KTH Electrical Engineering

Solution Examples to Exam in Pattern Recognition EN2202

Date: Wednesday Oct 30, 2013, 14:00 – 19:00

Place: E31, E33.

Allowed: Beta, calculator with empty memory, one page handwritten note.

Grades: A: 31p; B: 27p; C: 23p; D: 20p; E: 17; of max 23p + 10p project bonus.

Language: English.

Results: Friday, Nov 15.

Review: At KTH-S3/STEX, Osquldas v. 10.

Contact: Saikat, 073 891 3581; Jalil, 073 756 1933

Good Luck!

1 In a given pattern-classification application the signal source can be in one of two states, here called $S = 1$ and $S = 2$. The two source states are known to occur with equal probabilities. You can observe a feature vector $\boldsymbol{X} = (X_1, X_2)^T$ with two elements. Depending on the source state $S = i$, the feature vector has a Gaussian conditional distribution, defined by the mean vector μ_i and covariance matrix C_i, with known values

$$\mu_1 = \begin{pmatrix} 3 \\ 3 \end{pmatrix} \qquad C_1 = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$
$$\mu_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \qquad C_2 = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$

(a) Design an optimal classifier that can guess the source state with minimum error probability, and simplify the classifier to show that it is *possible* to make optimal decisions using a *linear* discriminant function of the type $g(x_1, x_2) = ax_1 + bx_2 + c$, together with a threshold mechanism. (4p)

Solution: As both source alternatives are equally probable, we use the *Maximum Likelihood* decision rule. The covariance matrices are equal, $C_1 = C_2 = C$, and we can define a single discriminant function simply as

$$\begin{aligned} g(\boldsymbol{x}) = \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|1) - \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|2) = \\ = (\boldsymbol{x} - \mu_2)^T C^{-1}(\boldsymbol{x} - \mu_2)/2 - (\boldsymbol{x} - \mu_1)^T C^{-1}(\boldsymbol{x} - \mu_1)/2 = \\ = \mu_1^T C^{-1}\boldsymbol{x} - \mu_2^T C^{-1}\boldsymbol{x} - \mu_1^T C^{-1}\mu_1/2 + \mu_2^T C^{-1}\mu_2/2 = \\ = (\mu_1 - \mu_2)^T C^{-1}\boldsymbol{x} - (\mu_1 - \mu_2)C^{-1}(\mu_1 + \mu_2)/2 = \\ = (\mu_1 - \mu_2)^T C^{-1}(\boldsymbol{x} - (\mu_1 + \mu_2)/2) \end{aligned}$$

Then, the optimal classifier decides $S = 1$, whenever $g(\boldsymbol{x}) > 0$ and vice versa. With the given covariance we have

$$C^{-1} = \frac{1}{15}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$$

and

$$g(\boldsymbol{x}) = \frac{1}{15}\begin{pmatrix} 4 & 4 \end{pmatrix}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}\left(\boldsymbol{x} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \frac{1}{3}\begin{pmatrix} 1 & 1 \end{pmatrix}\left(\boldsymbol{x} - \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \frac{1}{3}(x_1 + x_2 - 2)$$

Thus, the classifier will decide $S = 1$, whenever $x_1 + x_2 > 2$, and vice versa.

(b) What is the the conditional probability that source $S = 1$ was active, given an observed feature vector $(0, 0)$, i.e. $P(S = 1|\boldsymbol{X} = (0, 0)^T)$? (1p)

Solution: Omitting factors that are equal for both feature distributions, we find log-likelihood values

$$L_i = \ln f_{\boldsymbol{X}|S}(\boldsymbol{x}|i) = -(\boldsymbol{x} - \mu_i)^T C^{-1}(\boldsymbol{x} - \mu_i)/2 + \text{const.}$$

with

$$L_1 = -\frac{1}{30}\begin{pmatrix} -3 & -3 \end{pmatrix}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}\begin{pmatrix} -3 \\ -3 \end{pmatrix} = -3$$
$$L_2 = -\frac{1}{30}\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = -\frac{1}{3}$$

Thus the conditional probability for $S = 1$ is

$$P(S = 1|\boldsymbol{X} = (0, 0)^T) = \frac{e^{-3}}{e^{-3} + e^{-1/3}} \approx 0.065$$

2 Determine for each of the following statements whether it is *true* or *false*, and give a brief argument for your choice: (1p each) (5p)

(a) When designing an optimal classifier for a source with N_s source states and N_d decision alternatives, using a feature vector with K elements, the optimal performance can always be achieved using some K in the interval $1 \leq K \leq \max(N_s, N_d)$.

Solution: FALSE. Any number of features can be optimal, depending on the application.

(b) Given an observed output sequence $\underline{x} = (x_1, \ldots, x_T)$ from a Hidden Markov Model λ, we can use the results of the Forward algorithm to calculate the conditional probability density

$$P\left((x_{t+1}, \ldots, x_T)|(x_1, \ldots, x_t), \lambda\right)$$

for any $1 \leq t < T$.

Solution: TRUE. The Forward algorithm can calculate a sequence of scale factors defined as $c_t = P\left(x_t|x_1, \ldots, x_{t-1}, \lambda\right)$ for any t. Therefore, using Bayes' rule, we can calculate

$$P\left((x_{t+1}, \ldots, x_T)|(x_1, \ldots, x_t), \lambda\right) = \prod_{u=t+1}^{T} c_u$$

for any $t < T$.

(c) A hidden Markov model with the following initial state probabilities and state transition probabilities produces a *stationary* random sequence.

$$\text{Initial prob.:} \quad q = \begin{pmatrix} 0.9 \\ 0.1 \end{pmatrix}; \quad \text{Transition prob.:} \quad A = \begin{pmatrix} 0.99 & 0.01 \\ 0.05 & 0.95 \end{pmatrix};$$

Solution: FALSE. $q \neq A^T q$..

(d) For a scalar random variable X with Gaussian Mixture Model (GMM) probability density function

$$f_X(x) = \sum_{m=1}^{M} w_m g_m(x),$$

the combined density function must be limited as $0 \leq f_X(x) \leq 1$ for any x.
Solution: FALSE. Probability density functions can have any non-negative value, $0 \leq f_X(x)$, but no upper limit.

(e) It is possible to design classifier discriminant functions, that normally use all K elements of the feature vector, such that the classifier can allow one feature element to be *missing* but still make optimal use of the remaining features (although possibly with reduced performance).
Solution: TRUE. The discriminant functions only need to include a pre-designed variant that uses only $K - 1$ features.

3 (HMM) Consider a discrete hidden Markov source, whose internal state sequence $\underline{S} = (S_1, \ldots, S_t, \ldots)$ is unknown. We are able to observe some elements of the output sequence $\underline{x} = (x_1, \ldots, x_t, \ldots)$ from this source. The state transition probability matrix is given by

$$A = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix}, \text{with elements } a_{ij} = P(S_{t+1} = j | S_t = i).$$

The output probability matrix is given by

$$B = \begin{pmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ 0.2 & 0.5 & 0.5 & 0.1 \end{pmatrix}, \text{with elements } b_{ik} = P(x_t = k | S_t = i).$$

(a) Find the stationary state probability distribution for this HMM source. (1p)

Solution: We solve $A^T q = q$ and we find

$$q = \begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix}.$$

(b) Assume that the initial state probability $q = (q_1, q_2)^T$ is given by the stationary probability vector from question (a). We observe $x_4 = 4$. What is the probability to make this particular observation? (i.e., calculate $P(x_4 = 4)$). (2p)

Solution: We have

$$P(x_4 = 4) = \sum_{i=1}^{2} P(x_4 = 4 | S_4 = i) P(S_4 = i) = \frac{1}{3} b_{14} + \frac{2}{3} b_{24} = \frac{1}{5},$$

since the source is stationary.

(c) Given our observation $x_4 = 4$, compute the probabilities of being in a particular hidden state at time $t = 4$. That is, calculate $P(S_4 = i | x_4 = 4)$, for $i \in \{1, 2\}$.

Solution: We have

$$P(S_4 = i | x_4 = 4) = \frac{P(S_4 = i \cap x_4 = 4)}{P(x_4 = 4)} \\ \frac{P(x_4 = 4 | S_4 = i) P(S_4 = i)}{\sum_{i=1}^{2} P(x_4 = 4 | S_4 = i) P(S_4 = i)}.$$

Therefore

$$P(S_4 = 1 | x_4 = 4) = \frac{\frac{1}{3} b_{14}}{\frac{1}{3} b_{14} + \frac{2}{3} b_{24}} = \frac{2}{3}$$
$$P(S_4 = 2 | x_4 = 4) = \frac{\frac{2}{3} b_{24}}{\frac{1}{3} b_{14} + \frac{2}{3} b_{24}} = \frac{1}{3}.$$

4 (Expectation Maximization) Consider a set of D binary variables x_i, where $i = 1, \ldots, D$, each of which has a Bernoulli distribution with parameter μ_i, so that

$$p(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i} (1 - \mu_i)^{(1-x_i)}, \tag{1}$$

where $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)^{\mathrm{T}}$. The mean and covariance of this distribution are given by

$$\begin{aligned} \mathcal{E}[\mathbf{x}] &= \boldsymbol{\mu} \\ \mathrm{cov}[\mathbf{x}] &= \mathrm{diag}\{\mu_{\mathrm{i}}(1 - \mu_{\mathrm{i}})\}. \end{aligned}$$

A finite mixture of these distributions are given by

$$p(\mathbf{x} \mid \underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\pi}}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x} \mid \boldsymbol{\mu}_k), \tag{2}$$

where $\underline{\boldsymbol{\mu}} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$, $\underline{\pi} = \{\pi_1, \ldots, \pi_K\}$. The mean and covariance of the mixture are given by

$$\mathcal{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k, \tag{3}$$

$$\mathrm{cov}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \{\Sigma_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathrm{T}}\} - \mathcal{E}[\mathbf{x}]\mathcal{E}[\mathbf{x}]^{\mathrm{T}}. \tag{4}$$

Consider that we are given a data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_n$ is drawn from a mixture of Bernoulli distributions (2). Similar to the case of the Gaussian mixture model, we assume that associated with each observation $\mathbf{x}_n$ there is a latent switch variables $\mathbf{z}_n = (z_{n1}, \ldots, z_{nK})$ consisting of a binary K-dimensional variable having a single component equal to 1, with all other components equal to 0. Let $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ denote a set of latent switch variables. The conditional distribution of $\mathbf{X}$ given the latent variable $\mathbf{Z}$ is given by

$$p(\mathbf{X} \mid \mathbf{Z}, \underline{\boldsymbol{\mu}}) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_n \mid \boldsymbol{\mu}_k)^{z_{nk}} \tag{5}$$

while the prior distribution for the latent variables is given by

$$p(\mathbf{Z} \mid \underline{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}. \tag{6}$$

In the E-step, the posterior probabilities are evaluated using the Bayes' theorem and takes the form of

$$\gamma(z_{nk}) = \mathcal{E}[z_{nk}] = \frac{\pi_k p(\mathbf{x}_n \mid \boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n \mid \boldsymbol{\mu}_j)}. \tag{7}$$

(a) Derive the complete-data log likelihood function $\ln p(\mathbf{X}, \mathbf{Z} \mid \underline{\boldsymbol{\mu}}, \underline{\pi})$ and show that if we maximize the expected complete-data log likelihood function $\mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \underline{\boldsymbol{\mu}}, \underline{\pi})]$ with respect to the $\boldsymbol{\mu}_k$, we obtain the M-step equation for $\boldsymbol{\mu}_k$ given by

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n, \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{8}$$

(1p)

Solution:

This is easily shown by calculating the derivatives of $\mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \underline{\boldsymbol{\mu}}, \underline{\pi})]$ setting them to zero and solving for μ_{ki}.

$$\mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \underline{\boldsymbol{\mu}}, \underline{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right\}$$

Using standard derivation of the above expression, we get

$$\begin{aligned} \frac{\partial}{\partial \mu_{ki}} \mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \underline{\boldsymbol{\mu}}, \underline{\pi})] &= \sum_{n=1}^{N} \gamma(z_{nk}) \left(\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}} \right) \\ &= \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni} - \sum_{n=1}^{N} \gamma(z_{nk}) \mu_{ki}}{\mu_{ki}(1 - \mu_{ki})} \end{aligned}$$

Setting this to zero and solving for μ_{ki}, we get

$$\mu_{ki} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})},$$

which equals (8) when written in vector form.

(b) Show that if we maximize the expected complete-data log likelihood function $\mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \underline{\boldsymbol{\mu}}, \underline{\pi})]$ with respect to the mixing coefficients π_k, using a Lagrange multiplier to enforce the summation constraint, we obtain the M-step equation for π_k given by

$$\pi_k = \frac{N_k}{N}. \tag{9}$$

(2p)

Hint: To enforce the summation constraint $\sum_{k=1}^{K} \pi_k = 1$, you may add a Lagrange multiplier term to the expected complete-data log likelihood function which gives:

$$\lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) + \mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \underline{\boldsymbol{\mu}}, \underline{\pi})].$$

Solution: Using the hint, we first add a Lagrange multiplier term to $\mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \underline{\boldsymbol{\mu}}, \underline{\pi})]$ to enforce the constraint $\sum_{k=1}^{K} \pi_k = 1$ which gives:

$$\lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) + \mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \underline{\boldsymbol{\mu}}, \underline{\pi})]$$

By differentiating the above with respect to π_k, we get

$$\sum_{n=1}^{N} \gamma(z_{nk}) \frac{1}{\pi_k} + \lambda = \frac{N_k}{\pi_k} + \lambda.$$

Setting this equal to zero and rearranging, we get

$$N_k = -\pi_k \lambda.$$

Summing both sides over k, making use of $\sum_{k=1}^{K} \pi_k = 1$, we see that $-\lambda = N$ and thus

$$\pi_k = \frac{N_k}{N}.$$

(c) Using the re-estimation equations for the EM algorithm, show that a mixture of Bernoulli distributions, with its parameters set to values corresponding to a maximum of the likelihood function, has the property that

$$\mathcal{E}[\mathbf{X}] = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \equiv \overline{\mathbf{x}}. \tag{10}$$

Next show that if the parameters of this model are initialized such that all components have the same mean $\boldsymbol{\mu}_k = \widehat{\boldsymbol{\mu}}$ for $k = 1, \ldots, K$, then the EM algorithm will converge after one iteration, for any choice of the initial mixing coefficients, and that this solution has the property

$$\boldsymbol{\mu}_k = \overline{\mathbf{x}}. \tag{11}$$

(2p)

Hint: The expectation of $\mathbf{X}$ under the mixture distribution is given by (3). Next use the result of part (a) and (b) to show (10) holds true. For the second part, first compute the E-step and next the M-step to show that (11) holds.

Solution: The expectation of $\mathbf{X}$ under the mixture distribution is given by

$$\mathcal{E}[\mathbf{X}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$$

Now we make use of the result of part (a) and (b), that is (8) and (9). This gives,

$$\begin{aligned}\mathcal{E}[\mathbf{X}] &= \sum_{k=1}^{K} \pi_k \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\ &= \sum_{n=1}^{N} \mathbf{x}_n \frac{1}{N} \sum_{k=1}^{K} \gamma(z_{nk}) \\ &= \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \\ &= \bar{\mathbf{x}}\end{aligned} \tag{12}$$

where we have used the fact that $\gamma(z_{nk})$ are posterior probabilities and hence $\sum_{k=1}^{K} \gamma(z_{nk}) = 1$.

Now suppose we initialize a mixture of Bernoulli distributions by setting the means to a common value $\boldsymbol{\mu}_k = \widehat{\boldsymbol{\mu}}$ for $k = 1, \ldots, K$ and then run the EM algorithm, In the E-step we first compute the responsibilities which will be given by

$$\gamma(z_{nk}) = \frac{\pi_k p(\mathbf{x}_n \mid \boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n \mid \boldsymbol{\mu}_j)} = \frac{\pi_k}{\sum_{j=1}^{K} \pi_j} = \pi_k \tag{13}$$

and are therefore independent of n. In the subsequent M-step the revised means are given by

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\ &= \frac{1}{N_k} \pi_k \sum_{n=1}^{N} \mathbf{x}_n \\ &= \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \\ &= \bar{\mathbf{x}}.\end{aligned} \tag{14}$$

5 (Bayesian Learning) Consider the Bernoulli mixture model as discussed in the previous problem. Now we assign conjugate prior distributions over the model parameters $\underline{\boldsymbol{\mu}}$, $\underline{\pi}$. For this purpose we assign a beta distribution as the prior distribution over each of the parameter vectors $\boldsymbol{\mu}_k$ given by

$$\begin{aligned} p(\underline{\boldsymbol{\mu}}) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) &= \prod_{k=1}^{K} \text{Beta}(\boldsymbol{\mu}_k \mid a_k, b_k) \\ &= \prod_{k=1}^{K} \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} \boldsymbol{\mu}_k^{a_k - 1}(1 - \boldsymbol{\mu}_k)^{b_k - 1}, \end{aligned} \tag{15}$$

where a_k and b_k control the distribution of the parameters $\boldsymbol{\mu}_k$.

Next, we assign a Dirichlet prior over each of the mixture weights π_k given by

$$\begin{aligned} p(\underline{\pi}) =& \text{Dir}(\underline{\pi} \mid \boldsymbol{\alpha}) \\ =& \frac{\Gamma(\bar{\alpha})}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)} \prod_{k=1}^{K} \boldsymbol{\mu}_k^{\alpha_k - 1}, \end{aligned} \tag{16}$$

where $\alpha_1, \ldots, \alpha_K$ are the parameters of the distribution, $\boldsymbol{\alpha}$ denotes $(\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$, and $\bar{\alpha} = \sum_{k=1}^{K} \alpha_k$. Let $\gamma(z_{nk}) = \mathcal{E}[z_{nk}]$ denote the posterior probabilities obtained from the E-step.

(a) Derive the expected complete-data log likelihood function $\mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}, \underline{\boldsymbol{\mu}}, \underline{\pi} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha})]$. (1p)

Solution:

$$\begin{aligned} \mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}, \underline{\boldsymbol{\mu}}, \underline{\pi} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha})] = \text{const.} \\ &+ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \Big\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \Big\} \\ &+ \sum_{k=1}^{K} \sum_{i=1}^{D} \big\{ (a_k - 1) \ln \mu_{ki} + (b_k - 1) \ln(1 - \mu_{ki}) \big\} \\ &+ \sum_{k=1}^{K} (\alpha_k - 1) \ln \pi_k, \end{aligned} \tag{17}$$

and we have included terms independent of $\boldsymbol{\mu}_k$ and π_k in a constant term.

(b) Show that if we maximize the expected complete-data log likelihood function $\mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}, \underline{\boldsymbol{\mu}}, \underline{\pi} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha})]$ with respect to the $\boldsymbol{\mu}_k$, we obtain the M-step equation for $\boldsymbol{\mu}_k$ given by

$$\mu_{ki} = \frac{N_k \bar{x}_{ki} + a_k - 1}{N_k + a_k - 1 + b_k - 1}, \tag{18}$$

where $\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$, $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$. (2p)

Solution: Differentiating (17) w.r.t. μ_{ki} yields

$$\sum_{n=1}^{N} \gamma(z_{nk}) \left(\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}} \right) + \frac{a_k - 1}{\mu_{ki}} - \frac{1 - b_k}{1 - \mu_{ki}} = \frac{N_k \bar{x}_{ki} + a_k - 1}{\mu_{ki}} - \frac{N_k - N_k \bar{x}_{ki} + b_k - 1}{1 - \mu_{ki}}. \tag{19}$$

Setting this equal to zero and rearranging, we get

$$\mu_{ki} = \frac{N_k \bar{\mathbf{x}}_{ki} + a_k - 1}{N_k + a_k - 1 + b_k - 1}.$$

(c) Show that if we maximize the expected complete-data log likelihood function $\mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}, \underline{\boldsymbol{\mu}}, \underline{\pi} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha})]$ with respect to the mixing coefficients π_k, using a Lagrange multiplier to enforce the summation constraint, we obtain the M-step equation for π_k given by

$$\pi_k = \frac{N_k + \alpha_k - 1}{N + \bar{\alpha} - K}$$

(2p)

Hint: To enforce the summation constraint $\sum_{k=1}^{K} \pi_k = 1$, you may add a Lagrange multiplier term to the expected complete-data log likelihood function which gives:

$$\lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) + \mathcal{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}, \underline{\boldsymbol{\mu}}, \underline{\pi} \mid \mathbf{a}, \mathbf{b}, \boldsymbol{\alpha})].$$

Solution: When maximizing w.r.t. π_k, we need to enforce the constraint $\sum_{k=1}^{K} \pi_k = 1$, which we do by adding a Lagrange multiplier term to (17). Dropping terms independent of π_k we are left with

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln \pi_k + \sum_{k=1}^{K} (\alpha_k - 1) \ln \pi_k + \lambda \left(\sum_{j=1}^{K} \pi_j - 1 \right). \tag{20}$$

Differentiating with respect to π_k, we get

$$\frac{N_k + \alpha_k - 1}{\pi_k} + \lambda \tag{21}$$

and setting this equal to zero and rearranging, we have

$$N_k + \alpha_k - 1 = -\lambda \pi_k.$$

Summing both sides over k and using $\sum_{k=1}^{K} \pi_k = 1$, we see that $-\lambda = N + \bar{\alpha} - K$. And thus

$$\pi_k = \frac{N_k + \alpha_k - 1}{N + \bar{\alpha} - K}.$$