UML2015
Exercise set 6
Solutions to be presented in the 30.4.2015 session

Exercise 1:
The sum and product notations are

$$\sum_{m=1}^{M} a_m = a_1 + a_2 + \cdots + a_M \quad \text{and} \quad \prod_{m=1}^{M} a_m = a_1 a_2 \cdots a_M \tag{1}$$

Suppose you want to sum up values $f(a)$ where $a = (a_1, a_2, \ldots, a_M)$ is a vector, and each element a_m is supposed to run through all the values from the set $\{1, 2, \ldots, D\}$ with some $D \in \mathbb{N}$. We assume that the function has the form $f : \{1, 2, \ldots, D\}^M \to \mathbb{R}$. In the sum the number of terms will be D^M. According to Fubini's Theorem the sum can be written as an iterated sum:

$$\sum_{a \in \{1,2,\ldots,D\}^M} f(a) = \sum_{a_M=1}^{D} \Big(\cdots \Big(\sum_{a_2=1}^{D} \Big(\sum_{a_1=1}^{D} f(a) \Big) \Big) \cdots \Big) \tag{2}$$

It is common to leave the parentheses out, and we can also define a notation

$$\Big(\prod_{m=1}^{M} \sum_{a_m=1}^{D} \Big) f(a) = \sum_{a_M=1}^{D} \cdots \sum_{a_2=1}^{D} \sum_{a_1=1}^{D} f(a), \tag{3}$$

which is justified since in the iterated sum it looks as if the sum signs are being multiplied. The formula works for all f, and this can be emphasized by writing the Fubini's Theorem as

$$\sum_{a \in \{1,2,\ldots,D\}^M} = \prod_{m=1}^{M} \sum_{a_m=1}^{D} \tag{4}$$

1.1 Suppose we have M different functions $f_m : \{1, 2, \ldots, D\} \to \mathbb{R}$, and the f has been defined by a formula

$$f(a) = \prod_{m=1}^{M} f_m(a_m) \tag{5}$$

Prove the formula

$$\Big(\prod_{m=1}^{M} \sum_{a_m=1}^{D} \Big) \Big(\prod_{m'=1}^{M} f_{m'}(a_{m'}) \Big) = \prod_{m=1}^{M} \Big(\sum_{a_m=1}^{D} f_m(a_m) \Big) \tag{6}$$

1.2 Suppose we have functions $f_m : \mathbb{R} \to \mathbb{R}$, and have then defined $f : \mathbb{R}^M \to \mathbb{R}$ as

$$f(x) = \prod_{m=1}^{M} f_m(x_m) \tag{7}$$

In this context, explain the formulas

$$\int_{\mathbb{R}^M} dx = \prod_{m=1}^{M} \int_{-\infty}^{\infty} dx_m \tag{8}$$

and

$$\Big(\prod_{m=1}^{M} \int_{-\infty}^{\infty} dx_m \Big) \Big(\prod_{m'=1}^{M} f_{m'}(x_{m'}) \Big) = \prod_{m=1}^{M} \Big(\int_{-\infty}^{\infty} f_m(x_m) dx_m \Big) \tag{9}$$

Exercise 2:

If A is a discrete random variable, its probabilities are often denoted as $P(A = a)$. If A instead is a continuous random variable, its probability density is often denoted as $p_A(a)$. In both cases, the probability that A is in some set $\mathcal{A}$ is often denoted as $P(A \in \mathcal{A})$. However, we can also denote densities as $p(A = a)$, since this is only a matter notation. Also, we can denote probabilities with small "p" as $p(A \in \mathcal{A})$. If A is an N-dimensional random vector, its probability density (assuming the density exists and is continuous) can be written in terms of probabilities as

$$p(A = a) = \lim_{\varepsilon \to 0} \frac{p(A \in a + \varepsilon \mathcal{A})}{\varepsilon^N m(\mathcal{A})} \tag{10}$$

where $\mathcal{A}$ is some small set (such as a ball) containing the origin and $m(\mathcal{A})$ is its N-dimensional measure (generalized length, area or volume). The notation $a + \varepsilon \mathcal{A}$ means the set $\{a + \varepsilon x | x \in \mathcal{A}\}$ (or equivalently $\{y | \frac{y-a}{\varepsilon} \in \mathcal{A}\}$).

2.1 Assume that A is a continuous random variable (N-dimensional vector), while B can be either continuous or discrete. We can define probability quantities such as

$$p(A = a, B \in \mathcal{B}) = \lim_{\varepsilon \to 0} \frac{p(A \in a + \varepsilon \mathcal{A}, B \in \mathcal{B})}{\varepsilon^N m(\mathcal{A})} \tag{11}$$

which are probability density with respect to one variable, and probability with respect to the other variable. Assume that the conditional probability formula

$$p(A \in \mathcal{A} | B \in \mathcal{B}) = \frac{p(A \in \mathcal{A}, B \in \mathcal{B})}{p(B \in \mathcal{B})} \tag{12}$$

is known and prove the formula

$$p(A = a|B \in \mathcal{B}) = \frac{p(A = a, B \in \mathcal{B})}{p(B \in \mathcal{B})} \tag{13}$$

where the conditional density has been defined with a limit $\varepsilon \to 0$ similarly as in Equation (10).

2.2 Assume that B is a continuous random variable, while A can be either continuous or discrete. Singular conditions are defined without normalization factors:

$$p(A \in \mathcal{A}|B = b) = \lim_{\varepsilon \to 0} p(A \in \mathcal{A}|B \in b + \varepsilon\mathcal{B}) \tag{14}$$

where $\mathcal{B}$ is again some small set (such as a ball) containing the origin. Prove the formula

$$p(A \in \mathcal{A}|B = b) = \frac{p(A \in \mathcal{A}, B = b)}{p(B = b)}. \tag{15}$$

Advice: The essential is that you get the normalization factors right. The point of the exercise is not to focus on convergence issues or real analysis.

2.3 Assume that A is a continuous random variable, and B a discrete random variable, and that their distributions depend on some variable θ. Then suppose some observations a and b are available. How would you define the log-likelihood $\ell(\theta; a, b)$? (There is no strictly correct answer to this, since there are at least two possible definitions.) (In any case) Write the log-likelihood (also) so that its formula does not involve quantities which are simultaneously probability and probability density.

Exercise 3:

The EM-algorithm is a method to estimate parameters θ when you can only observe a subgroup $\mathbf{x}$ of all the variables $(\mathbf{x}, \mathbf{s})$ in the model. In other words, given the statistical model $p_{\mathbf{X},\mathbf{S}}(\mathbf{x}, \mathbf{s}; \theta)$ you want to estimate θ from the observations $\mathbf{x}(1), \ldots, \mathbf{x}(T)$, while not having any observations $\mathbf{s}(1), \ldots, \mathbf{s}(T)$ available. The s_i in the vector $\mathbf{s}$ are called latent variables.

We denote as $\mathbf{X}(t)$ and $\mathbf{S}(t)$ the random variables (random vectors), and as $\mathbf{x}(t)$ and $\mathbf{s}(t)$ some sample points. $\mathbf{X}$ and $\mathbf{S}$ are random matrices, whose columns are the random vectors $\mathbf{X}(t)$ and $\mathbf{S}(t)$, while $\mathbf{x}$ and $\mathbf{s}$ are ordinary matrices, whose columns contain the sample points.

If we were able to observe both $\mathbf{x}$ and $\mathbf{s}$, our full log-likelihood would be

$$\ell_{\text{full}}(\theta; \mathbf{x}, \mathbf{s}) = \log\left(p_{\mathbf{X},\mathbf{S}}(\mathbf{x}, \mathbf{s}; \theta)\right) \tag{16}$$

which we would maximize to find θ.

Since $\mathbf{s}$ are not known now, one solution for this problem would be to integrate out the latent variables to obtain $p_{\mathbf{X}}(\mathbf{x};\theta)$ and then to maximize the marginal likelihood

$$\ell_{\text{marg.}}(\theta;\mathbf{x}) = \log\big(p_{\mathbf{X}}(\mathbf{x};\theta)\big) = \log\Big(\int p_{\mathbf{X},\mathbf{S}}(\mathbf{x},\mathbf{s};\theta)d\mathbf{s}\Big) \tag{17}$$

in order to find θ. The EM-algorithm offers an alternative solution.

Assume that we have an initial estimate of θ available, call it θ_0. Then we can calculate a posterior

$$p_{\mathbf{S}}(\mathbf{s}|\mathbf{X}=\mathbf{x};\theta_0) = \frac{p_{\mathbf{X},\mathbf{S}}(\mathbf{x},\mathbf{s};\theta_0)}{p_{\mathbf{X}}(\mathbf{x};\theta_0)} \tag{18}$$

Even though we don't know the $\mathbf{s}$ precisely, we know something through this posterior distribution. Next, we replace the full likelihood with an expected (estimated) likelihood

$$\ell_{\text{estim.}}(\theta;\mathbf{x},\theta_0) = \int \ell_{\text{full}}(\theta;\mathbf{x},\mathbf{s})p_{\mathbf{S}}(\mathbf{s}|\mathbf{X}=\mathbf{x};\theta_0)d\mathbf{s} \tag{19}$$

This can be interpreted as an expectation value

$$\ell_{\text{estim.}}(\theta;\mathbf{x},\theta_0) = \mathbb{E}\big(\ell_{\text{full}}(\theta;\mathbf{x},\mathbf{S})\big) \tag{20}$$

if the expectation is defined with respect to the posterior distribution of $\mathbf{S}$ with the condition $\mathbf{X}=\mathbf{x}$ and the parameter θ_0.

This is called the E-step in the EM-algorithm. Maximization of $\ell_{\text{estim.}}(\theta;\mathbf{x},\theta_0)$, which is called the M-step, yields then a new estimate θ_1 for θ. We can then proceed iteratively and use a new posterior $p_{\mathbf{S}}(\mathbf{s}|\mathbf{X}=\mathbf{x},\theta_1)$ to obtain a new estimated log-likelihood and so on. The goal of this exercise is to show that the EM-iteration described above leads to estimate sequence $\theta_1,\theta_2,\theta_3,\ldots$ with the property $\ell_{\text{marg.}}(\theta_k;\mathbf{x}) \le \ell_{\text{marg.}}(\theta_{k+1};\mathbf{x})$ for all k.

3.1 The estimate θ_{k+1} is defined as a maximum of the mapping $\theta \mapsto \ell_{\text{estim.}}(\theta;\mathbf{x},\theta_k)$. Explain why in particular $\ell_{\text{estim.}}(\theta_k;\mathbf{x},\theta_k) \le \ell_{\text{estim.}}(\theta_{k+1};\mathbf{x},\theta_k)$.

3.2 We assume that the data $(\mathbf{x}(1),\mathbf{s}(1)),\ldots,(\mathbf{x}(T),\mathbf{s}(T))$ is iid. This means that with fixed indices t and t' such that $t \neq t'$ the sample points $(\mathbf{x}(t),\mathbf{s}(t))$ and $(\mathbf{x}(t'),\mathbf{s}(t'))$ are independent and from identical distributions. We don't assume that that $\mathbf{x}(t)$ and $\mathbf{s}(t)$ would be independent though. Use this iid assumption to show that the estimated likelihood can be written as

$$\begin{aligned}&\ell_{\text{estim.}}(\theta;\mathbf{x},\theta_k)\\ &= \sum_{t=1}^{T}\int \log\big(p_{\mathbf{X}(t),\mathbf{S}(t)}(\mathbf{x}(t),\mathbf{s}(t);\theta)\big)\; p_{\mathbf{S}(t)}(\mathbf{s}(t)|\mathbf{X}(t)=\mathbf{x}(t);\theta_k)d\mathbf{s}(t)\end{aligned} \tag{21}$$

$\Big($Or with simplified notation:

$$\ell_{\text{estim.}}(\theta; \mathbf{x}, \theta_k) = \sum_{t=1}^{T} \int \log\big(p(\mathbf{x}(t), \mathbf{s}(t); \theta)\big)\, p(\mathbf{s}(t)|\mathbf{x}(t); \theta_k) d\mathbf{s}(t) \quad \Big) \tag{22}$$

3.3 The Kullback-Leibler divergence of two functions is defined as

$$D_{\text{KL}}(f, g) = \int \log\left(\frac{f(x)}{g(x)}\right) f(x) dx \tag{23}$$

Prove that

$$\begin{aligned} &\ell_{\text{estim.}}(\theta_{k+1}; \mathbf{x}, \theta_k) - \ell_{\text{estim.}}(\theta_k; \mathbf{x}, \theta_k) \\ &= -\sum_{t=1}^{T} D_{\text{KL}}\big(p_{\mathbf{S}(t)}(\cdot|\mathbf{x}(t); \theta_k), p_{\mathbf{S}(t)}(\cdot|\mathbf{x}(t), \theta_{k+1})\big) + \ell_{\text{marg.}}(\theta_{k+1}; \mathbf{x}) - \ell_{\text{marg.}}(\theta_k; \mathbf{x}) \end{aligned} \tag{24}$$

3.4 Use the known fact that the Kullback-Leibler divergence is always non-negative to arrive at the conclusion:

$$\ell_{\text{marg.}}(\theta_0|\mathbf{x}) \leq \ell_{\text{marg.}}(\theta_1|\mathbf{x}) \leq \ell_{\text{marg.}}(\theta_2|\mathbf{x}) \leq \cdots \tag{25}$$

Advice: Exercise 3.2 is related to Exercise 1, and it shouldn't be possible to accomplish the given task without using the knowledge that

$$\int p(\mathbf{s}(t)|\mathbf{x}(t); \theta_k) d\mathbf{s}(t) = 1 \tag{26}$$

in intermediate steps. It is recommended that you also take a closer look at the quantities $p(\mathbf{s}(t)|\mathbf{x}; \theta_k)$ and $p(\mathbf{s}(t)|\mathbf{x}(t); \theta_k)$. In 3.3 you must begin from the left side of the equation, substitute the known formulas into estimated likelihoods, and simplify the expression with properties of logarithm. Then use the definition of conditional density $p(\mathbf{x}(t), \mathbf{s}(t); \theta_k) = p(\mathbf{s}(t)|\mathbf{x}(t); \theta_k) p(\mathbf{x}(t); \theta_k)$ (and same for θ_{k+1}) to proceed further.

Exercise 4:

In this exercise, we first set up a statistical model to do clustering and then use the EM-algorithm to estimate the parameters in the model. This material is treated in Section 12.7.

The random variables are $\mathbf{X}(1), \ldots, \mathbf{X}(T)$ and $R(1), \ldots, R(T)$. We denote the possible values by $\mathbf{x}(1), \ldots, \mathbf{x}(T) \in \mathbb{R}^N$ and $r(1), \ldots, r(T) \in \{1, 2, \ldots, K\}$, where $K \in \mathbb{N}$ is the number of clusters. The random variables $(\mathbf{X}(1), R(1)), (\mathbf{X}(2), R(2))$,

$\ldots, (\mathbf{X}(T), R(T))$ are iid. This means that the variable $(\mathbf{X}(t), R(t))$ is independent from the variable $(\mathbf{X}(t'), R(t'))$ when $t \neq t'$. Variables $\mathbf{X}(t)$ and $R(t)$ are not independent. $\mathbf{X}$ without t-index should be interpreted as an $(N \times T)$-matrix and R as a $(1 \times T)$-vector. (Random matrix and random vector.) The probabilities are given by

$$\begin{aligned} &p_{\mathbf{X}(t)}\big(\mathbf{x}(t)|R(t) = r(t); \theta\big) \\ &= \frac{1}{(2\pi)^{N/2}\det(C_{r(t)})^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}(t) - \boldsymbol{\mu}_{r(t)})^T C_{r(t)}^{-1}(\mathbf{x}(t) - \boldsymbol{\mu}_{r(t)})\Big) \end{aligned} \tag{27}$$

and

$$P\big(R(t) = r(t); \theta\big) = \pi_{r(t)}. \tag{28}$$

The parameter θ is a parameter containing the other parameters as $\theta = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, C_1, \ldots, C_K)$. Here $\pi_k \in \mathbb{R}$ are numbers such that $\pi_k \geq 0$ for all k, and $\sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\mu}_k \in \mathbb{R}^N$ are vectors, and $C_k \in \mathbb{R}^{N\times N}$ are covariance matrices (symmetric and positive definite). We also denote

$$\begin{aligned} &p(\mathbf{X}(t) = \mathbf{x}(t), R(t) = r(t); \theta) \\ &= p_{\mathbf{X}(t)}\big(\mathbf{x}(t)|R(t) = r(t); \theta\big)\ P\big(R(t) = r(t); \theta\big) \\ &= \frac{\pi_{r(t)}}{(2\pi)^{N/2}\det(C_{r(t)})^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}(t) - \boldsymbol{\mu}_{r(t)})^T C_{r(t)}^{-1}(\mathbf{x}(t) - \boldsymbol{\mu}_{r(t)})\Big) \end{aligned} \tag{29}$$

By the iid assumption we have

$$p\big(\mathbf{X} = \mathbf{x}, R = r; \theta\big) = \prod_{t=1}^{T} p\big(\mathbf{X}(t) = \mathbf{x}(t), R(t) = r(t); \theta\big) \tag{30}$$

The locations of the sample points $\mathbf{x}$ are known observations, while the cluster indices r are unknown latent variables (like $\mathbf{s}$ in the previous exercise). Next, our goal is to estimate θ using the observations $\mathbf{x}$.

4.1 Given data $(\mathbf{x}(1) \ldots \mathbf{x}(T))$ and $(r(1) \ldots r(T))$ write a formula for the full log-likelihood $\ell_{\text{full}}(\theta; \mathbf{x}, r)$.

4.2 Show that the posterior $p(R(t) = k|\mathbf{X}(t) = \mathbf{x}(t); \theta)$ is given by

$$\begin{aligned} &p(R(t) = k|\mathbf{X}(t) = \mathbf{x}(t); \theta) \\ &= \frac{\frac{\pi_k}{\sqrt{\det(C_k)}} \exp\big(-\frac{1}{2}(\mathbf{x}(t) - \boldsymbol{\mu}_k)^T C_k^{-1}(\mathbf{x}(t) - \boldsymbol{\mu}_k)\big)}{\sum_{k'=1}^{K} \frac{\pi_{k'}}{\sqrt{\det(C_{k'})}} \exp\big(-\frac{1}{2}(\mathbf{x}(t) - \boldsymbol{\mu}_{k'})^T C_{k'}^{-1}(\mathbf{x}(t) - \boldsymbol{\mu}_{k'})\big)} \end{aligned} \tag{31}$$

Advice: In a sense we only need the definition of a conditional probability, but this might be a good point to recall the point of exercise 2.2 and Equation (15).

4.3 Given an estimate θ_i, find a formula for the estimated log-likelihood $\ell_{\text{estim.}}(\theta; \mathbf{x}, \theta_i)$.

Advice: Recall the definition of the estimated log-likelihood from the exercise 3, and replace the old parameter $\mathbf{s}$ with r. To find the correct modification to the old integral $\int d\mathbf{s}$, recall summing notation from the exercise 1. A recommended possibility is that first the parameter r goes through all of its values when it is summed out, and that means in total K^T different terms. Then, the expression can be simplified by summing away some part. Alternatively, you can attempt to skip to a simplified result by using a result from the exercise 3. The answer will involve the expression $p(R(t) = r(t)|\mathbf{X}(t) = \mathbf{x}(t); \theta_i)$. Into this do not substitute the known formula, because it turns out that its details are not immediately needed.

4.4 Calculate the gradients $\nabla_{\boldsymbol{\mu}_k} \ell_{\text{estim.}}(\theta; \mathbf{x}; \theta_i)$ and $\nabla_{C_k} \ell_{\text{estim.}}(\theta; \mathbf{x}; \theta_i)$. Set the gradients to zero and obtain the following EM-update rules for $\boldsymbol{\mu}_k$ and C_k, $k = 1, \ldots, K$:

$$\boldsymbol{\mu}_k(i+1) = \frac{\sum_{t=1}^{T} P(R(t) = k|\mathbf{X}(t) = \mathbf{x}(t); \theta_i)\mathbf{x}(t)}{\sum_{t'=1}^{T} P(R(t') = k|\mathbf{X}(t') = \mathbf{x}(t'); \theta_i)} \tag{32}$$

$$C_k(i+1) = \frac{\sum_{t=1}^{T} P(R(t) = k|\mathbf{X}(t) = \mathbf{x}(t); \theta_i)(\mathbf{x}(t) - \boldsymbol{\mu}_k(i+1))(\mathbf{x}(t) - \boldsymbol{\mu}_k(i+1))^T}{\sum_{t'=1}^{T} P(R(t') = k|\mathbf{X}(t') = \mathbf{x}(t'); \theta_i)} \tag{33}$$

Advice: Notice that you must compute gradients with respect to quantities that are found in the parameter θ, while the quantities in the parameter θ_i are constants during this procedure. The results obtained in Set 3 (10.4.2015) Ex 1 and 2 (and possibly 3) can be reused here. Some Kronecker deltas between k and $r(t)$ might turn out useful too.

4.5 The optimization of $\ell_{\text{estim.}}(\theta; \mathbf{x}, \theta_i)$ with respect to the distribution of R, i.e. the weights π_k, is trickier because it is a constrained optimization problem: $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. There is a trick to convert the constrained optimization problem into an unconstrained one: Write π_k as

$$\pi_k = \frac{\exp(\gamma_k)}{\sum_{k'=1}^{K} \exp(\gamma_{k'})}, \tag{34}$$

where $\gamma_k \in \mathbb{R}$. First, verify that the trick works, i.e. that π_k as defined above satisfies the constraints for all γ_i. Then, find the derivative $\partial_{\gamma_k} \ell_{\text{estim.}}(\theta; \mathbf{x}, \theta_i)$, and

set it to zero to find the EM-update rule for π_k:

$$\pi_k(i+1) = \frac{1}{T}\sum_{t=1}^{T} P(R(t) = k|\mathbf{X}(t) = \mathbf{x}(t); \theta_i) \tag{35}$$