Derivation of Using Scalar Sigma in EM to update GMM

(a) **Derivation for Batch EM**
Assume the GMM composed of K Gaussian components, the pdf of the GMM($\theta = (\pi_k, \mu_k, \Sigma_k)$) is:

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\mu_k, \Sigma_k) \tag{1}$$

So the Likelihood Function of the GMM(N = the size of the dataset) should be:

$$\prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \{\sum_{k=1}^{K} \pi_k p(x_i|\mu_k, \Sigma_k)\} \tag{2}$$

Because the possibility of each point is usually a very small number, and the production of many small numbers will cause the underflow issue with floating point numbers, we will **take its log** to transform the production function to a summation function and get the **log-likelihood function** as following:

$$\sum_{i=1}^{N} \log\{\sum_{k=1}^{K} \pi_k p(x_i|\mu_k, \Sigma_k)\} \tag{3}$$

Because the equation(1) p(x) will be expanded as following:

$$p(x) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{(2\pi)^d |\Sigma_k|}} e^{-\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)} \tag{4}$$

And the Σ_k is changed to be a scalar parameter:

$$\begin{aligned} \Sigma_k &= \sigma_k^2 \\ &\Rightarrow |\Sigma_k| = (\sigma^2)^d \\ &\Rightarrow \Sigma_k^{-1} = \frac{1}{\sigma_k^2} \end{aligned}$$

Then the function becomes:

$$p(x) = \sum_{k=1}^{m} \frac{\pi_k}{\sqrt{(2\pi\sigma_k^2)^d}} e^{-\frac{1}{2}\frac{(x-\mu_k)^\top (x-\mu_k)}{\sigma_k^2}} \tag{5}$$

So the parameters in Gauss θ becomes: $\theta = (\pi_k, \mu_k, \sigma_k^2)$ and we can use $h_k = \sigma_k^2$ makes $\theta = (\pi_k, \mu_k, h_k)$

Because I need to maximize the equation(3), however, the $\log \Sigma$ is a challenge when doing the maximization. The normal way to do that is to use **Jensen Inequality** as following:
So the log-likelihood function becomes (z is the latent variable):

$\sum_{i=1}^{N} \log \sum_{k=1}^{K} p(x_i) = \sum_{i=1}^{N} \log\{\frac{\sum_{k=1}^{K} p(x_i,z_i=k|\theta)}{p(z_i=k|\theta_t)} p(z_i = k|\theta_t)$

$$\geq \sum_{i=1}^{N}\sum_{k=1}^{K} p(z_i = k|\theta_t) \log\{\frac{p(x_i, z_i = k|\theta)}{p(z_i = k|\theta_t)}\} \tag{6}$$

Represent $\gamma_{ik} = p(z_i = k|\theta_t)$ (γ_{ik} is the possibility that data x_i belonged to the kth Gauss; θ_t is the updated parameter after the current E-step)
Now I just need to maximize the equation (6) expanded as the following $L(\theta)$

$$L(\theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{jk} \left[\log \pi_k - \frac{d}{2} \log 2\pi h_k - \frac{1}{2}\frac{(x_i - \mu_k)^\top (x_i - \mu_k)}{h_k}\right] \tag{7}$$

And because we have a **constraint** that $\sum_{k=1}^{K} \pi_k = 1$, I use **Lagrange multiplier** to add this constraint in the maximization process as following $T(\theta)$.

$$T(\theta) = L(\theta) + \lambda(\sum_{k=1}^{K} \pi_k - 1) \tag{8}$$

Maximization: Get the derivation and set it to equal to zero.Then the parameters θ become:
$\frac{\partial T}{\partial \pi_k} = \sum_{i=1}^{N} \frac{\gamma_{ik}}{\pi_k} + \lambda = 0$

$$\Rightarrow \pi_k = \frac{\sum_{i=1}^{N} \gamma_{ik}}{\sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}} \tag{9}$$

$\frac{\partial T}{\partial \mu_k} = \sum_{i=1}^{N} \gamma_{ik} \frac{x_i - \mu_k}{h_k} = 0$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}} \tag{10}$$

$\frac{\partial \pi}{\partial h_k} = -\sum_{i=1}^{N} \gamma_{ik} \frac{d}{2}\frac{1}{h_k} + \frac{1}{2}\sum_{i=1}^{N} \gamma_{ik} \frac{(x_i-\mu_k)^\top (x_i-\mu_k)}{h_k^2} = 0$

$$\Rightarrow h_k = \frac{\sum_{i=1}^{N} \gamma_{ik} (x_i - \mu_k)^\top (x_i - \mu_k)}{d \sum_{i=1}^{N} \gamma_{ik}} \tag{11}$$

Finally, deviation result with $\Sigma_k = h_k I$ is as following:

The two Steps in EM become:

E-Step:
$\gamma_i k = p(z^j = k|\theta^t)$

M-Step:
$\pi_k = \frac{\sum_{i=1}^{N} \gamma_{ik}}{\sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}}$

$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}}$

$h_k = \frac{\sum_{i=1}^{N} \gamma_{ik} (x_i-\mu_k)^\top (x_i-\mu_k)}{d \sum_{i=1}^{N} \gamma_{ik}}$

In this case, dimension d = 2

(b) **Derivation for Online EM**[1]

Given the Gauvain's update formula for the matrix Σ_j of j-th Gaussian in the mixture reads:

$$\Sigma_j'' = \frac{b_j I + \Sigma_{q=0}^{N-1} \gamma_{qj}(s_q - \mu_j)(s_q - \mu_j)^\top}{(a_j - 2) + \Sigma_{q=0}^{N-1} \gamma_{qj}} \tag{12}$$

Because in the equation(12), scalars a,b and v are parameters of conjugate priors induced by the MAP solution[2]. I is the identity matrix. a $>$ d-1, b $>$ 0 are hyper-parameters, and d $=$ 2 is the dimension. Bishop[2006] provides details on the use of Dirichlet and Wishart distributions as conjugate priors.

So we use the same conjugate as the on-line paper[2] and by using the derivation result of h_k as above, we can get the$\Sigma_j$as following:

$$\Sigma_j' = \frac{b_j I + \Sigma_{q=0}^{N-1} \gamma_{qj}(s_q - \mu_j)^\top (s_q - \mu_j)}{(a_j - 2) + d\Sigma_{q=0}^{N-1} \gamma_{qj}} \tag{13}$$

By using simple algebra, we get:

$$(s_q - \mu_j)^\top (s_q - \mu_j) = s_q^\top s_q - s_q^\top \mu_j - \mu_j^\top s_q + \mu_j \mu_j^\top \tag{14}$$

Substituting (14) into (13) and multiplying both the nominator and denominator by $\frac{1}{N}$, the formula for Σ_j becomes:

$$\Sigma_j' = \Sigma_j' \frac{\frac{1}{N}}{\frac{1}{N}} \tag{15}$$

$$= \frac{\frac{b_j}{N} + \frac{1}{N}\Sigma_{q=0}^{N-1} \gamma_{qj} s_q^\top s_q - A + B}{\frac{(a_j-2)}{N} + \Sigma_{q=0}^{N-1} \frac{\gamma_{qj}}{N}} \tag{16}$$

$$\text{where,} \qquad A = \frac{1}{N}\Sigma_{q=0}^{N-1} \gamma_{qj} s_q^\top \mu_j + \frac{1}{N}\Sigma_{q=0}^{N-1} \gamma_{qj} \mu_j^\top s_q \tag{17}$$

$$B = \frac{1}{N}\Sigma_{q=0}^{N-1} \gamma_{qj} \mu_j^\top \mu_j \tag{18}$$

Then the fact to obtain MAP update formula[1] of the covariance matrix(the sufficient statistics: a triplet $u_i^j = ((u_\gamma)_i^j, (s)_i^j, (ss^\top)_i^j)$)is:

$$\Sigma_j' = \frac{\frac{b_j I}{N} + (s^\top s)_i^j - A + (u_\gamma)_i^j B}{\frac{(a_j-2)}{N} + d(u_\gamma)_i^j} \tag{19}$$

$$\text{where,} \qquad A = (s^\top)_i^j \mu_j + \mu_j^\top s_i^j \tag{20}$$

$$B = \mu_j^\top \mu_j \tag{21}$$

$$\tag{22}$$

This is a article ralted to Bibtex [1]
The second citition [**?**]

Refrence

[1] Vorba, Jiri, et al. "On-line Learning of Parametric Mixture Models for Light Transport Simulation?Supplemental material."

[2] Vorba, Ji?í, et al. "On-line learning of parametric mixture models for light transport simulation." ACM Transactions on Graphics (TOG) 33.4 (2014): 101.

[3] Bishop, Christopher M. "Pattern Recognition." Machine Learning (2006).

References

[1] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.