

## Gaussian Mixture Models Expectation-Maximization derivation

(Based on Notes by Stefanos Zafeiriou)

Note that in the following derivation, the dependence on the parameters ( $\theta$ ) is omitted for simplicity. Note that  $z_n$  is the hidden state that generated  $x_n$ .  $z_{nk}$  is one if the  $k^{\text{th}}$  hidden state generated  $x_n$ , otherwise it is zero. When summing over  $z_n$ , we sum over all possible values it can take.

We can write the joint probability as

$$p(x_n, z_n) = p(x_n | z_n) * p(z_n) = \prod_k N(x_n | \mu_k, \Sigma_k)^{z_{nk}} * \pi_k^{z_{nk}}$$

Taking the natural logarithm

$$\ln(p(x_n, z_n)) = \sum_k z_{nk} * \ln(N(x_n | \mu_k, \Sigma_k)^{z_{nk}}) + \sum_k z_{nk} * \ln(\pi_k)$$

Taking into account all samples:

$$\ln(p(x, z)) = \sum_n \sum_k z_{nk} * \ln(N(x_n | \mu_k, \Sigma_k)^{z_{nk}}) + \sum_n \sum_k z_{nk} * \ln(\pi_k)$$

If we wanted to optimize this, we couldn't, since  $z_{nk}$  values are unknown. We can replace them with their expectations over the posterior distribution,  $p(z_n | x_n)$  though.

$$p(z_n | x_n) = \frac{p(x_n, z_n)}{p(x_n)} = \frac{\prod_k N(x_n | \mu_k, \Sigma_k)^{z_{nk}} * \pi_k^{z_{nk}}}{\sum_{z_n} \prod_k N(x_n | \mu_k, \Sigma_k)^{z_{nk}} * \pi_k^{z_{nk}}}$$

Expectation of  $z_{nk}$  using the above distribution

$$\begin{aligned} E(z_{nk}) &= \frac{\sum_{z_n} z_{nk} * \prod_k N(x_n | \mu_k, \Sigma_k)^{z_{nk}} * \pi_k^{z_{nk}}}{\sum_{z_n} \prod_k N(x_n | \mu_k, \Sigma_k)^{z_{nk}} * \pi_k^{z_{nk}}} = \frac{N(x_n | \mu_k, \Sigma_k)^{z_{nk}} * \pi_k^{z_{nk}}}{\sum_{z_n} \prod_k N(x_n | \mu_k, \Sigma_k)^{z_{nk}} * \pi_k^{z_{nk}}} \\ &= \frac{N(x_n | \mu_k, \Sigma_k) * \pi_k}{\sum_l N(x_n | \mu_l, \Sigma_l) * \pi_l} \end{aligned}$$

Using this, our function to optimize is:

$$G(\theta) = \sum_n \sum_k E(z_{nk}) * \ln(N(x_n | \mu_k, \Sigma_k)^{z_{nk}}) + \sum_n \sum_k E(z_{nk}) * \ln(\pi_k)$$

Expansion of  $\ln(N(x_n | \mu_k, \Sigma_k)^{z_{nk}})$  yields:

$$\begin{aligned} G(\theta) &= \sum_n \sum_k E(z_{nk}) * \left\{ \frac{-1}{2} * (x - \mu_k)' \Sigma^{-1} (x - \mu_k) - \frac{1}{2} * (F \ln(2\pi) + \ln(|\Sigma_k|)) \right\} \\ &\quad + \sum_n \sum_k E(z_{nk}) * \ln(\pi_k) \end{aligned}$$

Update rule for  $\mu_k$

$$\frac{\delta G}{\delta \mu_k} = \frac{-\delta G}{\delta (x - \mu_k)} = \frac{1}{2} * \Sigma_k^{-1} * \sum_n E(z_{nk}) * (x - \mu_k) = 0$$

Multiplying by  $2 * \Sigma_k$  from the left and rearranging gives:

$$\mu_k = \frac{\sum_n E(z_{nk}) * x}{\sum_n E(z_{nk})}$$

Update rule for  $\Sigma_k$

$$\frac{\delta G}{\delta \Sigma_k} = \frac{-1}{2} * \frac{\delta \sum_n \sum_k E(z_{nk}) * (x - \mu_k)' \Sigma_k^{-1} (x - \mu_k)}{\delta \Sigma_k} - \frac{1}{2} * \frac{\delta \sum_n \sum_k E(z_{nk}) * \ln(|\Sigma_k|)}{\delta \Sigma_k}$$

Using the formulas  $\frac{\delta}{\delta X} (a^T X^{-1} b) = -X^{-T} a b^T X^{-T}$  and  $\frac{\delta}{\delta X} (\log(|X|)) = X^{-T}$  and using that  $\Sigma_k$  is symmetric we get

$$\frac{\delta G}{\delta \Sigma_k} = \frac{1}{2} * \sum_n E(z_{nk}) * \{ \Sigma_k^{-1} (x - \mu_k) (x - \mu_k)' \Sigma_k^{-1} - \Sigma_k^{-1} \} = 0$$

Multiplying by 2,  $\Sigma$  on the left and then  $\Sigma$  on the right gives:

$$\sum_n E(z_{nk}) * \{ (x - \mu_k) (x - \mu_k)' - \Sigma_k \} = 0$$

Rearranging gives:

$$\Sigma_k = \frac{\sum_n E(z_{nk}) * (x - \mu_k) (x - \mu_k)'}{\sum_n E(z_{nk})}$$

Update rule for  $\pi_k$ :

Since we have a constraint:  $\sum_k \pi_k = 1$  we will use the method of Lagrange multipliers and minimize the Lagrangian function:

$$L(\theta) = G(\theta) + \lambda * (1 - \sum_k \pi_k)$$

Its partial derivative w.r.t  $\pi_k$  yields:

$$\frac{\delta L(\theta)}{\delta \pi_k} = \sum_n E(z_{nk}) * \frac{1}{\pi_k} - \lambda = 0$$

After rearranging we get:

$$\sum_n E(z_{nk}) = \lambda * \pi_k$$

If we also sum over k, then:

$$\sum_k \sum_n E(z_{nk}) = \sum_k \lambda * \pi_k = \lambda$$

The left hand side is equivalent to:  $\sum_n \sum_k E(z_{nk}) = \sum_n 1 = N$  (the number of samples)

Hence  $\lambda = N$  and we get that

$$\pi_k = \frac{\sum_n E(z_{nk})}{N}$$