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).

(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 $k \in \{1, \ldots, K\}$.

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 $\mathrm{Var}(X_2 \mid X_1, X_3)$.

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 $\hat{\theta} = \psi(\hat{\eta})$.

(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.

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.

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

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

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

(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?

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

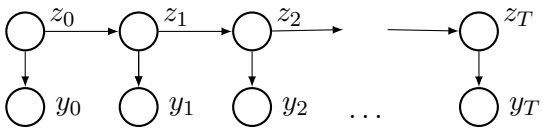

Figure 1: HMM for problem 5

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

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

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

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 $p(y_t|z_t)$.

(b) Prove or disprove the following statements:

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

(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.).

(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 $(z_0, \ldots, z_T)$ is exactly drawn from the desired distribution. We are not interested by *approximate* inference techniques.).