vector $x \in \mathbb{R}^n$ with precision matrix A and loading vector b. Write down the density of x.

We have $p(x) = \frac{|A|^{1/2}}{(2\pi)^{n/2}} \exp(-\frac{1}{2}x^\top Ax + b^\top x)$.

(b) What graphical model assumption is the tridiagonality of A equivalent to? Describe how to apply the sum-product algorithm to compute $\mu = A^{-1}b$.

Markov chain on which we can apply sum-product. After $2(n-1)$ messages, can get marginals and hence their expectations, which are exactly the components of μ.

(c) Show that the left-to-right message passing recursions may be written in the form (using integrals instead of sums)

$$m_{k\to k+1}(x_{k+1}) \propto \int_{-\infty}^{+\infty} \exp\Big(-\frac{1}{2}A_{k,k}x_k^2 + b_k x_k - A_{k,k+1}x_k x_{k+1}\Big) m_{k-1\to k}(x_k)dx_k.$$

Just apply the formula seen in class.

(d) Show that all left-to-right messages may be written in a form $m_{k-1\to k}(x_k) \propto \exp(-\frac{1}{2}c_k x_k^2 + d_k x_k)$, and provide a recursion between (c_k, d_k) and (c_{k+1}, d_{k+1}). What is the initialization?

The integral above is proportional to

$$\begin{aligned}
&\int_{-\infty}^{+\infty} \exp\Big(-\frac{1}{2}A_{k,k}x_k^2 + b_k x_k - A_{k,k+1}x_k x_{k+1}\Big) \exp(-\frac{1}{2}c_k x_k^2 + d_k x_k)dx_k \\
\propto\ &\int_{-\infty}^{+\infty} \exp\Big(-\frac{1}{2}[A_{k,k} + c_k]x_k^2 + [b_k + d_k - A_{k,k+1}x_{k+1}]x_k\Big)dx_k \\
\propto\ &\exp\Big(\frac{1}{2}[A_{k,k} + c_k]^{-1}[b_k + d_k - A_{k,k+1}x_{k+1}]^2\Big),
\end{aligned}$$

leading to $c_{k+1} = \frac{-A_{k,k+1}^2}{A_{k,k}+c_k}$ *and* $d_{k+1} = \frac{b_k - A_{k,k+1}d_k}{A_{k,k}+c_k}$. *This is initialized as* $c_0 = d_0 = 0$.

(e) Parameterize the right-to-left recursion and derive the update equations, as well as their initialization.

The right-to-left recursion for $m_{k+1\to k}(x_k) \propto \exp(-\frac{1}{2}e_k x_k^2 + f_k x_k)$ *is initialized as* $e_{n+1} = f_{n+1} = 0$ *and with recursion:* $e_k = \frac{-A_{k,k+1}^2}{A_{k,k}+e_{k+1}}$ *and* $f_k = \frac{b_{k+1} - A_{k+1,k+1}f_{k+1}}{A_{k+1,k+1}+e_{k+1}}$.

(f) Describe how to compute $\mu = A^{-1}b$ from all messages.

We simply have $p(x_k) \propto \exp\Big(-\frac{1}{2}A_{k,k}x_k^2 + b_k x_k\Big) m_{k+1\to k}(x_k)\ m_{k-1\to k}(x_k)$ *, which is equal to* $\exp\Big(-\frac{1}{2}[A_{k,k} + c_k + e_k]x_k^2 + [b_k + d_k + f_k]x_k\Big)$, *leading to* $\mu_k = \frac{b_k + d_k + f_k}{A_{k,k}+c_k+e_k}$.

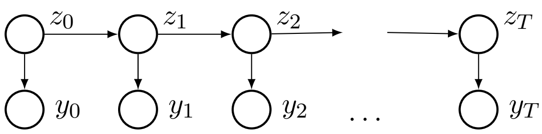

Figure 1: HMM for problem 5

(g) What is the running-time complexity of the algorithm above? Comment.

$O(n)$, compared to $O(n^3)$ for regular algorithms from numerical algebra.

(h) If we replace the sum-product algorithm by the max-product algorithm, what is changed?

Nothing, as up to constant terms, maximizing and marginalizing is the same.

5 Conditional independence and sampling in the HMM model

(8 points: 1 + 2 + 2 + 3)

Consider an HMM model as on Figure 1 with initial distribution π, with transition matrix A and emission probability $p(y_t|z_t)$. Assume that the values $y_0 = \bar{y}_0, \ldots, y_T = \bar{y}_T$ are observed, that the forward-backward algorithm has been run based on these observations and that all the messages $\alpha_t(z_t)$ and $\beta_t(z_t)$ are available for all t and all values of z_t.

(a) Consider a fixed value of $(y_0, \ldots, y_T)$. Show that if we define the distribution q by $q(z_0, \ldots, z_T) := p(z_0, \ldots, z_T \mid y_0, \ldots, y_T)$, then q factorizes according to an undirected chain graph with nodes indexed by $\{0, \ldots, T\}$ and edges $(t-1, t)$ for $t \in \{1, \ldots, T\}$. In particular, express the unary and binary potentials in terms of $p(z_t|z_{t-1})$, $p(z_0)$ and $p(y_t|z_t)$.

Take $\psi_t(z_t) = p(y_t|z_t)$ and $\psi_{t-1,t}(z_{t-1}, z_t) = p(z_t, z_{t-1})$. Then the partition function is $Z = p(y_0, \ldots, y_T)$.

(b) Prove or disprove the following statements:

- $p(z_t|z_0, \ldots, z_{t-1}, y_0, \ldots, y_T) = p(z_t|z_{t-1}, y_0, \ldots, y_T)$
- $p(z_t|z_0, \ldots, z_{t-1}, z_{t+1}, \ldots, z_T, y_0, \ldots, y_T) = p(z_t|z_{t-1}, z_{t+1}, y_0, \ldots, y_T)$

The statements are equivalent to $q(z_t|z_0, \ldots, z_{t-1}) = q(z_t|z_{t-1})$ and $q(z_t|z_{-t}) = q(z_t|z_{t-1}, z_{t+1})$. These statements are obviously true given that q is the distribution of a Markov chain, which we just proved in the previous question.

(c) Propose an algorithm to sample *exactly* $(z_0, \ldots, z_T)$ from its marginal distribution, given that you know A and π. (To sample *exactly* here means that after a finite number of operations we have that $(z_0, \ldots, z_T)$ is exactly drawn from the desired distribution. We are not interested by *approximate* inference techniques.).

We can use ***ancestral sampling****, with the conditional distribution given by $p(z_{t+1}|z_t) = z_t^\top A\, z_{t+1}$.*

(d) Propose an algorithm with complexity linear in T to sample *exactly* $(z_0, \ldots, z_T)$ from the distribution $p(z_0, \ldots, z_T \mid \bar{y}_0, \ldots, \bar{y}_T)$, knowing A,π and the value of $p(y_t|z_t)$ for all possible values of z_t. Furthermore, this algorithm should not use either rejection sampling or importance sampling. (As before sample *exactly* here means that after a finite number of operations we have that $(z_0, \dots, z_T)$ is exactly drawn from the desired distribution. We are not interested by *approximate* inference techniques.).

We have

$$\begin{cases} Z = \sum_{z_t} \alpha_t(z_t)\beta_t(z_t), \\ q(z_t) = \frac{1}{Z}\alpha_t(z_t)\beta_t(z_t), \\ q(z_t, z_{t+1}) = \frac{1}{Z}\alpha_t(z_t)p(z_{t+1}|z_t)p(y_{t+1}|z_{t+1})\beta_{t+1}(z_{t+1}) \end{cases}$$

so that

$$q(z_{t+1}|z_t) = p(z_{t+1}|z_t)p(y_{t+1}|z_{t+1})\frac{\beta_{t+1}(z_{t+1})}{\beta_t(z_t)}.$$

Given that $q(z_0, \dots, z_T) = q(z_0)q(z_1|z_0)\dots q(z_T, z_{T-1})$*, we can use ancestral sampling.*