13.2.1

The log-likelihood is

\[ -\frac{1}{2} \sum_{k=1}^{g} \sum_{i \in C_k} (x_i - \mu_k)' \Sigma^{-1} (x_i - \mu_k) - \frac{n}{2} \log(2\pi|\Sigma|^{-1}) \]

where

\[ C_k = \{ i = 1, \ldots, n : \gamma_i = k \} \]

The last term is constant (\( \Sigma \) is assumed known), so we are maximizing

\[ \sum_{k=1}^{g} \sum_{i \in C_k} (x_i - \mu_k)' \Sigma^{-1} (x_i - \mu_k) \]

over \( \mu_k \) and \( C_k \) for \( k = 1, \ldots, g \).

Fix \( \gamma \) (so that the \( C_k \) are given). The maximum is obtained for

\[ \hat{\mu}_k = \overline{x}_k = 1/n_k \sum_{i \in C_k} x_i \]

where \( n_k = \text{card}(C_k) \).

Therefore, \( \hat{\gamma} \) maximizes

\[ \sum_{k=1}^{g} \sum_{i \in C_k} (x_i - \overline{x}_k)' \Sigma^{-1} (x_i - \overline{x}_k) \]

Since

\[ (x_i - \overline{x}_k)' \Sigma^{-1} (x_i - \overline{x}_k) \]
is a scalar, it is equal to its trace, and therefore

\[(x_i - \mu_k)'\Sigma^{-1}(x_i - \mu_k) = \text{trace}[(x_i - \mu_k)'\Sigma^{-1}(x_i - \mu_k)]
= \text{trace}[(x_i - \mu_k)(x_i - \mu_k)'\Sigma^{-1}]
\]

by applying

\[\text{trace}(AB) = \text{trace}(BA)\]

Summing over \(k\) and \(i \in C_k\), we get

\[
\sum_{k=1}^{g} \sum_{i \in C_k} (x_i - \mu_k)'\Sigma^{-1}(x_i - \mu_k) = \text{trace}(W\Sigma^{-1})
\]

that \(\hat{\gamma}\) maximizes.

Now, observe that

\[
\sum_{i \in C_k} (x_i - \mu_k)'\Sigma^{-1}(x_i - \mu_k) = \sum_{i \in C_k} (x_i - \bar{x})'\Sigma^{-1}(x_i - \bar{x}) - n_k(\bar{x}_k - \bar{x})'\Sigma^{-1}(\bar{x}_k - \bar{x})
\]

and so

\[
\sum_{k=1}^{g} \sum_{i \in C_k} (x_i - \mu_k)'\Sigma^{-1}(x_i - \mu_k) = \sum_{i=1}^{n} (x_i - \bar{x})'\Sigma^{-1}(x_i - \bar{x}) - \sum_{k=1}^{g} n_k(\bar{x}_k - \bar{x})'\Sigma^{-1}(\bar{x}_k - \bar{x})
\]

And the first term does not depend on \(\gamma\).

**13.2.3**

This was inspired by the copy of Daniel Rubin.

For more details, see *Estimating the components of a mixture of normal distributions*, N.E. Day, Biometrika, 56.

Let

\[x^j \sim \mathcal{N}(\mu_j, \Sigma) \quad j = 1, 2\]

and

\[\epsilon^1 \sim \text{Bernoulli}(p_1)\]
jointly independent. Also, 

$$\epsilon^2 = 1 - \epsilon^1$$

Then \(x\) is equal in law to \(\epsilon^1 x^1 + \epsilon^2 x^2\), so we suppose that is how the \(x\) are generated, that is to say:

1. choose class \(j = 1, 2\) with probability \(p_j\);
2. draw a vector from \(\mathcal{N}(\mu_j, \Sigma)\).

(a) It follows that, by independence,

\[
E[x] = E[\epsilon^1 x^1 + \epsilon^2 x^2] = E[\epsilon^1]E[x_1] + E[\epsilon^2]E[x_2] = p_1 \mu_1 + p_2 \mu_2
\]

\[
E[xx'] = E[\epsilon^1]E[x_1 x_1'] + E[\epsilon^2]E[x_2 x_2'] = p_1 (\Sigma + \mu_1 \mu_1') + p_2 (\Sigma + \mu_2 \mu_2')
\]

That gives the values for \(m\) and \(V\).

Let us move to ML estimation. If the classes (i.e. the \(\epsilon_i^j\)) are given, the m.l.e s are easy to compute:

\[
\hat{p}_j = \frac{n_j}{n}
\]

\[
\hat{\mu}_j = \frac{1}{n_j} \sum_{i \in C_j} x_i
\]

\[
\hat{\Sigma} = \frac{1}{n} \left[ \sum_{i \in C_1} (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)' + \sum_{i \in C_2} (x_i - \hat{\mu}_2)(x_i - \hat{\mu}_2)' \right]
\]

where

\[C_j = \{i = 1, ..., n : \epsilon_i^j = 1\} \quad n_j = \text{card}(C_j)\]

Notice that they depend on \(\{\epsilon_i^j\}\). This dependence is not explicitly shown for clarity.

This gives

\[
\hat{m} = \hat{p}_1 \hat{\mu}_1 + \hat{p}_2 \hat{\mu}_2 = \bar{x}
\]

and using

\[
n = n_1 + n_2
\]

\[
\bar{x} = \frac{n_1 \mu_1 + n_2 \mu_2}{n}
\]
we get
\[ \hat{V} = S \]
Thus \( \hat{m} \) and \( \hat{V} \) do not depend on \( \{\epsilon_i^2\} \), so those are their values when the classes are not given.

To find the expressions for \( \hat{a} \) and \( \hat{b} \), just express \( a \) and \( b \) in terms of \( V \) instead of \( \Sigma \), i.e. replace \( \Sigma \) by
\[ V - p_1p_2(\mu_1 - \mu_2)(\mu_1 - \mu_2)' \]
and then plug in the m.l.e. estimator for the corresponding parameters.

(b) Again, if the classes are given,
\[ \hat{p}_j = n_j/n = 1/n \sum_{i=1}^{n} \{\epsilon_i^j = 1\} \]
and so
\[ \hat{p}_2/\hat{p}_1 = \frac{\sum_{i=1}^{n} \{\epsilon_i^2 = 1\}}{\sum_{i=1}^{n} \{\epsilon_i^1 = 1\}} \]
\[ \hat{\mu}_j = \frac{\sum_{i=1}^{n} x_i \{\epsilon_i^j = 1\}}{\sum_{i=1}^{n} \{\epsilon_i^j = 1\}} \]
\[ \hat{\Sigma} = 1/n \sum_{i=1}^{n} [(x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)'\{\epsilon_i^1 = 1\} + (x_i - \hat{\mu}_2)(x_i - \hat{\mu}_2)\{\epsilon_i^2 = 1\}] \]
If the classes are not given, we estimate the indicator r.v. by their m.l.e., which amounts to replace \( \{\epsilon_i^2 = 1\} \) by \( \hat{P}(j|x_i) \).

Now,
\[ \log[\hat{P}(1|x)/\hat{P}(2|x)] = \hat{a}'x + \hat{b} \]
so \( \hat{P}(1|x) \) is a function of \( x, \hat{a} \) and \( \hat{b} \). And by the formulas above, \( \hat{p}_j, \hat{\mu}_j \) are functions of \( \{x_i\}, \hat{a} \) and \( \hat{b} \), while \( \hat{V} = S \) is a function of \( \{x_i\} \) only. The conclusion follows.
13.3.2

The following figures are self-explanatory:

![Diagram of Single linkage method]

0.0 0.1 0.2 0.3 0.4 0.5 0.6

2 4 6 7
We see that there is a typo in the book: reverse the results.

From those results, we see that complete linkage (that produces (1,3,6), (2,4) and (5)) is closer to (1,3,6) and (2,4,5).