Machine Learning 2 **Due: 23:59, September 23, 2019**

Homework 3

Instructors: Adeel Pervez, Pim de Haan, Noud de Kroon, David Zhang
Email: a.a.pervez@uva.nl, pim.de.haan@uva.nl, a.a.w.m.dekroon@uva.nl, w.d.zhang@uva.nl

Problem 1. (0.5 + 0.5 = 1 pts)

1. Given the definition of the entropy, conditional entropy

$$H(X) = \mathop{\mathrm{E}}_{p(x)}\big[-\log p(x)\big] \tag{1}$$
$$H(Y|X) = \mathop{\mathrm{E}}_{p(x,y)}\big[-\log p(y|x)\big] \tag{2}$$

   show that the following equalities hold:

$$H(X,Y) = H(X) + H(Y|X) \tag{3}$$
$$= H(Y) + H(X|Y). \tag{4}$$

2. Consider the conditional mutual information defined by

$$I(X;Y|Z) = \mathop{\mathrm{E}}_{p(z)}\big[KL\big(p(x,y|z)||p(x|z)p(y|z)\big)\big] \tag{5}$$

   and show that

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{6}$$
$$= H(Y|Z) - H(Y|X,Z). \tag{7}$$

Problem 2. (0.75 + 0.75 + 0.75 + 0.75 = 3 pts) Consider the multinomial distribution:

$$\mathrm{Mult}(\mathbf{x}|\boldsymbol{\pi}) = \frac{M!}{x_1!x_2!\cdots x_K!}\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_K^{x_K}$$

whre x_i are non-negative integers such that $\sum_{i=1}^K x_i = M$ and π_i are constants with $\pi_i > 0$ and $\sum_{i=1}^K \pi_i = 1$.

1. Show that it is a member of an exponential family. Derive the minimal representation, i.e. express it in terms of a minimal number of parameters, sufficient statistics and log partition function.

2. Derive the mean and covariance from the log partition function.

3. Construct the conjugate prior family of this distribution (up to a normalization constant). To which family belongs this conjugate prior?

4. For n i.i.d. multinomial observations write down the prior-to-posterior update rule for the hyperparameters.

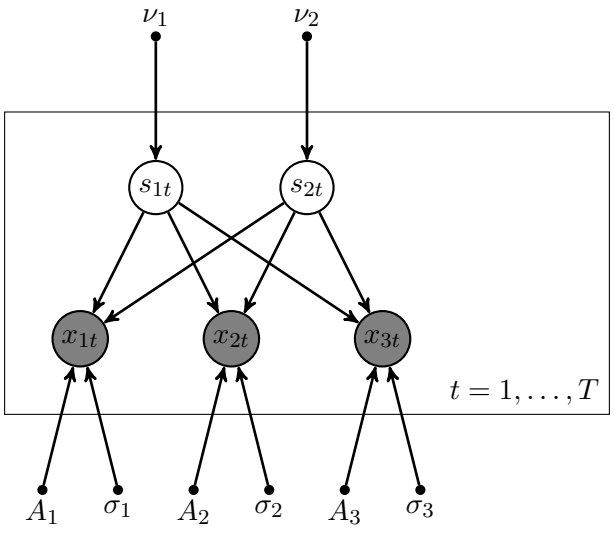
Problem 3. (0.25 + 0.25 + 0.25 + 0.25 + 0.25 + 0.5 + 0.25 + 0.25 + 0.25 = 2.5 pts)
Consider a time sequence of T samples from K_s *statistically independent* sound sources: $\{s_{it}\} = (s_{i1}, ..s_{iT})$, where i labels the source and t the time it was emitted. We record K_x *noisy* linear mixtures of these sound sources:

$$x_{kt} = \sum_{i=1}^{K_s} A_{ki} s_{it} + \epsilon_{kt} \quad t = 1..T, \quad k = 1..K_x$$
$$s_{it} \sim \mathcal{T}(0, \nu_i), \qquad \epsilon_{kt} \sim \mathcal{N}(0, \sigma_k^2).$$

where s_{it} is distributed as a zero mean Student's T distribution with ν_i degrees of freedom and ϵ_{kt} a noise random variable drawn from a zero mean normal (Gaussian) distribution with standard deviation σ_k. We assume that the sources were generated independently and we also assume that there are no statistical dependencies between samples generated at different points in time.

1. Explain why this is an ICA model.

For $K_s = 2$ sound sources and $K_x = 3$ recordings we have the following graphical model:



Where $A_k = (A_{k1}, A_{k2})$ and we used the plate notation to indicate the replication over T.

2. Write a general (Bayesian network) expression for the joint probability distribution

$$p(\{s_{1t}\}, \{s_{2t}\}, \{x_{1t}\}, \{x_{2t}\}, \{x_{3t}\}), \quad t = 1..T$$

Factorize the distribution into smaller conditional and marginal distributions as much as possible. Use explicit (conditional) distributions such as Normal and Student's T distributions instead of a generic form "p" as much as possible.

3. Explain what the term "explaining away" means and indicate if this explaining away phenomenon is present in the ICA model under discussion.

4. Since samples across time t are independent, we will ignore the index t in the following two questions (you may imagine $t = 1$). For all of the (conditional) independence expressions below, state if they are true or (typically) false:

   (a) $x_1 \perp\!\!\!\perp x_2|\varnothing$

   (b) $s_1 \perp\!\!\!\perp s_2|\varnothing$

   (c) $x_1 \perp\!\!\!\perp s_1|\varnothing$

   (d) $x_1 \perp\!\!\!\perp x_2|\{s_1, s_2\}$

   (e) $x_1 \perp\!\!\!\perp x_2|s_1$

   (f) $s_1 \perp\!\!\!\perp s_2|\{x_1, x_2, x_3\}$

   (g) $s_1 \perp\!\!\!\perp s_2|x_1$

   (h) $x_1 \perp\!\!\!\perp s_1|\{s_2, x_2, x_3\}$

5. What is the Markov blanket of s_1? What is the Markov blanket of x_1?

From now on we will assume that the number of sources and the number of recordings are the same (complete ICA), i.e. $K_x = K_s = K$. We will also assume that the relation between sources and recordings is deterministic, i.e.

$$x_{kt} = \sum_{i=1}^{K} A_{ki} s_{it} \qquad t = 1..T, \qquad k = 1..K$$
$$s_{it} \sim \mathcal{T}(0, \nu_i).$$

We call $W = A^{-1}$ the inverse of the mixing matrix A (aka the "unmixing matrix"), such that

$$s_{it} = \sum_{k=1}^{K} W_{ik} x_{kt} \qquad t = 1, ..T, \qquad i = 1..K$$

In the following question, you may use the general expression:

$$p_X(x) = p_S(s(x)) \; |\det Jac(s \to x)|$$

where $Jac(s \to x)$ is the Jacobian for a transformation from the random variable s to x.

6. Write an explicit expression in terms of W and the sources' student's T distributions $\mathcal{T}(s_i|0, \nu_i)$ of the probability:
$$p(\{x_{kt}\}|W, \{\nu_i\}) \qquad t = 1..T, \quad k = 1..K$$

7. Write down the *log-likelihood* of the complete deterministic ICA model above.

8. Explain in detail the "stochastic gradient ascent" optimization algorithm to maximize the log-likelihood of the previous question. Note: you do not have to derive or provide the expression of the gradient; instead you can provide a general description of the algorithm.

9. In which limit do you expect overfitting: $K \gg T$ or $T \gg K$? Explain your answer.

Problem 4. (0.5 + 0.5 + 0.75 + 0.75 = 2.5 points)

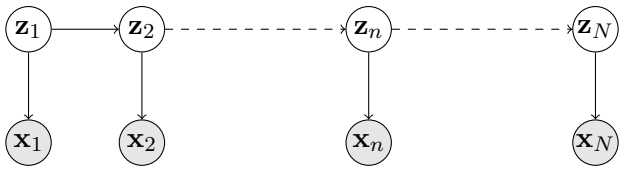


Figure 1: Markov chain of latent variables.

Given a graphical model in Figure 1. Show that:

1. $p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{x}_n, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{z}_n)$
2. $p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1})$
3. $p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_n, \mathbf{z}_{n+1}) = p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N | \mathbf{z}_{n+1})$
4. $p(\mathbf{z}_{N+1} | \mathbf{z}_N, \mathbf{X}) = p(\mathbf{z}_{N+1} | \mathbf{z}_N)$, where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Use d-separation for the first two equalities and use the factorization properties of the graphical model for the third and fourth equalities.