$c-1$ them the unknown $P(\omega_j)$ reduced by the single constraint $\sum_{j=1}^{c} P(\omega_j) = 1$. Thus, the problem is not identifiable if $2c - 1 > m$.

2. PROBLEM NOT YET SOLVED

3. We are given the mixture density

$$P(x|\boldsymbol{\theta}) = P(\omega_1)\frac{1}{\sqrt{2\pi}\sigma_1}e^{-x^2/(2\sigma_1^2)} + (1-P(\omega_1))\frac{1}{\sqrt{2\pi}\sigma_2}e^{-x^2/(2\sigma_2^2)}.$$

(a) When $\sigma_1 = \sigma_2$, then $P(\omega_1)$ can take any value in the range $[0,1]$, leaving the same mixture density. Thus the density is completely unidentifiable.

(b) If $P(\omega_1)$ is fixed (and known) but not $P(\omega_1) = 0, 0.5$, or 1.0, then the model is identifiable. For those three values of $P(\omega_1)$, we cannot recover parameters for the first distribution. If $P(\omega_1) = 1$, we cannot recover parameters for the second distribution. If $P(\omega_1) = 0.5$, the parameters of the two distributions are interchangeable.

(c) If $\sigma_1 = \sigma_2$, then $P(\omega_1)$ cannot be identified because $P(\omega_1)$ and $P(\omega_2)$ are interchangeable. If $\sigma_1 \neq \sigma_2$, then $P(\omega_1)$ can be determined uniquely.

Section 10.3

4. We are given that $\mathbf{x}$ is a binary vector and that $P(\mathbf{x}|\theta)$ is a mixture of c multivariate Bernoulli distributions:

$$P(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{c} P(\mathbf{x}|\omega_i, \boldsymbol{\theta})P(\omega_i),$$

where

$$P(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = \prod_{j=1}^{d} \theta_{ij}^{x_{ij}}(1-\theta_{ij})^{1-x_{ij}}.$$

(a) We consider the log-likelihood

$$\ln P(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = \sum_{j=1}^{d} \left[x_{ij}\ln\theta_{ij} + (1-x_{ij})\ln(1-\theta_{ij})\right],$$

and take the derivative

$$\begin{aligned}
\frac{\partial \ln P(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)}{\partial \theta_{ij}} &= \frac{x_{ij}}{\theta_{ij}} - \frac{1-x_{ij}}{1-\theta_{ij}} \\
&= \frac{x_{ij}(1-\theta_{ij}) - \theta_{ij}(1-x_{ij})}{\theta_{ij}(1-\theta_{ij})} \\
&= \frac{x_{ij} - x_{ij}\theta_{ij} - \theta_{ij} + \theta_{ij}x_{ij}}{\theta_{ij}(1-\theta_{ij})} \\
&= \frac{x_{ij} - \theta_{ij}}{\theta_{ij}(1-\theta_{ij})}.
\end{aligned}$$

We set this to zero, which can be expressed in a more compact form as

$$\sum_{k=1}^{n} \hat{P}(\omega_i|x_k, \hat{\boldsymbol{\theta}}_i)\frac{x_k - \hat{\boldsymbol{\theta}}_i}{\hat{\boldsymbol{\theta}}_i(1-\hat{\boldsymbol{\theta}}_i)} = 0.$$

(b) Equation 7 in the text shows that the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_i$ must satisfy

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\nabla_{\boldsymbol{\theta}_i}\ln P(x_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = 0.$$

We can write the equation from part (a) in component form as

$$\nabla_{\boldsymbol{\theta}_i}\ln P(x_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = \frac{x_k\hat{\boldsymbol{\theta}}_i}{\hat{\boldsymbol{\theta}}_i(1-\hat{\boldsymbol{\theta}}_i)},$$

and therefore we have

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\frac{\mathbf{x}_k - \hat{\boldsymbol{\theta}}_i}{\hat{\boldsymbol{\theta}}_i(1-\hat{\boldsymbol{\theta}}_i)} = 0.$$

We assume $\hat{\boldsymbol{\theta}}_i \in (0,1)$, and thus we have

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)(\mathbf{x}_k - \hat{\boldsymbol{\theta}}_i) = 0,$$

which gives the solution

$$\hat{\boldsymbol{\theta}}_i = \frac{\sum\limits_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)x_k}{\sum\limits_{k=1}^{n} \hat{P}(\omega_i|x_k, \hat{\boldsymbol{\theta}}_i)}.$$

(c) Thus $\hat{\boldsymbol{\theta}}_i$, the maximum-likelihood estimate of $\boldsymbol{\theta}_i$, is a weighted average of the $\mathbf{x}_k$'s, with the weights being the posteriori probabilities of the mixing weights $\hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)$ for $k = 1, \ldots, n$.

5. We have a c-component mixture of Gaussians with each component of the form

$$p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) \sim N(\boldsymbol{\mu}_i, \sigma_i^2\mathbf{I}),$$

or more explicitly,

$$p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = \frac{1}{(2\pi)^{d/2}\sigma_i^d}\exp\left[-\frac{1}{2\sigma_i^2}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i)\right].$$

We take the logarithm and find

$$\ln p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i) = -\frac{d}{2}\ln(2\pi) - \frac{d}{2}\ln\sigma_i^2 - \frac{1}{2\sigma_i^2}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i),$$

and the derivative with respect to the variance is

$$\begin{aligned}\frac{\partial \ln p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)}{\partial\sigma_i^2} &= -\frac{d}{2\sigma_i^2} + \frac{1}{2\sigma_i^4}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i) \\ &= \frac{1}{2\sigma_i^4}(-d\sigma_i^2 + \|\mathbf{x}-\boldsymbol{\mu}_i\|^2).\end{aligned}$$

The maximum-likelihood estimate $\hat{\boldsymbol{\theta}}_i$ must satisfy Eq. 12 in the text, that is,

$$\sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\nabla_{\boldsymbol{\theta}_i} \ln p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i) = \mathbf{0}.$$

We set the derivative with respect to σ_i^2 to zero, that is,

$$\begin{aligned} \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\frac{\partial \ln p(\mathbf{x}_k|\omega_i, \hat{\boldsymbol{\theta}}_i)}{\partial \sigma_i^2} &= \\ \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\frac{1}{2\hat{\sigma}_i^4}(-d\hat{\sigma}_i^2 + \|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\|^2) &= 0, \end{aligned}$$

rearrange, and find

$$d\hat{\sigma}_i^2 \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i) = \sum_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\|^2.$$

The solution is

$$\hat{\sigma}_i^2 = \frac{\frac{1}{d}\sum\limits_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)\|\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i\|^2}{\sum\limits_{k=1}^{n} \hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)},$$

where $\hat{\boldsymbol{\mu}}_i$ and $\hat{P}(\omega_i|\mathbf{x}_k, \hat{\boldsymbol{\theta}}_i)$, the maximum-likelihood estimates of $\boldsymbol{\mu}_i$ and $P(\omega_i|\mathbf{x}_k, \boldsymbol{\theta}_i)$, are given by Eqs. 11–13 in the text.

6. Our c-component normal mixture is

$$p(\mathbf{x}|\alpha) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j, \alpha)P(\omega_j),$$

and the sample log-likelihood function is

$$l = \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\alpha).$$

We take the derivative with respect to α and find

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= \sum_{k=1}^{n} \frac{\partial \ln p(\mathbf{x}_k|\alpha)}{\partial \alpha} = \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k, \alpha)} \frac{\partial p(\mathbf{x}_k, \alpha)}{\partial \alpha} \\ &= \sum_{k=1}^{n} \frac{1}{p(\mathbf{x}_k, \alpha)} \frac{\partial}{\partial \alpha} \sum_{l=1}^{c} p(\mathbf{x}_k|\omega_j, \alpha)P(\omega_j) \\ &= \sum_{k=1}^{n}\sum_{j=1}^{c} \frac{p(\mathbf{x}_k|\omega_j, \alpha)P(\omega_j)}{p(\mathbf{x}_k, \alpha)} \frac{\partial}{\partial \alpha} \ln p(\mathbf{x}_k|\omega_j, \alpha) \\ &= \sum_{k=1}^{n}\sum_{j=1}^{c} P(\omega_j|\mathbf{x}_k, \alpha)\frac{\partial \ln p(\mathbf{x}_k|\omega_j, \alpha)}{\partial \alpha}, \end{aligned}$$

$$
\begin{aligned}
& -\frac{1}{n}\sum_{j=1}^{c}\sum_{k:x_k\in\Omega_j}\frac{1}{2\sigma^2}(x_k-\mu_j)^2 \\
= & \frac{1}{n}\sum_{j=1}^{c}P(\omega_j)n_j - \frac{1}{2}\ln\,(2\pi\sigma^2) - \frac{1}{n}\sum_{j=1}^{c}\sum_{k:x_k\in\Omega_j}(x_k-\mu_j)^2,
\end{aligned}
$$

where $n_j = \sum\limits_{k:x_k\in\Omega_j} 1$ is the number of points in the interval Ω_j. The result above implies

$$
\begin{aligned}
& \max_{\mu_1,\ldots,\mu_c}\frac{1}{n}\ln\,p(x_1,\ldots,x_n|\mu_1,\ldots,\mu_c) \\
\simeq & \frac{1}{n}\sum_{j=1}^{c}n_j\ln\,P(\omega_j) - \frac{1}{2}\ln\,(2\pi\sigma^2) + \frac{1}{n}\sum_{j=1}^{c}\max_{\mu_j}\sum_{k:x_k\in\Omega_j}[-(x_k-\mu_j)^2].
\end{aligned}
$$

However, we note the fact that

$$
\max_{\mu_j}\sum_{k:x_k\in\Omega_j}[-(x_k-\mu_j)^2]
$$

occurs at

$$
\begin{aligned}
\hat{\mu}_j &= \frac{\sum\limits_{k:x_k\in\Omega_j}x_k}{\sum\limits_{k:x_k\in\Omega_j}1} \\
&= \frac{\sum\limits_{k:x_k\in\Omega_j}x_k}{n_j} \\
&= \bar{x}_j,
\end{aligned}
$$

for some interval, j say, and thus we have

$$
\begin{aligned}
& \max_{\mu_1,\ldots,\mu_c}\frac{1}{n}p(x_1,\ldots,x_n|\mu_1,\ldots,\mu_c) \\
\simeq & \frac{1}{n}\sum_{j=1}^{n}n_j\ln\,P(\omega_j) - \frac{1}{2}\ln\,(2\pi\sigma^2) - \frac{1}{2\sigma^2}\frac{1}{n}\sum_{j=1}^{c}\sum_{k:x_k\in\Omega_j}(x_k-\bar{x}_j)^2 \\
= & \frac{1}{n}\sum_{j=1}^{c}n_j\ln\,P(\omega_j) - \frac{1}{2}\ln\,(2\pi\sigma^2) - \frac{1}{2\sigma^2}\frac{1}{n}\sum_{j=1}^{c}n_j\frac{1}{n_j}\sum_{k':x_k\in\Omega_j}(x_k-\bar{x}_j)^2.
\end{aligned}
$$

Thus if $n \to \infty$ (i.e., the number of independently drawn samples is very large), we have $n_j/n =$ the proportion of total samples which fall in Ω_j, and this implies (by the law of large numbers) that we obtain $P(\omega_j)$.

14. We let the mean value be denoted

$$
\bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.
$$

Then we have

$$
\frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{x})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_k-\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\mathbf{x})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_k-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\mathbf{x})
$$

$$\begin{aligned}
&= \frac{1}{n}\Big[\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_k-\bar{\mathbf{x}}) \\
&\quad +2(\bar{\mathbf{x}}-\mathbf{x})^t\mathbf{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}}) + n(\bar{\mathbf{x}}-\mathbf{x})^t\mathbf{\Sigma}^{-1})(\bar{\mathbf{x}}-\mathbf{x})\Big] \\
&= \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k\bar{\mathbf{x}})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_k\bar{\mathbf{x}}) + (\bar{\mathbf{x}}-\mathbf{x})^t\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}-\mathbf{x}) \\
&\geq \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_k-\bar{\mathbf{x}}),
\end{aligned}$$

where we used

$$\sum_{k=1}^{n}(\mathbf{x}_k-\bar{\mathbf{x}}) = \sum_{k=1}^{n}\mathbf{x}_k - n\bar{\mathbf{x}} = n\bar{\mathbf{x}} - n\bar{\mathbf{x}} = \mathbf{0}.$$

Since $\mathbf{\Sigma}$ is positive definite, we have

$$(\bar{\mathbf{x}}-\mathbf{x})^t\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}-\mathbf{x}) \geq 0,$$

with strict inequality holding if and only if $\mathbf{x} \neq \bar{\mathbf{x}}$. Thus

$$\frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{x})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_k-\mathbf{x})$$

is minimized at $\mathbf{x} = \bar{\mathbf{x}}$, that is, at

$$\mathbf{x} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.$$

15. Problem not yet solved

16. The basic operation of the algorithm is the computation of the distance between a sample and the center of a cluster which takes $O(d)$ time since each dimension needs to be compared seperately. During each iteration of the algorithm, we have to classify each sample with respect to each cluster center, which amounts to a total number of $O(nc)$ distance computations for a total complexity $O(ncd)$. Each cluster center than needs to be updated, which takes $O(cd)$ time for each cluster, therefore the update step takes $O(cd)$ time. Since we have T iterations of the classification and update step, the total time complexity of the algorithm is $O(Tncd)$.

17. We derive the equations as follows.

(a) From Eq. 14 in the text, we have

$$\ln p(\mathbf{x}_k|\omega_i,\boldsymbol{\theta}_i) = \ln\frac{|\mathbf{\Sigma}_i^{-1}|^{1/2}}{(2\pi)^{d/2}} - \frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu}_i)^t\mathbf{\Sigma}_i^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_i).$$

It was shown in Problem 11 that

$$\frac{\partial \ln p(\mathbf{x}_k|\omega_i,\boldsymbol{\theta}_i)}{\partial\sigma_{pq}(i)} = \left(1-\frac{\delta_{pq}}{2}\right)[\sigma_{pq}(i) - (x_p(k)-\mu_p(i))(x_q(k)-\mu_q(i))].$$