Q1. For the mixture model the joint distribution can be written

$$p(x_a, x_b) = \sum_{k=1}^{K} \pi_k p(x_a, x_b | k).$$

We can find the conditional density $p(x_b | x_a)$ by making use of relation

$$p(x_b | x_a) = \frac{p(x_a, x_b)}{p(x_a)}$$

For mixture model the marginal density of x_a is given by

$$p(x_a) = \sum_{k=1}^{K} \pi_k p(x_a | k) \quad \text{where} \quad p(x_a | k) = \int p(x_a, x_b | k) dx_b$$

Thus we can write the conditional density in the form

$$p(x_b | x_a) = \frac{\sum_{k=1}^{K} \pi_k p(x_a, x_b | k)}{\sum_{j=1}^{K} \pi_j p(x_a | j)}$$

Now one can decompose the number using

$$p(x_a, x_b | k) = p(x_b | x_a, k) p(x_a | k).$$

which allows us finally to write the conditional density as a mixture model of the form

$$p(x_b | x_a) = \sum_{k=1}^{K} \lambda_k p(x_b | x_a, k).$$

where the mixture coefficientes are given by

$$\lambda_k = p(k | x_a) = \frac{\pi_k p(x_a | k)}{\sum_j \pi_j p(x_a | j)}.$$

Q2. a) $\hat{a} = \frac{\frac{1}{2}}{\frac{1}{2} + \mu} h \quad \hat{b} = \frac{\mu}{\frac{1}{2} + \mu} h.$ b) $\hat{\mu} = \frac{h - a + c}{6(h - a + c + d)}$

Q3. a)

$$P(Z=1 | X=1) \propto P(X=1 | Z=1) P(Z=1) = \pi_1 e^{-1}$$

$$P(Z=2 | X=1) \propto P(X=1 | Z=2) P(Z=2) = \pi_2 4e^{-2}$$

$$P(Z=3 | X=1) \propto P(X=1 | Z=3) P(Z=3) = \pi_3 16e^{-4}$$

$$P(Z=1 | X=1) = \frac{\pi_1 e^{-1}}{(\pi_1 e^{-1} + \pi_2 4e^{-2} + \pi_3 16e^{-4})}$$

b) For each $X = x$.

$$P(Z=k | X=x) = \frac{P(X=x | Z=k) P(Z=k)}{\sum_{k'} P(X=x | Z=k') P(Z=k')} = \frac{\beta_k^2 x e^{-\beta_k x} \pi_k}{\sum_{k'} \beta_{k'}^2 x e^{-\beta_{k'} x} \pi_{k'}}$$

Q4 Consider first the maximization with respects to the componentes π_k of π, To do thus we must take account of the summation constraint

$$\sum_{k=1}^{K} \pi_k = 1.$$

We therefore first omit terms from $Q(\theta, \theta^{old})$ which are independent of π, and then add a Lagrange multipleer term to enforce the constraint, giving the following function to be maximized.

$$\tilde{Q} = \sum_{k=1}^{K} \gamma(z_{ik}) \ln \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right).$$

Setting the derivertive with ~~re~~respect to π_k equal to zero we obtain. $0 = \gamma(z_{ik}) \frac{1}{\pi_k} + \lambda.$

We now multiply through by π_k and then sum over k and make use of the summation constraint to give $\lambda = - \sum_{k=1}^{K} \gamma(z_{ik}).$

Substituting back into (1) and solving for λ we obtain that.

For the maximization with respect to A we follow the same step and first omit terms from $Q(\theta, \theta_{old})$ which are independent of A, and then add appropriate Lagrange multiplier terms to enforce the summation constraints. In this case there are K constraints to be satisfied since we must have $\sum_{k=1}^{K} A_{jk} = 1$. for $j = 1 \dots K$, we introduce K Lagrange multiplier λ_j for $j = 1, \dots, K$, and maximize the following function

$$\hat{Q} = \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} + \sum_{j=1}^{K} \lambda_j \left(\sum_{k=1}^{K} A_{jk} - 1 \right).$$

Setting the derivative of $\hat{Q}$ with respect to A_{jk} to zero we obtain

$$0 = \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk}) \frac{1}{A_{jk}} + \lambda_j$$

Again we multiply through by A_{jk} and then sum over k and make use of the summation constraint give $\lambda_j = -\sum_{n=2}^{N} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk})$.

Substituting for λ_j in (2).

Q5. Only the final term of $Q(\theta, \theta_{old})$ given by (13.17) depends on the parameters of the emission model. For the multinomial variable x, whose D components are all zero except for a single entry of 1

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(x_n | \phi_k) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \sum_{i=1}^{D} x_{ni} \ln \mu_{ki}$$

Now when we maximize with respect to μ_{ki} we have to take account of the constraint that, for each value of k the components of μ_k must sum to one. We therefore introduce Lagrange multiples $\{\lambda_k\}$ and maximize the modified function given by.

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \sum_{i=1}^{D} x_{ni} \ln \mu_{ki} + \sum_{k=1}^{K} \lambda_k \left(\sum_{i=1}^{D} \mu_{ki} - 1 \right).$$

Setting the derivative with respect to μ_{ki} to zero. we obtain

$0 = \sum_{n=1}^{N} \gamma(z_{nk}) \frac{x_{ni}}{\mu_{ki}} + \lambda_k$. Multiplying through by μ_{ki}, summing over i, and making use of the constraint on μ_{ki} together with the result $\sum_i x_{ni} = 1$ we have $\lambda_k = -\sum_{n=1}^{N} \gamma(z_{nk})$.

Finally, back-substituting for λ_k and solving for μ_{ki} we again obtain (13.23).

Similarly, for the case of a multivariate Bernoulli observed variable x whose D components independently take the value of 0 or 1, using the standard expression for the multivariate Bernoulli distribution we have $\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(x_n | \phi_k) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \sum_{i=1}^{D} \{ x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki}) \}$.

Maximizing with respect to μ_{ki} we obtain $\mu_{ki} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})}$ which is equivalent to (13.23).

Q6. First of all, note that for every observed variable there is a ~~compon~~ corresponding latent variable, and so for every sequence $X^{(r)}$ of observed variables there is a corresponding sequence $Z^{(r)}$ of latent variables. These sequences are assumed to be independent given the model parameters, and so the joint distribution of all latent and observed variable will be given by $p(X, Z | \theta) = \prod_{r=1}^{R} p(X^{(r)}, Z^{(r)} | \theta)$ where X denotes $\{X^{(r)}\}$ and Z denotes $\{Z^{(r)}\}$. Using the sum and product rules of probability we then see that posterior distribution for the latent sequences that factorizes with respect to those sequences, so that.

$$p(Z | X, \theta) = \frac{p(X, Z | \theta)}{\sum_Z p(X, Z | \theta)} = \frac{\prod_{r=1}^{R} p(X^{(r)}, Z^{(r)} | \theta)}{\sum_{Z^{(1)}} \cdots \sum_{Z^{(R)}} \prod_{r=1}^{R} p(X^{(r)}, Z^{(r)} | \theta)} = \prod_{r=1}^{R} \left\{ \frac{p(X^{(r)}, Z^{(r)} | \theta)}{\sum_{Z^{(r)}} p(X^{(r)}, Z^{(r)} | \theta)} \right\} = \prod_{r=1}^{R} p(Z^{(r)}, X^{(r)} | \theta)$$

Thus the ~~equiv~~ evaluation of the posterior distribution of the latent variables, corresponding to the E-step of the EM algorithm, can be done independently for each of the sequences (using the standard alpha-beta recursions).

Now consider the M-step. We use the posterior distribution computed in the E-step using θ_{old} to evaluate the expectation of the complete-data ~~log~~ log likelihood. From our expression for the joint distribution we see that this is given by

$$Q(\theta, \theta_{old}) = E_Z[\ln p(X, Z \mid \theta)] = E_Z\left[\sum_{r=1}^{R} \ln p(X^{(r)}, Z^{(r)} \mid \theta)\right] = \sum_{r=1}^{R} p(Z^{(r)}, X^{(r)} \mid \theta_{old}) \ln p(X, Z \mid \theta)$$

$$= \sum_{r=1}^{R}\sum_{k=1}^{K} \gamma(z_{1k}^{(r)}) \ln \pi_k + \sum_{r=1}^{R}\sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} \xi(z_{n-1}^{(r)}) \ln A_{jk} + \sum_{r=1}^{R}\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk}^{(r)}) \ln p(x_n^{(r)} \mid \phi_k)$$

We now maximize ~~this~~ the quantity with respect to π and A in the usual way, with Lagrange multipliers to takes account of the summation constraints, ~~ege~~ yielding (13.124) and (13.125). The ~~mo~~ M-step results for the mean of the Gaussian follow in the usual way also.