UML2015
Solutions and other comments
to some exercises from sets 4 and 6

Set 4 Exercise 1:
1.1

$$\mathbb{E}(\mathcal{Y}_1\mathcal{Y}_2) \approx \frac{1}{K}\sum_{k=1}^{K}(\mathbf{y}_1)_k(\mathbf{y}_2)_k = \frac{1}{K}\mathbf{y}_1 \cdot \mathbf{y}_2 \tag{1}$$

Assume that A is some real random variable, and then assume that $(A_1, A_2, \ldots, A_K)$ is iid each A_k having the same distribution as A. If $(a_1, a_2, \ldots, a_K)$ is a sample draw, an approximation

$$\mathbb{E}(A) \approx \frac{1}{K}\sum_{k=1}^{K} a_k \tag{2}$$

holds. This is true for all random variables for which the mean exist, and in particular for $A = \mathcal{Y}_1\mathcal{Y}_2$ too.

If $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are uncorrelated, then $\mathbb{E}(\mathcal{Y}_1\mathcal{Y}_2) = 0$. This will not imply that that the vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ would be orthogonal, but since it does imply $\frac{1}{K}\mathbf{y}_1 \cdot \mathbf{y}_2 \approx 0$, we can say that the vectors are approximately orthogonal in this sense. The point of the exercise is to understand that theoretical uncorrelatedness of $\mathcal{Y}_1$ and $\mathcal{Y}_2$ implies approximate orthogonality for $\mathbf{y}_1$ and $\mathbf{y}_2$.

1.2

$$\frac{1}{K}\overline{XX}^T = \frac{1}{K}(U^TX)(X^TU) = U^T\underbrace{\left(\frac{1}{K}XX^T\right)U}_{=U\Lambda} = \underbrace{U^TU}_{=\mathrm{id}}\Lambda = \Lambda \tag{3}$$

where Λ is an $M \times M$ diagonal matrix containing the most significant eigenvalues of $\frac{1}{K}XX^T$ on its diagonal.

$$\implies \quad \overline{X}_{m*} \cdot \overline{X}_{m'*} = (\overline{XX}^T)_{mm'} = 0 \quad \forall m \neq m' \tag{4}$$

Here we used the knowledge that eigenvectors (here U_{*m}) of a symmetric matrix (here $\frac{1}{K}XX^T$) can always be assumed to be orthogonal. Notice that we did not assume that U would have fully diagonalized the $\frac{1}{K}XX^T$ (which is $N \times N$). The calculation works under the assumption that some of the eigenvectors were placed into the rows of U.

A little mistake: Eigenvectors of a symmetric matrix are always orthogonal if the corresponding eigenvalues are distinct. The eigenvectors are not necessarily orthogonal, if some of the eigenvalues appear multiple times (some eigenvalues have degree greater than 1). The eigenvectors are never unique, and even if some

eigenvalues appear multiple times, the eigenvectors can be chosen so that they are orthogonal. In this exercise, it was not mentioned that the eigenvalues of $\frac{1}{K}XX^T$ would be distinct, and also it was not mentioned that the columns of U would be assumed orthogonal. This can be seen as a mistake in the exercise, and it should have been clarified that the columns are assumed to be orthogonal so that $U^TU = \mathrm{id}$.

If X was a real data sample, it would be extremely unlikely that some of the eigenvalues of $\frac{1}{K}XX^T$ would be the same, so in this sense the mistake was very small.

Further comments to the explanation request in 1.1. If $\mathbf{X}$ was a zero mean $N \times 1$ random vector, whose elements $X_1, \ldots, X_N$ were not uncorrelated, it would be nice, if we could find a matrix U with orthogonal columns such that a new random vector $\overline{\mathbf{X}} = U^T\mathbf{X}$ did have uncorrelated elements so that $\mathbb{E}(\overline{X}_m\overline{X}_{m'}) = 0$ for all $m \neq m'$. It is not possible to find this kind of U based on any $N \times K$ sample X due to the randomness of the sample points. However, we learned that it is possible to find U such that the rows of U^TX were precisely orthogonal. In doing this, we found U such that the elements of $\overline{\mathbf{X}} = U^T\mathbf{X}$ are approximately uncorrelated so that $\mathbb{E}(\overline{X}_m\overline{X}_{m'}) \approx 0$ for all $m \neq m'$. So the relationship of uncorrelatedness and orthogonality works in this direction too.

Set 4 Exercise 2:

2.1 This exercise deals with basics of linear algebra. We assume that A is a real $N \times M$ matrix and $M < N$, and that the columns of A are linearly independent. The given task is to prove *that AA^T has M positive eigenvalues, and zero appears as the eigenvalues $N - M$ times.* The trick to this is that we don't attempt to solve the non-zero eigenvalues, and also don't solve any eigenvectors either, but we seek to prove that $\dim(\ker(AA^T)) = N - M$, which will be sufficient to accomplish the given task. Notice the obvious: $\mathbf{x} \in \ker(AA^T)$ is equivalent with $AA^T\mathbf{x} = 0$, which is equivalent with $AA^T\mathbf{x} = \lambda\mathbf{x}$ for $\lambda = 0$. Hence the kernel contains the eigenvectors corresponding to the eigenvalue zero and nothing else (ignoring the origin). By considering the diagonalization of AA^T we see that the order of the eigenvalue zero (how many times it appears on the diagonal after diagonalization) is the same as the dimension of the kernel.

First we seek to prove that

$$\ker(AA^T) = \ker(A^T) \tag{5}$$

One possibility to (5): First we prove that $\ker(A) = \{\mathbf{0}\}$. So we want to prove

$$A\mathbf{y} = \mathbf{0} \quad \Longleftrightarrow \quad \mathbf{y} = \mathbf{0}. \tag{6}$$

Notice that $A\mathbf{y}$ is the same thing as $\sum_{m=1}^{M} A_{*m} y_m$, which is a linear combination of the columns of A with real coefficients y_m. By definition of the linear independence, this kind of linear combination cannot be zero with non-zero coefficients, so $\ker(A) = \{\mathbf{0}\}$ is clear. Now we know that

$$(AA^T)\mathbf{x} = \mathbf{0} \quad \Longleftrightarrow \quad A(A^T\mathbf{x}) = \mathbf{0} \quad \Longleftrightarrow \quad A^T\mathbf{x} = \mathbf{0}. \tag{7}$$

which is the same thing as (5).

Second possibility to (5): The direction

$$AA^T\mathbf{x} = \mathbf{0} \quad \Longleftarrow \quad A^T\mathbf{x} = \mathbf{0} \tag{8}$$

is obvious, so it sufficies to prove the other direction. We assume $AA^T\mathbf{x} = \mathbf{0}$. Then also $0 = \mathbf{x}^T(AA^T\mathbf{x}) = (\mathbf{x}^T A)(A^T\mathbf{x}) = \|A^T\mathbf{x}\|^2$, which implies $A^T\mathbf{x} = \mathbf{0}$.

Once (5) has been proven, we seek to prove

$$\dim \ker(A^T) = N - M \tag{9}$$

next. By the rank-nullity theorem equation

$$N = \dim \operatorname{im}(A^T) + \dim \ker(A^T) \tag{10}$$

holds. Therefore it will be sufficient to prove

$$\dim \operatorname{im}(A^T) = M. \tag{11}$$

We make an anti-thesis that $\dim \operatorname{im}(A^T) < M$ would hold. Another basic result is that if $V \subset \mathbb{R}^M$ is some vector space, $V + V^\perp = \mathbb{R}^M$ holds where $V^\perp$ is the orthogonal complement of V, and the sum of the dimensions of V and $V^\perp$ equals the M. The anti-thesis $\dim \operatorname{im}(A^T) < M$ will then imply that the orthogonal complement of the image must have a dimension greater than or equal to 1: $\dim(\operatorname{im}(A^T))^\perp \geq 1$. This implies that in particular $(\operatorname{im}(A^T))^\perp \neq \{\mathbf{0}\}$, and at least one non-zero vector $\mathbf{x} \in (\operatorname{im}(A^T))^\perp$ can be found. In other words, the anti-thesis implies that there exists a vector $\mathbf{x} \in \mathbb{R}^M \setminus \{0\}$ with the property, that it is orthogonal to all vectors in the $\operatorname{im}(A^T)$. Then, in particular it will be orthogonal to the columns of A^T (which belong to the image), and

$$0 = \mathbf{x} \cdot (A^T)_{*n} \; \forall 1 \leq n \leq N \quad \Longrightarrow \quad \mathbf{0} = \mathbf{x}^T A^T \quad \Longrightarrow \quad \mathbf{0} = A\mathbf{x} \tag{12}$$

and this contradicts the assumption that the columns of A are linearly independent.

We have now proven both equations (5) and (9), and together they imply

$$\dim \ker(AA^T) = N - M. \tag{13}$$

This in turn implies that the matrix AA^T has M non-zero eigenvalues. The exercise requested a proof for a claim that AA^T has M positive eigenvalues. To accomplish this, we must prove that negative eigenvalues cannot exists. Assume $\mathbf{x}$ is a normalized eigenvector of AA^T with an eigenvalue λ. Then

$$\lambda = \lambda\|\mathbf{x}\|^2 = \mathbf{x}^T(\lambda\mathbf{x}) = \mathbf{x}^T AA^T\mathbf{x} = \|A^T\mathbf{x}\|^2 \geq 0, \tag{14}$$

so we see that the eigenvalues of AA^T can never be negative.

2.2 The answer is 2^N.

2.3

$$\overline{A} = A(A^TA)^{-\frac{1}{2}} \tag{15}$$

$$\overline{A}^T\overline{A} = \big((A^TA)^{-\frac{1}{2}}A^T\big)\big(A(A^TA)^{-\frac{1}{2}}\big) = (A^TA)^{-\frac{1}{2}}\underbrace{A^TA(A^TA)^{-\frac{1}{2}}}_{=(A^TA)^{\frac{1}{2}}} = \mathrm{id} \tag{16}$$

Set 6 Exercise 1:
1.1: We first check the case $M = 2$.

$$\begin{aligned}
\Big(\prod_{m=1}^{2}\sum_{a_m=1}^{D}\Big)\Big(\prod_{m'=1}^{2} f_{m'}(a_{m'})\Big) &= \sum_{a_2=1}^{D}\sum_{a_1=1}^{D}\big(f_1(a_1)f_2(a_2)\big) \\
&= \sum_{a_2=1}^{D}\Big(\sum_{a_1=1}^{D}\big(f_1(a_1)f_2(a_2)\big)\Big) \\
&= \sum_{a_2=1}^{D}\Big(f_2(a_2)\sum_{a_1=1}^{D} f_1(a_1)\Big) \\
&= \Big(\sum_{a_1=1}^{D} f_1(a_1)\Big)\Big(\sum_{a_2=1}^{D} f_2(a_2)\Big) \\
&= \prod_{m=1}^{2}\Big(\sum_{a_m=1}^{D} f_m(a_m)\Big)
\end{aligned} \tag{17}$$

We write the sum as an iterated sum so that the sum over a_1 is inside, and sum over a_2 is outside. In the inner sum we see that $f_2(a_2)$ is a constant with respect to the sum over a_1, and hence it can be taken in front of the sum. Then the entire sum $\sum_{a_1=1}^{D} f_1(a_1)$ is a constant with respect to the outer sum over a_2, and hence this can be taken in front as a constant too.

For arbitrary M the result can be proven with an induction step.

$$\begin{aligned}
\Big(\prod_{m=1}^{M}\sum_{a_m=1}^{D}\Big)\Big(\prod_{m'=1}^{M} f_{m'}(a_{m'})\Big) &= \sum_{a_M=1}^{D}\cdots\sum_{a_1=1}^{D}\big(f_1(a_1)\cdots f_M(a_M)\big) \\
&= \sum_{a_M=1}^{D}\cdots\sum_{a_2=1}^{D}\Big(\sum_{a_1=1}^{D}\big(f_1(a_1)f_2(a_2)\cdots f_M(a_M)\big)\Big) \\
&= \sum_{a_M=1}^{D}\cdots\sum_{a_2=1}^{D}\Big(f_2(a_2)\cdots f_M(a_M)\sum_{a_1=1}^{D} f_1(a_1)\Big) \\
&= \Big(\sum_{a_1=1}^{D} f_1(a_1)\Big)\Big(\sum_{a_M=1}^{D}\cdots\sum_{a_2=1}^{D} f_2(a_2)\cdots f_M(a_M)\Big) \qquad (18) \\
&= \Big(\sum_{a_1=1}^{D} f_1(a_1)\Big)\Big(\Big(\prod_{m=2}^{M}\sum_{a_m=1}^{D}\Big)\Big(\prod_{m'=2}^{M} f_{m'}(a_{m'})\Big)\Big) \\
&= \Big(\sum_{a_1=1}^{D} f_1(a_1)\Big)\Big(\prod_{m=2}^{M}\Big(\sum_{a_m=1}^{D} f_m(a_m)\Big)\Big) \\
&= \prod_{m=1}^{M}\Big(\sum_{a_m=1}^{D} f_m(a_m)\Big)
\end{aligned}$$

We write the sum as an iterated sum, placing the sum with respect to a_1 inside. Then $f_2(a_2)\cdots f_M(a_M)$ is a constant with respect to the sum, and can be taken in front. Then the sum $\sum_{a_1=1}^{D} f_1(a_1)$ is a constant with respect to all other sums, so it can be taken in front too. Then we use the induction assumption, that the result is known for $M-1$ dimensional sum, and finally compose the product into one product over m.

Set 6 Exercise 2:

2.3: The obvious answer is

$$\ell(\theta; a, b) = \log\big(p(A = a, B = b; \theta)\big) \qquad (19)$$

where the $p(A = a, B = b; \theta)$ is probability density with respect to a, and probability with respect to b. To the second request the intended answer is

$$\ell(\theta; a, b) = \log\big(p_A(a|B = b; \theta)\big) + \log\big(P(B = b; \theta)\big) \qquad (20)$$

Here $p_A(a|B = b; \theta)$ is probability density, and $P(B = b; \theta)$ is probability, so there are no mixed probability quantities present.

Set 6 Exercise 4:
4.1 The answer is:

$$\begin{aligned}\ell_{\text{full}}(\theta;\mathbf{x},r) = -\frac{1}{2}\sum_{t=1}^{T}\Big(\log\ \det\ C_{r(t)} + (\mathbf{x}(t)-\boldsymbol{\mu}_{r(t)})^T C_{r(t)}^{-1}(\mathbf{x}(t)-\boldsymbol{\mu}_{r(t)})\Big) \\ + \sum_{t=1}^{T}\log\pi_{r(t)} + \text{const.}\end{aligned} \tag{21}$$

4.3

$$\begin{aligned}\ell_{\text{estim.}}(\theta;\mathbf{x},\theta_i) &= \sum_{r\in\{1,2,\ldots,K\}^T}\ell_{\text{full}}(\theta;\mathbf{x},r)p(R=r|\mathbf{X}=\mathbf{x};\theta_i) \\ &= \sum_{t=1}^{T}\sum_{r\in\{1,2,\ldots,K\}^T}\log\big(p(\mathbf{x}(t),r(t);\theta)\big)p(R=r|\mathbf{X}=\mathbf{x};\theta_i) \\ &= \sum_{t=1}^{T}\Big(\prod_{t'=1}^{T}\sum_{r(t')=1}^{K}\Big)\log\big(p(\mathbf{x}(t),r(t);\theta)\big)\Big(\prod_{t''=1}^{T}p(r(t'')|\mathbf{x}(t'');\theta_i)\Big) \\ &= \sum_{t=1}^{T}\Big(\sum_{r(t)=1}^{K}\log\big(p(\mathbf{x}(t),r(t);\theta)\big)p(r(t)|\mathbf{x}(t);\theta_i)\Big) \\ &\quad \Big(\prod_{\substack{t''=1\\t''\neq t}}^{T}\Big(\underbrace{\sum_{r(t'')=1}^{K}p(r(t'')|\mathbf{x}(t'');\theta_i)}_{=1}\Big)\Big) \\ &= \sum_{t=1}^{T}\sum_{r(t)=1}^{K}\log\big(p(\mathbf{x}(t),r(t);\theta)\big)p(r(t)|\mathbf{x}(t);\theta_i)\end{aligned} \tag{22}$$

First we write the definition of the estimated log-likelihood, which is otherwise the same as definition in Equation (19) in the original exercise sheet, except that the integral over $\mathbf{s}$ has been replaced with a sum over r. In the sum r goes over all possible vectors $r = (r(1), r(2), \ldots, r(T))$, where each $r(t)$ can obtain values from the set $\{1, 2, \ldots, K\}$. The total amount of terms in this sum is K^T. We don't immediately substitute the full formula of ℓ_{full}, because it would make the equations unnecessarily complicated. Instead we only substitute the iid assumption, and take the product out of the logarithm, making it a sum. Then the order of sums can be changed. Then we use the iid assumption to $p(R = r|\mathbf{X} = \mathbf{x};\theta_i)$ and denote

the sum over r in the same way as in the first exercise of set 6. Then with fixed t the product is arranged like

$$\Big(\log \big(p(\mathbf{x}(t), r(t); \theta)\big) p(r(t)|\mathbf{x}(t); \theta_i) \Big) \Big(\prod_{\substack{t''=1 \\ t'' \neq t}}^{T} p(r(t'')|\mathbf{x}(t''); \theta_i) \Big) \tag{23}$$

and the result from first exercise is used to change the order of sums and products. In the end we can substitute the known formula for $\log(p(\mathbf{x}(t), r(t); \theta))$ and get

$$\begin{aligned} \ell_{\text{estim.}}(\theta; \mathbf{x}, \theta_i) = \sum_{t=1}^{T} \sum_{r(t)=1}^{K} \Big(&- \frac{1}{2} \log \det C_{r(t)} - \frac{1}{2} (\mathbf{x}(t) - \boldsymbol{\mu}_{r(t)})^T C_{r(t)}^{-1} (\mathbf{x}(t) - \boldsymbol{\mu}_{r(t)}) \\ &+ \log \pi_{r(t)} \Big) p(r(t)|\mathbf{x}(t); \theta_i) + \text{const.} \end{aligned} \tag{24}$$

This is precisely the same thing as

$$\begin{aligned} \ell_{\text{estim.}}(\theta; \mathbf{x}, \theta_i) = \sum_{t=1}^{T} \sum_{k=1}^{K} \Big(&- \frac{1}{2} \log \det C_k - \frac{1}{2} (\mathbf{x}(t) - \boldsymbol{\mu}_k)^T C_k^{-1} (\mathbf{x}(t) - \boldsymbol{\mu}_k) \\ &+ \log \pi_k \Big) p(R(t) = k|\mathbf{x}(t); \theta_i) + \text{const.} \end{aligned} \tag{25}$$

which is perhaps clearer.

4.4 The most difficult part of exercise 4 of set 6 was probably getting the sums right in 4.3. Once the correct formula has been found for $\ell_{\text{estim.}}$, its gradient can be computed with the same formulas that were used earlier in the exercise set 3. The final place where to avoid mistakes is to remember, that if the expression of $\ell_{\text{estim.}}$ contains a sum over k, then you must not use the same index for partial derivatives. So if you want to apply $\nabla_{\boldsymbol{\mu}_k}$ and ∇_{C_k} on $\ell_{\text{estim.}}$, then the sum must be replaced into a sum over k' for example (or keep $r(t)$ which was the original index). The Kronecker delta $\delta_{kk'}$ (or $\delta_{k,r(t)}$) will appear in very trivial manner in 4.4, but in 4.5 it is not so trivial.

4.5

$$\pi_k = \frac{e^{\gamma_k}}{\sum_{k'=1}^{K} e^{\gamma_{k'}}} \tag{26}$$

$$\ell_{\text{estim.}}(\theta; \mathbf{x}, \theta_i) = \sum_{t=1}^{T} \sum_{k=1}^{K} \log \Big(\frac{e^{\gamma_k}}{\sum_{k'=1}^{K} e^{\gamma_{k'}}} \Big) p\big(R(t) = k \mid \mathbf{x}(t); \theta_i\big) + \text{const.} \tag{27}$$

Here we have denoted as constants all terms not depending on $\pi_1, \ldots, \pi_K$.

$$\frac{\partial \ell_{\text{estim.}}(\theta; \mathbf{x}, \theta_i)}{\partial \gamma_m} = \sum_{t=1}^{T} \sum_{k=1}^{K} \Big(\frac{\partial}{\partial \gamma_m} \log \Big(\frac{e^{\gamma_k}}{\sum\limits_{k'=1}^{K} e^{\gamma_{k'}}} \Big) \Big) p\big(R(t) = k \mid \mathbf{x}(t); \theta_i\big) \tag{28}$$

We take a closer look at the derivative inside:

$$\begin{aligned} \frac{\partial}{\partial \gamma_m} \log \Big(\frac{e^{\gamma_k}}{\sum\limits_{k'=1}^{K} e^{\gamma_{k'}}} \Big) &= \frac{1}{\pi_k} \frac{\partial}{\partial \gamma_m} \frac{e^{\gamma_k}}{\sum\limits_{k'=1}^{K} e^{\gamma_{k'}}} \\ &= \frac{1}{\pi_k} \Big(\frac{e^{\gamma_k} \delta_{km}}{\sum\limits_{k'=1}^{K} e^{\gamma_{k'}}} - \frac{e^{\gamma_k} e^{\gamma_m}}{\Big(\sum\limits_{k'=1}^{K} e^{\gamma_{k'}} \Big)^2} \Big) \\ &= \delta_{km} - \pi_m \end{aligned} \tag{29}$$

Here the relation (26) was used multiple times. We set the partial derivative as zero:

$$\begin{aligned} 0 = \frac{\partial \ell_{\text{estim.}}(\theta; \mathbf{x}, \theta_i)}{\partial \gamma_m} &= \sum_{t=1}^{T} \sum_{k=1}^{K} (\delta_{km} - \pi_m) p\big(R(t) = k \mid \mathbf{x}(t);\ \theta_i\big) \\ &= \sum_{t=1}^{T} p\big(R(t) = m \mid \mathbf{x}(t);\ \theta_i\big) - \pi_m \underbrace{\sum_{t=1}^{T} \overbrace{\sum_{k=1}^{K} p\big(R(t) = k \mid \mathbf{x}(t); \theta_i\big)}^{=1}}_{=T} \end{aligned} \tag{30}$$

The π_m can be solved out of this, and this is how the iteration formula is found.