Expectation maximization

Course of Machine Learning
Master Degree in Computer Science
University of Rome "Tor Vergata"
a.a. 2024-2025

Giorgio Gambosi

1 The case of a treatable $p(\mathrm{z}|\mathrm{x})$ and the EM algorithm

In the case of hypothesis 2 holding,[1] that is if the conditional probability $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$ is easy to evaluate, then the approach described above results into:

- first computing
$$q^{(k)}(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(k)})$$
- next, deriving
$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(k)}), \mathbf{x}, \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underset{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(k)})}{\mathbb{E}} [\, p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\,]$$

The idea here is to address the maximization of the log-likelihood $\log p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$ of the joint distribution – that is not possible since the value $\mathbf{z}$ of the latent variable is unknown by definition – by referring to the expectation of $p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$ with respect to $\mathbf{z} \sim p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$.

Given a set $\mathbf{X}$ of observations, the method is usually described by the following two steps for each iteration:

Expectation. Given a current value $\boldsymbol{\theta}^{(k)}$ of $\boldsymbol{\theta}$, derive for each observation $\mathbf{x}_i$ the expectation of the joint distribution $p(\mathbf{x}_i,\mathbf{z};\boldsymbol{\theta})$ with respect to $\mathbf{z}$, distributed as $p(\mathbf{z}|\mathbf{x}_i;\boldsymbol{\theta}^{(k)})$: this is a function

$$\underset{p(\mathbf{z}|\mathbf{x}_i;\boldsymbol{\theta}^{(k)})}{\mathbb{E}} [\, p(\mathbf{x}_i,\mathbf{z};\boldsymbol{\theta})\,]$$

of $\boldsymbol{\theta}$

Maximization. Maximize the function $\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)},\mathrm{X}) = \sum_{i=1}^n \mathbb{E}_{p(\mathbf{z}|\mathbf{x}_i;\boldsymbol{\theta}^{(k)})}[\, p(\mathbf{x}_i,\mathbf{z};\boldsymbol{\theta})\,]$ wrt $\boldsymbol{\theta}$, obtaining a new value

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^n \underset{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(k)})}{\mathbb{E}} [\, p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})\,]$$

Such value provides a new conditional distribution $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(i+1)})$ and a new function of $\boldsymbol{\theta}$ to maximize.

$$\mathcal{Q}(\boldsymbol{\theta};\boldsymbol{\theta}^{(i+1)},\mathrm{X}) = \sum_{i=1}^n \underset{p(\mathbf{z}|\mathbf{x}_i;\boldsymbol{\theta}^{(i)})}{\mathbb{E}} [\, \log p(\mathbf{x}_i,\mathbf{z};\boldsymbol{\theta})\,]$$

[1] Observe that in this case the gradient of the log-likelihood can also be evaluated, and a local maximum $\boldsymbol{\theta}^*$ can be computed, making all distributions $p(\mathbf{z}_i|\mathbf{x};\boldsymbol{\theta}^*)$ computable too. However, the EM algorithm introduced here has several advantages wrt gradient methods, such as naturally handling probabilistic constraints and providing a guarantee of convergence. Also, it does not make use of a "step" hyperparameter η, thus avoiding the consequent tuning problem.

The iterative algorithm then starts from any initial value $\boldsymbol{\theta}^{(0)}$ of $\boldsymbol{\theta}$ and performs a sequence of steps, where the k-th step computes $\boldsymbol{\theta}^{(k)}$ from $\boldsymbol{\theta}^{(k-1)}$ by applying the Expectation and the Maximization step in sequence.

We now show that in this case the algorithm monotonically increases (or at least does not decrease) the log-likelihood $\log p(\mathbf{X};\boldsymbol{\theta})$. For simplicity, we will again refer to the case of a single observation $\mathbf{x}$: we already saw how this is immediately extended to the case of a dataset $\mathbf{X}$ with more that one items.

As we know, for any distribution q and parameter value $\boldsymbol{\theta}$, the ELBO decomposition of the log-likelihood holds.

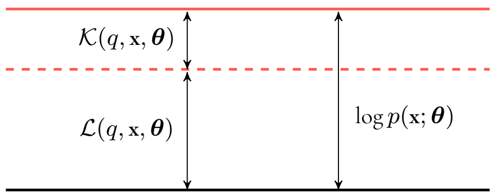

Figure 1: Log-likelihood decomposition

The situation is visualized in Figure 1 where, for a given $\boldsymbol{\theta}$, the gap from the black line to the red line corresponds to the log-likelihood of the observable value, which is independent from the distribution q. The gap between the black and the dashed line (which in any case lies between the black and red ones) corresponds instead to $\mathcal{L}(\boldsymbol{\theta})$ and depends also on the choice of q.

Given $\boldsymbol{\theta}$, setting $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$ provides the maximum lower bound of $\log p(\mathbf{x};\boldsymbol{\theta})$ attainable, since by definition

$$\mathcal{K}(p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}),\mathbf{x},\boldsymbol{\theta}) = 0$$

The k-th step of the iteration includes the following substeps:

E-step

We set $q^*(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(k)})$, obtaining the following situation, sketched in Figure 2,

$$\begin{aligned}
&\mathcal{K}(q^*,\mathbf{x},\boldsymbol{\theta}^{(k)}) = 0\\
&\log p(\mathbf{x};\boldsymbol{\theta}^{(k)}) = \mathcal{L}(q^*,\mathbf{x},\boldsymbol{\theta}^{(k)}) = \mathcal{L}(p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}^{(k)}),\mathbf{x},\boldsymbol{\theta}^{(k)})
\end{aligned}$$

and there is no gap between the blue and red line in Figure 2.

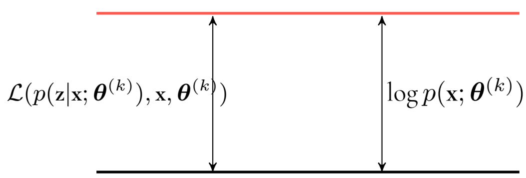

Figure 2: After the E-step

M-step

Since

$$\log p(\mathbf{x};\boldsymbol{\theta}) = \mathcal{L}(q,\mathbf{x},\boldsymbol{\theta}) + \mathcal{K}(q,\mathbf{x},\boldsymbol{\theta})$$

for any distribution q, this is in particular true for the special case when $q(\mathrm{z}) = p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)})$, which implies, in the notation defined above,

$$\log p(\mathrm{x}; \boldsymbol{\theta}) = \mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta}) + \mathcal{K}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta})$$

and the usual lower bound

$$\log p(\mathrm{x}; \boldsymbol{\theta}) \geq \mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \boldsymbol{\theta})$$

Let us consider the maximization of such lower bound with respect to $\boldsymbol{\theta}$.

As already observed, since we may decompose $\mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta})$ as follows

$$\mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta}) = \underset{p(\mathrm{z}|\mathrm{x};\boldsymbol{\theta}^{(k)})}{\mathbb{E}}[\log p(\mathrm{x}, \mathrm{z}; \boldsymbol{\theta})] + \underset{p(\mathrm{z}|\mathrm{x};\boldsymbol{\theta}^{(k)})}{\mathbb{H}}[\mathrm{z}]$$

and since the entropy

$$\underset{p(\mathrm{z}|\mathrm{x};\boldsymbol{\theta}^{(k)})}{\mathbb{H}}[\mathrm{z}]$$

is independent from $\boldsymbol{\theta}$, this is equivalent to maximizing

$$\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \mathrm{x}) = \underset{p(\mathrm{z}|\mathrm{x};\boldsymbol{\theta}^{(k)})}{\mathbb{E}}[\log p(\mathrm{x}, \mathrm{z}; \boldsymbol{\theta})]$$

with respect to $\boldsymbol{\theta}$.

Let us now consider

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \mathrm{x})$$

Since $\boldsymbol{\theta}^{(k+1)}$ is the value of $\boldsymbol{\theta}$ which provides the maximum value for $\mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta})$, we have

$$\mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta}^{(k+1)}) \geq \mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta})$$

for all possible values $\boldsymbol{\theta}$. As a particular case, it holds then that (see Figure 3)

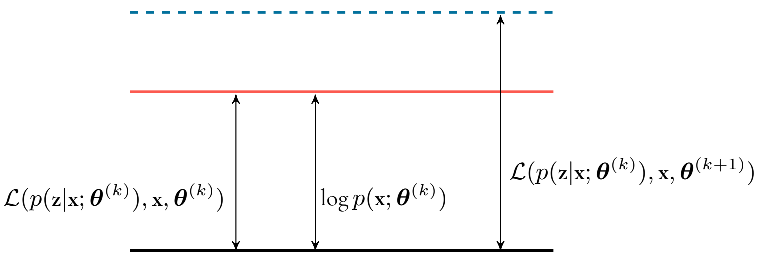

Figure 3: After the M-step

$$\mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta}^{(k+1)}) \geq \mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta}^{(k)}) = \log p(\mathrm{x}; \boldsymbol{\theta}^{(k)})$$

Since in general $p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}) \neq p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k+1)})$, we have $D_{KL}\left(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta})||p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k+1)})\right) > 0$ and, as a consequence, the lower bound is strict, that is (see Figure 4)

$$\log p(\mathrm{x}; \boldsymbol{\theta}^{(k+1)}) > \mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta}^{(k+1)})$$

We may then verify that, after an E-step followed by an M-step, the estimated log-likelihood becomes larger. In particular, it increases from

$$\log p(\mathrm{x}; \boldsymbol{\theta}^{(k)}) = \mathcal{L}(p(\mathrm{z}|\mathrm{x}; \boldsymbol{\theta}^{(k)}), \mathrm{x}, \boldsymbol{\theta}^{(k)})$$

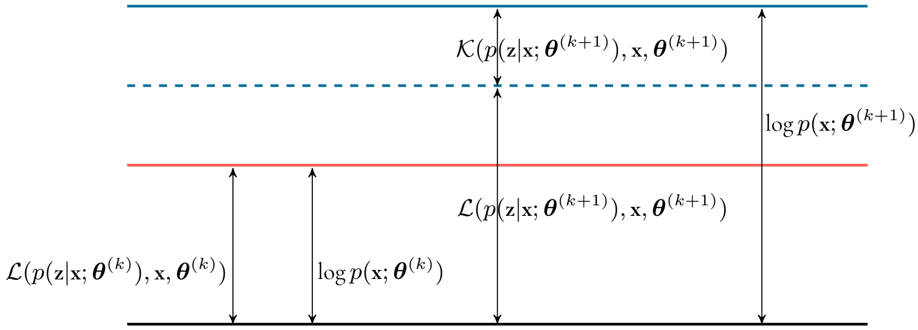

Figure 4: Decomposition of the new log-likelihood

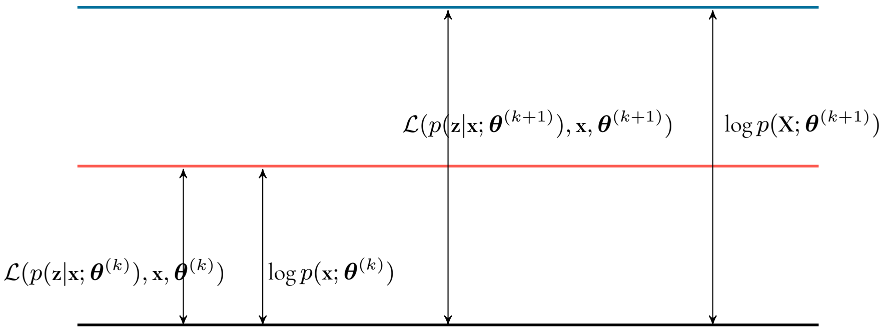

Figure 5: After a new E-step, where $q^*(\mathrm{z}) = p(\mathrm{z}|\mathrm{x};\boldsymbol{\theta}^{(k+1)})$

to

$$\begin{aligned}
\log p(\mathrm{x};\boldsymbol{\theta}^{(k+1)}) &= \mathcal{L}(p(\mathrm{z}|\mathrm{x};\boldsymbol{\theta}^{(k)}),\mathrm{x},\boldsymbol{\theta}^{(k+1)}) + \mathcal{K}(p(\mathrm{z}|\mathrm{x};\boldsymbol{\theta}^{(k)}),\mathrm{x},\boldsymbol{\theta}^{(k+1)}) \\
&\geq \mathcal{L}(p(\mathrm{z}|\mathrm{x};\boldsymbol{\theta}^{(k)}),\mathrm{x},\boldsymbol{\theta}^{(k+1)}) \\
&\geq \mathcal{L}(p(\mathrm{z}|\mathrm{x};\boldsymbol{\theta}^{(k)}),\mathrm{x},\boldsymbol{\theta}^{(k)}) = \log p(\mathrm{x};\boldsymbol{\theta}^{(k)})
\end{aligned}$$

where the last equality is just $\leq$ in the general case.

Mixtures as latent variable models

Discrete mixture models can be seen also as latent variable models where hypothesis 2 holds and the EM algorithm can then be applied.

We remind that in a mixture model the marginal distribution is defined as

$$p(\mathrm{x};\boldsymbol{\pi},\boldsymbol{\Theta}) = \sum_{i=1}^{K} \pi_i q(\mathrm{x};\boldsymbol{\theta}_i)$$

A mixture can be modeled, in terms of latent variables, according to the graphical model in Figure 6, where for each element $\mathbf{x}_i$ a discrete scalar latent variable z_i is introduced with domain $\{1, \ldots, K\}$ which is assumed distributed according to a categorical distribution $p(z) = Cat(z; \boldsymbol{\pi})$, such that $\pi_k = p(z = k)$. We shall denote as $\boldsymbol{\psi}$ the set of all parameters, i.e. $\boldsymbol{\psi} = \boldsymbol{\pi} \cup \boldsymbol{\Theta}$.

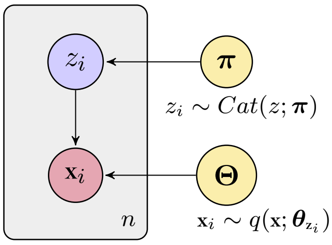

Figure 6: Graphical model of a mixture

By introducing the latent variable $z \in \mathcal{Z} = \{1, \ldots, K\}$, we define the joint distribution

$$p(\mathbf{x}, z; \boldsymbol{\psi}) = p(z; \boldsymbol{\pi}) p(\mathbf{x}|z; \boldsymbol{\theta})$$

The corresponding marginal probability is given by

$$p(\mathbf{x}; \boldsymbol{\psi}) = \sum_{i=1}^{K} p(z = i; \boldsymbol{\pi}) p(\mathbf{x}|z = i; \boldsymbol{\Theta})$$

from which the interpretations $\pi_i = p(z = i; \boldsymbol{\pi})$ and $q(\mathbf{x}; \boldsymbol{\theta}_i) = p(\mathbf{x}|z = i; \boldsymbol{\Theta})$ of the mixture components result.

As we may check, the conditional probability $p(z|\mathbf{x})$ can be computed here, assuming the distributions $q(\mathbf{x}|z)$ can be evaluated. In fact, for $j = 1, \ldots, K$,

$$p(z = j|\mathbf{x}; \boldsymbol{\psi}) = \frac{p(\mathbf{x}|z = j; \boldsymbol{\psi}) p(z = j; \boldsymbol{\psi})}{p(\mathbf{x}; \boldsymbol{\psi})} = \frac{q(\mathbf{x}; \boldsymbol{\theta}_j)\pi_j}{\sum_{r=1}^{K} q(\mathbf{x}; \boldsymbol{\theta}_r)\pi_r}$$

This makes it possible to apply the EM algorithm, since, as shown before, in correspondence to the k-th expectation step the conditional probabilities values

$$\gamma_j^{(k-1)}(\mathbf{x}_i) \triangleq p(z_i = j|\mathbf{x}_i; \boldsymbol{\psi}^{(k-1)})$$

must be computed for $i = 1, \ldots, n$ and $j = 1, \ldots, K$. That is, the values

$$\gamma_j^{(k-1)}(\mathbf{x}_i) = \frac{q(\mathbf{x}_i; \boldsymbol{\theta}_j^{(k-1)})\pi_j^{(k-1)}}{\sum_{r=1}^{K} q(\mathbf{x}_i; \boldsymbol{\theta}_r^{(k-1)})\pi_r^{(k-1)}}$$

must be computed.

From the discussion on the expectation-maximization algorithm, this results into the following function to be maximized in the M-step:

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k-1)}, \mathrm{X}) &= \sum_{i=1}^{n} \sum_{j=1}^{K} \log p(\mathbf{x}_i, z_i; \boldsymbol{\psi}) p(z_i = j|\mathbf{x}_i; \boldsymbol{\psi}^{(k-1)}) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} \gamma_j^{(k-1)}(\mathbf{x}_i) \log \left(\pi_j q(\mathbf{x}_i; \boldsymbol{\theta}_j)\right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} \gamma_j^{(k-1)}(\mathbf{x}_i) \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{K} \gamma_j^{(k-1)}(\mathbf{x}_i) \log q(\mathbf{x}_i; \boldsymbol{\theta}_j)
\end{aligned}$$

First, let us take a look at the maximization wrt the component probabilities π_j.
As already shown, the maximization with respect to $\boldsymbol{\pi}$ provides

$$\pi_r^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_r^{(k-1)}(\mathbf{x}_i)$$

Let us now remaind that the maximization wrt component parameters $\boldsymbol{\theta}_r$ results into

$$\nabla_{\boldsymbol{\theta}_r} L(\boldsymbol{\Theta}, \lambda) = \sum_{i=1}^{n}\frac{\gamma_r^{(k-1)}(\mathbf{x}_i)}{q(\mathbf{x}_i;\boldsymbol{\theta}_r)}\nabla_{\boldsymbol{\theta}_r} q(\mathbf{x}_i;\boldsymbol{\theta}_r) = 0$$

Gaussian mixtures

In this case, we have $\boldsymbol{\theta}_r = \{\boldsymbol{\mu}_r, \Sigma_r\}$, the mean and covariance matrix of the r-th gaussian

$$q(\mathbf{x};\boldsymbol{\mu}_r,\Sigma_r) \triangleq \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_r,\Sigma_r) = \frac{1}{(2\pi)^{d/2}}\frac{1}{|\Sigma_r|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_r)^T\Sigma_r^{-1}(\mathbf{x}-\boldsymbol{\mu}_r)\right)$$

In the E-step, given the current values $\boldsymbol{\pi}^{(k)}$, $\boldsymbol{\Theta}^{(k)}$, the coefficients $\gamma_j^{(k-1)}(\mathbf{x}_i)$ are computed as already shown when gaussian mixtures were introduced, that is as

$$\gamma_j^{(k-1)}(\mathbf{x}_i) = \frac{\pi_j^{(k-1)}\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_j^{(k-1)},\Sigma_j^{(k-1)})}{\sum_{r=1}^{K}\pi_r^{(k-1)}\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_r^{(k-1)},\Sigma_r^{(k-1)})}$$

In the M-step, new values $\boldsymbol{\pi}^{(k)}$, $\boldsymbol{\Theta}^{(k)}$ are computed by maximization of the log-likelihood. As already shown this results into:

$$\pi_j^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_j^{(k-1)}(\mathbf{x}_i)$$

The maximization wrt $\boldsymbol{\mu}_j$ corresponds to solving

$$\sum_{i=1}^{n}\frac{\gamma_j(\mathbf{x}_i)}{\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_j,\Sigma_j)}\nabla_{\boldsymbol{\mu}_j}\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_j,\Sigma_j) = 0$$

which we already saw is

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{n}\gamma_j(\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n}\gamma_j(\mathbf{x}_i)}$$

As a consequence, we have

$$\boldsymbol{\mu}_j^{(k)} = \frac{\sum_{i=1}^{n}\gamma_j^{(k-1)}(\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{n}\gamma_j^{(k-1)}(\mathbf{x}_i)}$$

Similarly, the next value for Σ_j derives in general from the solution of

$$\sum_{i=1}^{n}\frac{\gamma_j(\mathbf{x}_i)}{\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_j,\Sigma_j)}\nabla_{\Sigma_j}\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_j,\Sigma_j) = 0$$

which can be proved to be

$$\Sigma_j = \frac{1}{\sum_{i=1}^n \gamma_j(\mathbf{x}_i)} \sum_{i=1}^n \gamma_j(\mathbf{x}_i)(\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T$$
$$= \frac{1}{\sum_{i=1}^n \gamma_j(\mathbf{x}_i)} \sum_{i=1}^n \gamma_j(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T - \boldsymbol{\mu}_j\boldsymbol{\mu}_j^T$$

As a consequence, we have then

$$\Sigma_j^{(k)} = \frac{1}{\sum_{i=1}^n \gamma_j^{(k-1)}(\mathbf{x}_i)} \sum_{i=1}^n \gamma_j^{(k-1)}(\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^T - \boldsymbol{\mu}_j^{(k)}\boldsymbol{\mu}_j^{(k)T}$$

Notice that, indeed,

1. knowing $\pi_j^{(k)}, \boldsymbol{\mu}_j^{(k)}, \Sigma_j^{(k)}$ for $j = 1, \ldots, K$ makes it possible, in the E-step, to compute $\gamma_j^{(k)}(\mathbf{x}_i)$ for $j = 1, \ldots, K$ and $i = 1, \ldots, n$
2. also, knowing $\gamma_j^{(k)}(\mathbf{x}_i)$ for $j = 1, \ldots, K$ and $i = 1, \ldots, n$ allows, in the M-step, to compute $\pi_j^{(k+1)}$, $\boldsymbol{\mu}_j^{(k+1)}$, $\Sigma_j^{(k+1)}$

Mixtures of Poissons

In the case of a mixture of K Poisson distributions both $\mathcal{Z}$ and $\mathcal{X}$ are discrete, thus implying that $p(\mathbf{z})$ and $p(\mathbf{x}|\mathbf{z})$ are both discrete distributions (in this case categorical and Poisson distributions). In terms of marginal distribution, we have a mixture, again:

$$p(x; \boldsymbol{\pi}, \boldsymbol{\Lambda}) = \sum_{i=1}^K \pi_i q(x; \lambda_i)$$

with

$$q(x; \lambda_k) = \frac{e^{-\lambda_k}\lambda_k^x}{x!},$$

In the EM algorithm, the expectation step requires computing

$$\gamma_j^{(k-1)}(x_i) = \frac{\pi_j^{(k)} \frac{e^{-\lambda_j^{(k)}}\lambda_j^{(k)x_i}}{x_i!}}{\sum_{r=1}^K \pi_r^{(k)} \frac{e^{-\lambda_r^{(k)}}\lambda_r^{(k)x_i}}{x_i!}}.$$

For what regards the maximization step, the new values $\boldsymbol{\pi}^{(k)}$ are still given by

$$\pi_j^{(k)} = \frac{1}{n}\sum_{i=1}^n \gamma_j^{(k-1)}(\mathbf{x}_i)$$

while the new values $\lambda_j^{(k)}$ derive by setting

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \gamma_j^{(k-1)}(x_i) \frac{\partial}{\partial \lambda_j} \log q(x_i; \lambda_j) \\
&= \sum_{i=1}^{n} \gamma_j^{(k-1)}(x_i) \frac{\partial}{\partial \lambda_j} (-\lambda_j + x_i \log \lambda_j - \log x_i!) \\
&= \sum_{i=1}^{n} \gamma_j^{(k-1)}(x_i) \left(-1 + \frac{x_i}{\lambda_j} \right) \\
&= -\sum_{i=1}^{n} \gamma_j^{(k-1)}(x_i) + \frac{1}{\lambda_j} \sum_{i=1}^{n} \gamma_j^{(k-1)}(x_i) x_i
\end{aligned}$$

which results into

$$\lambda_j^{(k)} = \frac{\sum_{i=1}^{n} \gamma_j^{(k-1)}(x_i) x_i}{\sum_{i=1}^{n} \gamma_j^{(k-1)}(x_i)}$$