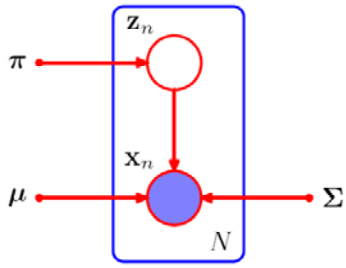
GMM with EM

Zhi-Sheng Chen 2012/8/20

The Gaussian mixture model could be formulated as $p(x)=\sum_{k=1}^{K}\pi_k \mathrm{N}(x \mid \mu_k, \Sigma_k)$, where the parameters are $\theta=\{\pi_k, \mu_k, \Sigma_k \mid k=1 \sim K\}$. The graphic model is shown below:

Figure 9.6 Graphical representation of a Gaussian mixture model for a set of N i.i.d. data points $\{\mathbf{x}_n\}$, with corresponding latent points $\{\mathbf{z}_n\}$, where $n=1,\ldots,N$.



A complete i^{th} observation is (x_i, z_i), where $z_i=\{1,2,\ldots,K\}$, and $z_i=k$ indicates the explicit observation x_i comes from k^{th} Gaussian component. So the complete data log-likelihood $p(x_i, z_i=k \mid \theta)=p(z_i=k \mid \theta)p(x_i \mid z_i=k,\theta)=\pi_k \mathrm{N}(x_i \mid \mu_k, \Sigma_k)$.

Following the E-step of standard EM procedure, we first set

$$q(z_i=k)=p(z_i \mid x_i,\theta^{old})=\frac{p(z_i,x_i \mid \theta^{old})}{p(x_i \mid \theta^{old})}=\frac{p(z_i=k\theta^{old})p(x_i \mid z_i=k,\theta^{old})}{\sum_{l=1}^{K}p(z_l=k\theta^{old})p(x_i \mid z_l=k,\theta^{old})}$$

$$=\frac{\pi_k^{old}\mathrm{N}(x_i \mid \mu_k^{old},\Sigma_k^{old})}{\sum_{l=1}^{K}\pi_l^{old}\mathrm{N}(x_i \mid \mu_l^{old},\Sigma_l^{old})}$$

Then for the M-step, we are to maximize

$$\sum_{i=1}^{N}\sum_{k=1}^{K}q(z_i=k)\ln p(x_i,z_i=k \mid \theta) \text{ subject to } 1=\sum_{k=1}^{K}\pi_k .$$

$$L=\sum_{i=1}^{N}\sum_{k=1}^{K}q(z_i=k)\ln \pi_k \mathrm{N}(x_i \mid \mu_k,\Sigma_k)+\alpha\left(1-\sum_{k=1}^{K}\pi_k\right)$$

1) $\partial L/\partial \pi_k=0 \Rightarrow$

$\left(\sum_{i=1}^{N}q(z_i=k)\right)/\pi_k=\alpha$, and according to $1=\sum_{k=1}^{K}\pi_k$, we have

$$\pi_k=\frac{\sum_{i=1}^{N}q(z_i=k)}{\sum_{k=1}^{K}\sum_{i=1}^{N}q(z_i=k)}.$$

2) $\partial L / \partial \mu_k = 0 \Rightarrow$

$$\frac{\partial}{\partial \mu_k}\left(\sum_{i=1}^{N} \sum_{k=1}^{K} q(z_i = k)\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma^{-1} (x_i - \mu_k) \right) \right) = 0$$
$$\Rightarrow \sum_{i=1}^{N} q(z_i = k) \Sigma^{-1} (x_i - \mu_k) = 0$$
$$\Rightarrow \sum_{i=1}^{N} q(z_i = k) x_i = \left[\sum_{i=1}^{N} q(z_i = k) \right] \mu_k$$
$$\Rightarrow \mu_k = \sum_{i=1}^{N} q(z_i = k) x_i \Big/ \sum_{i=1}^{N} q(z_i = k)$$

3) $\partial L / \partial \Sigma_k = 0$

We need three rules:

A is not a function of X	[7] $\frac{\partial \operatorname{tr}(\mathbf{X}^{-1}\mathbf{A})}{\partial \mathbf{X}} =$	$-(\mathbf{X}^{-1})^{\mathrm{T}} \mathbf{A} (\mathbf{X}^{-1})^{\mathrm{T}}$	$-\mathbf{X}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{X}^{-1}$

(from wikipedia.org “Matrix calculus”)

, $\alpha^t A \alpha = tr(A \alpha \alpha^T)$, and $\frac{\partial}{\partial A} \ln|A| = (A^{-1})^T$

Then we can calculate the derivatives as follows,

$$\partial L / \partial \Sigma_k = 0 \Rightarrow \frac{\partial}{\partial \Sigma_k}\left[\sum_{i=1}^{N} \sum_{k=1}^{K} q(z_i = k)\left(-\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right) \right] = 0$$
$$\Rightarrow \frac{\partial}{\partial \Sigma_k}\left[\sum_{i=1}^{N} \sum_{k=1}^{K} q(z_i = k)\left(-\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2} tr\left(\Sigma_k^{-1} (x_i - \mu_k)(x_i - \mu_k)^T \right) \right) \right] = 0$$
$$\Rightarrow \sum_{i=1}^{N} q(z_i = k)\left(-\frac{1}{2}\Sigma_k^{-1} + \frac{1}{2}\Sigma_k^{-1}(x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1} \right) = 0$$
$$\Rightarrow \left(\sum_{i=1}^{N} q(z_i = k) \right) \Sigma_k^{-1} = \Sigma_k^{-1} \left(\sum_{i=1}^{N} q(z_i = k)(x_i - \mu_k)(x_i - \mu_k)^T \right) \Sigma_k^{-1}$$
$$\Rightarrow \Sigma_k = \frac{\sum_{i=1}^{N} q(z_i = k)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} q(z_i = k)}$$

Demos could be seen from
http://neural.cs.nthu.edu.tw/jang/matlab/toolbox/machineLearning/