Exercises for the tutorials: 1(a-d).

The other exercises are for self-study and exam preparation. All material is examinable unless otherwise mentioned.

Exercise 1. *EM algorithm for mixture models*

Mixture models are statistical models of the form

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x};\boldsymbol{\theta}_k) \tag{1}$$

where each $p_k(\mathbf{x};\boldsymbol{\theta}_k)$ is itself a statistical model parameterised by $\boldsymbol{\theta}_k$ and the $\pi_k \geq 0$ are mixture weights that sum to one. The parameters $\boldsymbol{\theta}$ of the mixture model consist of the parameters $\boldsymbol{\theta}_k$ of each mixture component and the mixture weights π_k, i.e. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \pi_1, \dots, \pi_K)$. An example is a mixture of Gaussians where each $p_k(\mathbf{x};\boldsymbol{\theta}_k)$ is a Gaussian with parameters given by the mean vector $\boldsymbol{\mu}_k$ and a covariance matrix $\boldsymbol{\Sigma}_k$.

The mixture model in (1) can be considered to be the marginal distribution of a latent variable model $p(\mathbf{x}, h;\boldsymbol{\theta})$ where h is an unobserved variable that takes on values $1, \dots, K$ and $p(h = k) = \pi_k$. Defining $p(\mathbf{x}|h = k;\boldsymbol{\theta}) = p_k(\mathbf{x};\boldsymbol{\theta}_k)$, the latent variable model corresponding to (1) thus is

$$p(\mathbf{x}, h = k;\boldsymbol{\theta}) = p(\mathbf{x}|h = k;\boldsymbol{\theta})p(h = k) = \pi_k p_k(\mathbf{x};\boldsymbol{\theta}_k). \tag{2}$$

In particular note that marginalising out h gives $p(\mathbf{x};\boldsymbol{\theta})$ in (1).

(a) *Verify that the latent variable model in (2) can be written as*

$$p(\mathbf{x}, h;\boldsymbol{\theta}) = \prod_{k=1}^{K} \left[\pi_k p_k(\mathbf{x};\boldsymbol{\theta}_k)\right]^{\mathbb{1}(h=k)} \tag{3}$$

where h takes values in $1, \dots, K$.

Solution. Since $\mathbb{1}(h = k)$ is one if $h = k$ and zero otherwise, we have

$$p(\mathbf{x}, h = j;\boldsymbol{\theta}) = \prod_{k=1}^{K} \left[\pi_k p_k(\mathbf{x};\boldsymbol{\theta}_k)\right]^{\mathbb{1}(j=k)} = \pi_j p_j(\mathbf{x};\boldsymbol{\theta}_j) \tag{S.1}$$

for any $j \in \{1, \dots, K\}$, which matches (2).

(b) *Since the mixture model in (1) can be seen as the marginal of a latent-variable model, we can use the expectation maximisation (EM) algorithm to estimate the parameters $\boldsymbol{\theta}$.*

For a general model $p(\mathcal{D}, \mathbf{h};\boldsymbol{\theta})$ where $\mathcal{D}$ are the observed data and $\mathbf{h}$ the corresponding unobserved variables, the EM algorithm iterates between computing the expected complete-data log-likelihood $J^l(\boldsymbol{\theta})$ and maximising it with respect to $\boldsymbol{\theta}$:

$$\textbf{\textit{E-step at iteration l:}} \quad J^l(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}^l)}[\log p(\mathcal{D}, \mathbf{h};\boldsymbol{\theta})] \tag{4}$$

$$\textbf{\textit{M-step at iteration l:}} \quad \boldsymbol{\theta}^{l+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} J^l(\boldsymbol{\theta}) \tag{5}$$

Here $\boldsymbol{\theta}^l$ is the value of $\boldsymbol{\theta}$ in the l-th iteration. When solving the optimisation problem, we also need to take into account constraints on the parameters, e.g. that the π_k correspond to a pmf.

Assume that the data $\mathcal{D}$ consists of n iid data points $\mathbf{x}_i$, that each $\mathbf{x}_i$ has associated with it a scalar unobserved variable h_i, and that the tuples $(\mathbf{x}_i, h_i)$ are all iid. What is $J^l(\boldsymbol{\theta})$ under these additional assumptions?

Solution. Since the $(\mathbf{x}_i, h_i)$ are iid, we have that $p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{x}_i, h_i; \boldsymbol{\theta})$. Hence

$$
\begin{aligned}
J^l \boldsymbol{\theta}) &= \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}^l)}[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})] && \text{(S.2)}\\
&= \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}^l)}\left[\sum_{i=1}^n \log p(\mathbf{x}_i, h_i; \boldsymbol{\theta})\right] && \text{(S.3)}\\
&= \sum_{i=1}^n \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}^l)}[\log p(\mathbf{x}_i, h_i; \boldsymbol{\theta})] && \text{(S.4)}\\
&= \sum_{i=1}^n \mathbb{E}_{p(h_i|\mathcal{D};\boldsymbol{\theta}^l)}[\log p(\mathbf{x}_i, h_i; \boldsymbol{\theta})] && \text{(S.5)}\\
&= \sum_{i=1}^n \mathbb{E}_{p(h_i|\mathbf{x}_i;\boldsymbol{\theta}^l)}[\log p(\mathbf{x}_i, h_i; \boldsymbol{\theta})] && \text{(S.6)}
\end{aligned}
$$

where in the second last step, we have used that each $\log p(\mathbf{x}_i, h_i; \boldsymbol{\theta})]$ only involves one latent variable h_i so that we only need to take the expectation over $p(h_i|\mathcal{D}; \boldsymbol{\theta}^l)$, and in the last step, we have used that $h_i \perp\!\!\!\perp \mathbf{x}_j$, for $j \neq i$.

(c) *Show that for the latent variable model in (3), $J^l(\boldsymbol{\theta})$ equals*

$$
\begin{aligned}
J^l(\boldsymbol{\theta}) &= \sum_{i=1}^n \sum_{k=1}^K w_{ik}^l \log[\pi_k p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)], && (6)\\
w_{ik}^l &= \frac{\pi_k^l p_k(\mathbf{x}_i; \boldsymbol{\theta}_k^l)}{\sum_{k=1}^K \pi_k^l p_k(\mathbf{x}_i; \boldsymbol{\theta}_k^l)} && (7)
\end{aligned}
$$

Note that the w_{ik}^l are defined in terms of the parameters π_k^l and $\boldsymbol{\theta}_k^l$ from iteration l. They are equal to the conditional probabilities $p(h = k|\mathbf{x}_i; \boldsymbol{\theta}^l)$, i.e. the probability that $\mathbf{x}_i$ has been sampled from component $p_k(\mathbf{x}_i; \boldsymbol{\theta}_k^l)$.

Solution. We consider a single term $\mathbb{E}_{p(h|\mathbf{x};\boldsymbol{\theta}^l)}[\log p(\mathbf{x}, h; \boldsymbol{\theta})]$ in (S.6).
Given the form of the model in (3), we have that

$$
\log p(\mathbf{x}, h; \boldsymbol{\theta}) = \sum_{k=1}^K \mathbb{1}(h = k) \log[\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)] \tag{S.7}
$$

and hence

$$
\begin{aligned}
\mathbb{E}_{p(h|\mathbf{x};\boldsymbol{\theta}^l)}[\log p(\mathbf{x}, h; \boldsymbol{\theta})] &= \mathbb{E}_{p(h|\mathbf{x};\boldsymbol{\theta}^l)}\left[\sum_{k=1}^K \mathbb{1}(h = k) \log[\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)]\right] && \text{(S.8)}\\
&= \sum_{k=1}^K \mathbb{E}_{p(h|\mathbf{x};\boldsymbol{\theta}^l)}\left[\mathbb{1}(h = k)\right] \log[\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)] && \text{(S.9)}\\
&= \sum_{k=1}^K p(h = k|\mathbf{x}; \boldsymbol{\theta}^l) \log[\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)] && \text{(S.10)}
\end{aligned}
$$

where we have used that the expectation over an indicator event equals the probability for the event to happen, i.e. $\mathbb{E}_{p(h|\mathbf{x};\boldsymbol{\theta}^l)}\left[\mathbb{1}(h = k)\right] = p(h = k|\mathbf{x}; \boldsymbol{\theta}^l)$.

The probability $p(h = k|\mathbf{x}; \boldsymbol{\theta}^l)$ can be determined via the product (Bayes') rule and Equations (2) and (1)

$$p(h = k|\mathbf{x}; \boldsymbol{\theta}^l) = \frac{p(\mathbf{x}, h = k, \boldsymbol{\theta}^l)}{p(\mathbf{x}; \boldsymbol{\theta}^l)} \tag{S.11}$$

$$= \frac{\pi_k^l p_k(\mathbf{x}; \boldsymbol{\theta}_k^l)}{\sum_{k=1}^K \pi_k^l p_k(\mathbf{x}; \boldsymbol{\theta}_k^l)} \tag{S.12}$$

Note that the superscript ^{l} indicates that the π_k^l are the mixture weights and the $\boldsymbol{\theta}_k^l$ the model parameters from iteration l.

The objective $J^l(\boldsymbol{\theta})$ sums over n terms $\mathbb{E}_{p(h|\mathbf{x}_i;\boldsymbol{\theta}^l)}[\log p(\mathbf{x}_i, h; \boldsymbol{\theta})]$. Let us denote $p(h = k|\mathbf{x}_i; \boldsymbol{\theta}^l)$ from (S.12) by w_{ik}^l so that

$$\mathbb{E}_{p(h|\mathbf{x}_i;\boldsymbol{\theta}^l)}[\log p(\mathbf{x}_i, h; \boldsymbol{\theta})] = \sum_{k=1}^K w_{ik}^l \log[\pi_k p_k(\mathbf{x}; \boldsymbol{\theta}_k)] \tag{S.13}$$

and

$$J^l(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K w_{ik}^l \log[\pi_k p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)]. \tag{S.14}$$

The objective $J^l(\boldsymbol{\theta})$ takes the form of a weighted log-likelihood. In more detail, since $\sum_k w_{ik}^l = 1$ for all data points $\mathbf{x}_i$ (and $w_{ik}^l \geq 0$), $\sum_{k=1}^K w_{ik}^l \log[\pi_k p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)]$ is a convex combination. This means that the different components of the mixture model compete with each other: larger weights for some components mean smaller weights for others. In the extreme case, some components may contribute in a negligible way to the i-th term of the log-likelihood.

The weights w_{ik}^l are sometimes, in particular for mixture of Gaussians, called "soft-assignments" because they specify to which extent a data points $\mathbf{x}_i$ "belongs" to a mixture component p_k. Alternatively, we can interpret the w_{ik}^l to be the "responsibilities" of each mixture component p_k for a datapoint $\mathbf{x}_i$.

In some cases, e.g. for computational reasons, we may determine which of the K weights $w_{i1}^l, \ldots, w_{iK}^l$ is the largest and then set it to one while setting the other weights to zero. This corresponds to "hard-assignments" (and "hard EM") where a data point $\mathbf{x}_i$ is exclusively assigned to a single mixture component p_k.

(d) *Assume that the different mixture components $p_k(\mathbf{x}; \boldsymbol{\theta}_k), k = 1, \ldots, K$ do not share any parameters. Show that the updated parameter values $\boldsymbol{\theta}_k^{l+1}$ are given by weighted maximum likelihood estimates.*

Solution. We interchange the order of the summations in (6) so that

$$J^l(\boldsymbol{\theta}) = \sum_{k=1}^K \sum_{i=1}^n w_{ik}^l \log[\pi_k p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)] \tag{S.15}$$

$$= \sum_{k=1}^K \sum_{i=1}^n w_{ik}^l \log \pi_k + \sum_{k=1}^K \underbrace{\sum_{i=1}^n w_{ik}^l \log p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)}_{\ell_k^l(\boldsymbol{\theta}_k)} \tag{S.16}$$

When we update the parameters $\boldsymbol{\theta}_k$ of the mixture components, the first term is a constant. The second term is a sum over weighted log-likelihoods $\ell_k^l(\boldsymbol{\theta}_k)$, one for each mixture component. If the mixture components do not share parameters, we thus have

$$\boldsymbol{\theta}_k^{l+1} = \underset{\boldsymbol{\theta}_k}{\operatorname{argmax}}\, J^l(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}_k}{\operatorname{argmax}}\, \ell_k^l(\boldsymbol{\theta}_k) \tag{S.17}$$

This means that we can compute $\boldsymbol{\theta}_k^{l+1}$ as if we performed maximum likelihood estimation for the model $p_k(\mathbf{x}; \boldsymbol{\theta}_k)$, expect that the data points $\mathbf{x}_i$ are weighted by the w_{ik}^l.

(e) *Show that maximising $J^l(\boldsymbol{\theta})$ with respect to the mixture weights π_k gives the update rule*

$$\pi_k^{l+1} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}^l \tag{8}$$

Solution. We start with (6) and drop additive terms that do not depend on the π_k. Since

$$J^l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^l \log \pi_k + \text{terms not depending on the } \pi_k \tag{S.18}$$

we can focus on the objective

$$J_\pi^l(\pi_1, \ldots, \pi_K) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^l \log \pi_k \tag{S.19}$$

$$= \sum_{k=1}^{K} \underbrace{\left(\sum_{i=1}^{n} w_{ik}^l \right)}_{\omega_k^l} \log \pi_k \tag{S.20}$$

$$= \sum_{k=1}^{K} \omega_k^l \log \pi_k. \tag{S.21}$$

Taking into account that the $\pi_k = p(h = k)$ define a pmf, the optimisation problem to solve is

$$\text{maximise} \quad \sum_{k=1}^{K} \omega_k^l \log \pi_k \tag{S.22}$$

$$\text{subject to} \quad \pi_k \geq 0 \tag{S.23}$$

$$\sum_{k=1}^{K} \pi_k = 1 \tag{S.24}$$

The constrained optimisation problem could be solved via Lagrange multipliers. But we here take another approach and solve the optimisation problem by phrasing it in terms of a KL-divergence minimisation problem.

First, note that the π_k that maximise $J_\pi^l(\pi_1, \ldots, \pi_K)$ will also maximise the re-scaled objective

$$\frac{1}{\sum_{k=1}^{K} \omega_k^l} J_\pi^l(\pi_1, \ldots, \pi_K) = \frac{1}{\sum_{k=1}^{K} \omega_k^l} \sum_{k=1}^{K} \omega_k^l \log \pi_k \tag{S.25}$$

$$= \sum_{k=1}^{K} q_k^l \log \pi_k \tag{S.26}$$

where we introduced

$$q_k^l = \frac{\omega_k^l}{\sum_{k=1}^K \omega_k^l}. \tag{S.27}$$

The q_k^l are non-negative and sum to one. Hence, we can consider them to define a pmf. Second, note that the π_k that maximise $J_\pi^l(\pi_1, \ldots, \pi_K)$ will also maximise

$$\sum_{k=1}^K q_k^l \log \pi_k - \sum_{k=1}^K q_k^l \log q_k^l = \sum_{k=1}^K q_k^l \log \frac{\pi_k}{q_k^l} \tag{S.28}$$

$$= -\sum_{k=1}^K q_k^l \log \frac{q_k^l}{\pi_k} \tag{S.29}$$

$$= -\mathrm{KL}(q^l, \pi) \tag{S.30}$$

since adding constants does not change the solution. Hence, the optimal π_k are given by the pmf π that minimises the KL-divergence $\mathrm{KL}(q^l, \pi)$. This means that the optimal π_k are

$$\pi_k = q_k^l = \frac{\omega_k^l}{\sum_{k=1}^K \omega_k^l} = \frac{\sum_{i=1}^n w_{ik}^l}{\sum_{k=1}^K \sum_{i=1}^n w_{ik}^l}. \tag{S.31}$$

The denominator can be simplified by noting that, with (7), $\sum_{k=1}^K w_{ik}^l = 1$ so that

$$\sum_{k=1}^K \sum_{i=1}^n w_{ik}^l = \sum_{i=1}^n \sum_{k=1}^K w_{ik}^l = n \tag{S.32}$$

The requested update rule thus is

$$\pi_k^{l+1} = \frac{1}{n} \sum_{i=1}^n w_{ik}^l \tag{S.33}$$

The update rule does not depend directly on the statistical model $p_k(\mathbf{x}; \boldsymbol{\theta}_k)$ that we may choose for the mixture components. Their influence occurs indirectly via the w_{ik}^l.

(f) *Summarise the EM-algorithm to learn the parameters $\boldsymbol{\theta}$ of the mixture model in (1) from iid data $\mathbf{x}_1, \ldots, \mathbf{x}_n$.*

Solution. We collect and summarise the results from the previous questions:

- **E-step at iteration l:** Compute the posterior probabilities (soft assignments)

$$w_{ik}^l = \frac{\pi_k^l p_k(\mathbf{x}_i; \boldsymbol{\theta}_k^l)}{\sum_{k=1}^K \pi_k^l p_k(\mathbf{x}_i; \boldsymbol{\theta}_k^l)} \tag{S.34}$$

  for all data points $\mathbf{x}_i$ and and mixture components k. Then formulate the objective function $J^l(\boldsymbol{\theta})$

$$J^l(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K w_{ik}^l \log[\pi_k p_k(\mathbf{x}_i; \boldsymbol{\theta}_k)] \tag{S.35}$$

- **M-step at iteration l:** Compute the new mixture weights

$$\pi_k^{l+1} = \frac{1}{n}\sum_{i=1}^{n} w_{ik}^l \tag{S.36}$$

To compute the new mixture parameters $\boldsymbol{\theta}_k^{l+1}$, maximise $J^l(\boldsymbol{\theta})$ if some parameters are shared or tied. If the $p_k(\mathbf{x}; \boldsymbol{\theta}_k)$ do not share parameters, the new parameters $\boldsymbol{\theta}_k^{l+1}$ are obtained by maximising a weighted log-likelihood for each mixture component separately:

$$\boldsymbol{\theta}_k^{l+1} = \underset{\boldsymbol{\theta}_k}{\operatorname{argmax}} \sum_{i=1}^{n} w_{ik}^l \log p_k(\mathbf{x}_i; \boldsymbol{\theta}_k) \tag{S.37}$$

for $k = 1, \ldots, K$.

Exercise 2. ***EM algorithm for mixture of Gaussians***

We here use the results from Exercise 1 to derive the EM update rules for a mixture of Gaussians. This is a mixture model where each mixture component is a Gaussian distribution, i.e.

$$p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{9}$$

We consider the case where each $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ can be individually changed (no tying of parameters). The overall parameters of the model are given by the $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ and the mixture weights $\pi_k \geq 0$, $k = 1, \ldots K$. As in the case of general mixture models, the mixture weights sum to one.

(a) *Determine the maximum likelihood estimates for a multivariate Gaussian $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for iid data $\mathcal{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ when each data point $\mathbf{x}_i$ has a weight w_i. The weights are non-negative but do not necessarily sum to one.*

Solution. The weighted log-likelihood is

$$\begin{aligned} \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{i=1}^{n} w_i \log \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) && \text{(S.38)} \\ &= \sum_{i=1}^{n} w_i \log |\det 2\pi\boldsymbol{\Sigma}|^{-1/2} - \frac{1}{2}\sum_{i=1}^{n} w_i (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) && \text{(S.39)} \end{aligned}$$

Introducing the normalised weights $W_i = w_i / \sum_{i=1}^{n} w_i$, we have

$$\frac{1}{\sum_{i=1}^{n} w_i}\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log |\det 2\pi\boldsymbol{\Sigma}|^{-1/2} - \frac{1}{2}\sum_{i=1}^{n} W_i (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \tag{S.40}$$

Let us write out the quadratic term

$$(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_i - 2\mathbf{x}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{S.41}$$

Hence

$$\sum_{i=1}^n W_i(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) = \sum_{i=1}^n W_i\mathbf{x}_i^\top\boldsymbol{\Sigma}^{-1}\mathbf{x}_i - 2\sum_{i=1}^n W_i\mathbf{x}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \underbrace{\sum_{i=1}^n W_i}_{=1}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{S.42}$$

$$= \operatorname{tr}\left[\left(\sum_{i=1}^n W_i\mathbf{x}_i\mathbf{x}_i^\top\right)\boldsymbol{\Sigma}^{-1}\right] - 2\left(\sum_{i=1}^n W_i\mathbf{x}_i\right)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{S.43}$$

$$= \operatorname{tr}\left(\mathbf{R}\boldsymbol{\Sigma}^{-1}\right) - 2\mathbf{b}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{S.44}$$

where $\mathbf{R} = \sum_{i=1}^n W_i\mathbf{x}_i\mathbf{x}_i^\top$ and $\mathbf{b} = \sum_{i=1}^n W_i\mathbf{x}_i$. Hence

$$\frac{1}{\sum_{i=1}^n w_i}\ell(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \log|\det 2\pi\boldsymbol{\Sigma}|^{-1/2} - \frac{1}{2}\operatorname{tr}\left(\mathbf{R}\boldsymbol{\Sigma}^{-1}\right) + \mathbf{b}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{S.45}$$

This has exactly the same form as the unweighted likelihood function, just the sufficient statistics $\mathbf{R}$ and $\mathbf{b}$ are computed using the weights. Hence, the maximum likelihood estimates, when expressed in terms of $\mathbf{R}$ and $\mathbf{b}$ remain the same as in the unweighted case:

$$\hat{\boldsymbol{\mu}} = \mathbf{b} = \sum_{i=1}^n W_i\mathbf{x}_i \tag{S.46}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{R} - \mathbf{b}\mathbf{b}^\top = \sum_{i=1}^n W_i\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{b}\mathbf{b}^\top \tag{S.47}$$

Moreover, since

$$\sum_{i=1}^n W_i(\mathbf{x}_i-\mathbf{b})(\mathbf{x}_i-\mathbf{b})^\top = \sum_{i=1}^n W_i\mathbf{x}_i\mathbf{x}_i^\top - \underbrace{\sum_{i=1}^n W_i\mathbf{x}_i}_{\mathbf{b}}\mathbf{b}^\top - \mathbf{b}\underbrace{\sum_{i=1}^n W_i\mathbf{x}_i^\top}_{\mathbf{b}^\top} + \mathbf{b}\mathbf{b}^\top \tag{S.48}$$

$$= \mathbf{R} - \mathbf{b}\mathbf{b}^\top - \mathbf{b}\mathbf{b}^\top + \mathbf{b}\mathbf{b}^\top \tag{S.49}$$

$$= \mathbf{R} - \mathbf{b}\mathbf{b}^\top \tag{S.50}$$

we find that the weighted maximum likelihood estimates are the weighted average and weighted covariance matrix:

$$\hat{\boldsymbol{\mu}} = \sum_{i=1}^n W_i\mathbf{x}_i \qquad \hat{\boldsymbol{\Sigma}} = \sum_{i=1}^n W_i(\mathbf{x}_i-\hat{\boldsymbol{\mu}})(\mathbf{x}_i-\hat{\boldsymbol{\mu}})^\top \qquad W_i = \frac{w_i}{\sum_{i=1}^n w_i} \tag{S.51}$$

(b) *Use the results from Exercise 1 to derive the EM update rules for the parameters of the Gaussian mixture model.*

Solution. From the solution to Exercise 1(f) and the derived weighted MLE solutions, we find:

- **E-step at iteration l:** Compute the posterior probabilities (soft assignments)

$$w_{ik}^l = \frac{\pi_k^l\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_k^l,\boldsymbol{\Sigma}_k^l)}{\sum_{k=1}^K \pi_k^l\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_k^l,\boldsymbol{\Sigma}_k^l)} \tag{S.52}$$

  for all data points $\mathbf{x}_i$ and and mixture components k.

- **M-step at iteration l:**
  - Determine the weighted MLEs

$$\boldsymbol{\mu}_k^{l+1} = \sum_{i=1}^{n} W_{ik}^l \mathbf{x}_i \qquad \boldsymbol{\Sigma}_k^{l+1} = \sum_{i=1}^{n} W_{ik}^l (\mathbf{x}_i - \boldsymbol{\mu}_k^{l+1})(\mathbf{x}_i - \boldsymbol{\mu}_k^{l+1})^\top \tag{S.53}$$

    where $W_{ik}^l = w_{ik}^l / (\sum_{i=1}^n w_{ik}^l)$.
  - Compute the new mixture weights

$$\pi_k^{l+1} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}^l \tag{S.54}$$