Qn 3) If we have a prior distribution on parameter, then, we are trying to solve the problem

$$\theta^*_{max} = \arg\max_{\theta} \left\{ \log\left[P(y|\theta)\right] + \log\left[P(\theta)\right] \right\}$$

say, $F(q,\theta) = \sum_x q(x) \log\left(P(y,x|\theta)/q(x)\right)$.

$$KL\left(q \| p(x|y,\theta)\right) := \sum_x q(x) \log\left(\frac{q(x)}{P(x|y,\theta)}\right).$$

then, we have

$$F(q,\theta) = \log\left[P(y|\theta)P(\theta)\right] - KL\left(q \| p(x|y,\theta)\right) - \log\left[P(\theta)\right].$$

$\left\{ \begin{matrix} \therefore F(q,\theta) \\ + \log\left[P(\theta)\right] \end{matrix} \right\}$ is a lower bound of $\log\left[P(y|\theta)P(\theta)\right]$

with equality only when $q(x)$ is the true posterior.

(a) E-step :

Need to design 'q' to maximize

$$\{F(q, \theta_i) + P(\theta)\}.$$

where,

$P(\theta)$ = independent of data.

$\therefore$ Equivalent to maximizing

$$F(q, \theta_i).$$

$\therefore$ q is still = posterior of labels given y & θ.

(Same as regular EM algorithm)

(b) M-step: Find 'θ' to maximize

$$\{F(q, \theta_i) + P(\theta)\}$$

We know, $F(q, \theta_i) = E_q[\log(P(x, y)|\theta)] + H(q)$

independent of θ.

$\therefore$ We need to find

$$\boxed{\theta^* = \underset{\theta}{\operatorname{argmax}} \{Q(\theta; \theta_i) + P(\theta)\}}$$

Reference: Conjugate Prior Penalized learning

c) (i) We use conjugate priors as prior for parameters of GMM.

Conjugate prior of a single gaussian is a product of gaussian density N and wishart density W.

Also, conjugate prior of mixing probabilities is a Dirichlet density D.

∴ We obtain,

$$p(\Theta) = D(p(k) \mid r_k) \prod_{k=1}^{K} N(\mu_k \mid v_k, \eta_k^{-1} \Sigma_k)$$
$$\times W(\Sigma_k^{-1} \mid \alpha_k, \beta_k)$$

where, D = Dirichlet density
N = Gaussian density
W = Wishart density.

ii) If we use the above prior, the M-step updates obtained are as follows,

$$w_k = \hat{p}(k) = \frac{\sum_{n=1}^{N} p(k \mid x^n) + r_k - 1}{\sum_{k=1}^{K} \left(\sum_{n=1}^{N} p(k \mid x^n) + r_k - 1 \right)}$$

$$\hat{\mu}_k = \frac{\sum_{n=1}^{N} p(k|x^n)\, x^n + \eta_k \nu_k}{\sum_{n=1}^{N} p(k|x^n) + \eta_k}.$$

And,

$$\hat{\Sigma}_k = \frac{\sum_{n=1}^{N} \left\{ p(k|x^n)(x^n - \hat{\mu}_k)(x^n - \hat{\mu}_k)^T + \eta_k (\hat{\mu}_k - \nu_k)(\hat{\mu}_k - \nu_k)^T + 2\beta_k \right\}}{\sum_{n=1}^{N} p(k|x^n) + 2\alpha_k - d}.$$

(C_k)