Homework 4

JD Gallego Posada, *University of Amsterdam* 09/05/2017

Problem 1

Collaborator: D Kianfar

1. First consider the Lagrangian for this problem given by:

$$\mathcal{V} = \sum_n \sum_k \gamma(z_{nk}) \left\{ \log \pi_k + \log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} + \lambda \left(\sum_k \pi_k - 1 \right)$$

Now, optimizing with respect to the mixture coefficients,

$$\frac{\partial \mathcal{V}}{\partial \pi_k} = \sum_n \frac{\gamma(z_{nk})}{\pi_k} + \lambda = 0 \Rightarrow \pi_k = \frac{\sum_n \gamma(z_{nk})}{-\lambda} =: \frac{N_k}{-\lambda}$$

$$-\lambda \sum_k \pi_k = \sum_k \sum_n \gamma(z_{nk}) = \sum_k N_k \Rightarrow \lambda = -N$$

Which gives the desired result $\pi_k = \frac{N_k}{N}$.

Note that the only terms depending on the parameters of the Gaussians is:

$$\log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = -\frac{D}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

Now, optimizting with respect to $\boldsymbol{\mu}_k$:

$$\begin{aligned}
\frac{\partial \mathcal{V}}{\partial \boldsymbol{\mu}_k} &= \sum_n \gamma(z_{nk}) \frac{\partial \log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k} \\
&= \sum_n -\frac{1}{2} \gamma(z_{nk}) \frac{\partial}{\partial (\mathbf{x}_n - \boldsymbol{\mu}_k)} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \frac{\partial (\mathbf{x}_n - \boldsymbol{\mu}_k)}{\partial \boldsymbol{\mu}_k} \\
&= \sum_n -\frac{1}{2} \gamma(z_{nk}) 2 \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)(-\mathbb{I}) \\
&= \sum_n \gamma(z_{nk}) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0
\end{aligned}$$

Solving for $\boldsymbol{\mu}_k$ gives:

$$\boldsymbol{\Sigma}_k^{-1} \sum_n \gamma(z_{nk}) \mathbf{x}_n = \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \sum_n \gamma(z_{nk}) \Rightarrow \boldsymbol{\mu}_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) \mathbf{x}_n$$

Finally, optimizing with respect to $\boldsymbol{\Sigma}_k$, and using the results 57 and 61 from the Matrix Cookbook,

$$\begin{aligned}
\frac{\partial \mathcal{V}}{\partial \boldsymbol{\Sigma}_k} &= \sum_n \gamma(z_{nk}) \frac{\partial \log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\Sigma}_k} \\
&= \sum_n -\frac{1}{2} \gamma(z_{nk}) \left[\frac{\partial \log |\boldsymbol{\Sigma}_k|}{\partial \boldsymbol{\Sigma}_k} + \frac{\partial}{\partial \boldsymbol{\Sigma}_k} (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right] \\
&= \sum_n -\frac{1}{2} \gamma(z_{nk}) \left[\boldsymbol{\Sigma}_k^{-T} - \boldsymbol{\Sigma}_k^{-T} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-T} \right] \\
&= -\frac{1}{2} \left[\boldsymbol{\Sigma}_k^{-T} \sum_n \gamma(z_{nk}) - \boldsymbol{\Sigma}_k^{-T} \left(\sum_n \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \right) \boldsymbol{\Sigma}_k^{-T} \right] = 0
\end{aligned}$$

From which we get:

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_n \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

2. The final part of the derivation changes to:

$$
\begin{aligned}
\frac{\partial \mathcal{V}}{\partial \boldsymbol{\Sigma}} &= \sum_k \sum_n \gamma(z_{nk}) \frac{\partial \log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}_k} \\
&= \sum_k \sum_n -\frac{1}{2}\gamma(z_{nk}) \left[\frac{\partial \log |\boldsymbol{\Sigma}|}{\partial \boldsymbol{\Sigma}_k} + \frac{\partial}{\partial \boldsymbol{\Sigma}_k}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\right] \\
&= \sum_k \sum_n -\frac{1}{2}\gamma(z_{nk}) \left[\boldsymbol{\Sigma}^{-T} - \boldsymbol{\Sigma}^{-T}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-T}\right] \\
&= -\frac{1}{2}\left[\boldsymbol{\Sigma}^{-T} \sum_k \sum_n \gamma(z_{nk}) - \boldsymbol{\Sigma}^{-T} \left(\sum_k \sum_n \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\right) \boldsymbol{\Sigma}^{-T}\right] = 0
\end{aligned}
$$

From which it is easy to conclude that:

$$
\boldsymbol{\Sigma} = \frac{1}{N} \sum_k \sum_n \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T
$$

Problem 2

First note that $\log p(\boldsymbol{\theta}|\mathbf{X}) = \log p(\boldsymbol{\theta}, \mathbf{X}) - \log p(\mathbf{X})$ and $\log p(\boldsymbol{\theta}, \mathbf{X}) = \log p(\mathbf{X}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$.

$$
\begin{aligned}
\log p(\boldsymbol{\theta}|\mathbf{X}) &= \log p(\boldsymbol{\theta}, \mathbf{X}) - \log p(\mathbf{X}) \\
&= \log p(\mathbf{X}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathbf{X}) \\
&= \mathcal{L}(q, \boldsymbol{\theta}) + \mathbb{KL}(q||p) + \log p(\boldsymbol{\theta}) - \log p(\mathbf{X}) \\
&\geq \mathcal{L}(q, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathbf{X})
\end{aligned}
$$

Therefore, the E-step remains the same as before since q only appears in $\mathcal{L}$. After the E-step, the lower bound for the M-step has the form:

$$
\mathcal{L}(q, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{const} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{const}
$$

Problem 3

From Problem 2, in the M-step we want to optimize $\mathbb{E}_{\mathbf{Z}}\left[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\right] + \log p(\boldsymbol{\theta})$. In this case, this has the form:

$$
\Psi = \sum_n \sum_k \gamma(z_{nk}) \left\{\log \pi_k + \sum_i x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki})\right\} + \sum_k \log \text{Beta}(\boldsymbol{\mu}_k | a_k, b_k) + \log \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})
$$

The final two terms in Ψ can be rewritten as:

$$
\Psi = \ldots + \sum_k \sum_i (a_k - 1) \log \mu_{ki} + (b_k - 1) \log(1 - \mu_{ki}) + f(a_k, b_k) + \sum_k (\alpha_k - 1) \log \pi_k + g(\alpha_k)
$$

Now, optimizing with respect to μ_{ki} we have:

$$
\frac{\partial \Psi}{\partial \mu_{ki}} = \sum_n \gamma(z_{nk}) \left(\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}}\right) + \frac{a_k - 1}{\mu_{ki}} - \frac{b_k - 1}{1 - \mu_{ki}} = 0
$$

$$
\begin{aligned}
\frac{1}{\mu_{ki}}\left(\sum_n \gamma(z_{nk})x_{ni} + a_k - 1\right) &= \frac{1}{1-\mu_{ki}}\left(\sum_n \gamma(z_{nk})(1-x_{ni}) + b_k - 1\right) \\
(1-\mu_{ki})\left(\sum_n \gamma(z_{nk})x_{ni} + a_k - 1\right) &= \mu_{ki}\left(\sum_n \gamma(z_{nk})(1-x_{ni}) + b_k - 1\right) \\
\sum_n \gamma(z_{nk})x_{ni} + a_k - 1 \cancel{-\mu_{ki}\sum_n \gamma(z_{nk})x_{ni}} - \mu_{ki}(a_k - 1) &= \cancel{-\mu_{ki}\sum_n \gamma(z_{nk})x_{ni}} + \mu_{ki}\sum_n \gamma(z_{nk}) + \mu_{ki}(b_k - 1) \\
\sum_n \gamma(z_{nk})x_{ni} + a_k - 1 &= \mu_{ki}(a_k - 1 + N_k + b_k - 1)
\end{aligned}
$$

From which we can conclude:

$$
\mu_{ki} = \frac{\sum_n \gamma(z_{nk})x_{ni} + a_k - 1}{N_k + a_k + b_k - 2}
$$

Recall that the optimization with respect to the mixing coefficient implies the use of a Lagrange multiplier to ensure the constraint $\sum_k \pi_k = 1$. Thus,

$$
\frac{\partial}{\partial \pi_k}\left[\Psi + \lambda\left(\sum_k \pi_k - 1\right)\right] = \sum_n \frac{\gamma(z_{nk})}{\pi_k} + \frac{\alpha_k - 1}{\pi_k} + \lambda = 0 \Rightarrow \pi_k = \frac{\sum_n \gamma(z_{nk}) + \alpha_k - 1}{-\lambda} = \frac{N_k + \alpha_k - 1}{-\lambda}
$$

$$
-\lambda \sum_k \pi_k = \sum_k \sum_n \gamma(z_{nk}) + \alpha_k - 1 = N + \sum_k \alpha_k - K \Rightarrow -\lambda = N + \sum_k \alpha_k - K
$$

Which gives the desired result

$$
\pi_k = \frac{N_k + \alpha_k - 1}{N + \sum_k \alpha_k - K}.
$$