Solutions to the Exercises of Section 6.2.

6.2.1. First we show the hint: for $z > 0$, $1 - z + \log z \leq 0$ (note the inequality is backward in the text). Let $g(z) = 1 - z + \log z$. Then $g'(z) = -1 + (1/z)$ and $g''(z) = -1/z^2$. Thus, $g(z)$ reaches its maximum at $z = 1$, so for all $z > 0$, $g(z) \leq g(1) = 0$.
Now for $f(x, y) = xy(\log y - \log x)/(y - x)$ for $x \neq y$, (and we may define $f(x, x) = x$ by continuity), we may compute

$$\frac{\partial f}{\partial y} = \frac{(y-x)[x(\log y - \log x) + x] - xy(\log y - \log x)}{(y-x)^2} = \frac{-x^2}{(y-x)^2} g(y/x) \geq 0.$$

Hence, $f(x, y)$ is increasing in y for each x. And since $f(y, x) = f(x, y)$, we also have that $f(x, y)$ is increasing in x for each y.

6.2.2. (a) An optimal classification rule is $\phi(i|x) = I_{[\mu_i, \mu_{i+1})}(x)$ for $i = 1, \ldots, k$, where $\mu_{k+1} = \infty$. In words, classify into the population with μ_i nearest of those to the left of x.

(b) Optimal classification rules are of the form,

$$\phi(i|x) = \begin{cases} 0 & \text{if} \quad x < \mu_i \\ \text{any} & \text{if} \quad x \geq \mu_i. \end{cases}$$

In words, classify into any population with μ_i to the left of x.

6.2.3. We classify x into the population with the smallest value of $(1/2)(\mathbf{x} - \mu_i)^T(\mathbf{x} - \mu_i) - \log p_i$. Let S_i denote the region for classification into population i. In all three cases, S_1 is a square centered at the origin, and S_2 is the set $\{(x_1, x_2) : x_1 > 0, |x_2| < x_1\}$ with S_1 removed, and each of S_3, S_4 and S_5 is the set S_2 rotated $90°$, $180°$ and $270°$ respectively. The problem is to find the set S_1. We concentrate on finding the boundary between S_1 and S_2. We prefer 1 to 2 if $(\mu_2 - \mu_1)^T x > (1/2)(\|\mu_2\|^2 - \|\mu_1\|^2) + \log(p_i/p_j)$, which reduces to $x_1 > .5 + \log(p_1/p_2)$.

(a) We have $\log(p_1/p_2) = \log 2 = .693\ldots$, so that S_1 is the square $\{(x_1, x_2) : |x_i| < 1.193\ldots\}$. It is interesting to note that S_i does not contain μ_i for $i = 2, 3, 4, 5$.

(b) We have $\log(p_1/p_2) = 0$, so that S_1 is the unit square $\{(x_1, x_2) : |x_1| < .5, |x_2| < .5\}$.

(c) We have $\log(p_1/p_2) = \log(1/2) = -.693\ldots$, so that S_1 is empty.

6.2.4. We rank the populations into the order (1,2) if

$$\frac{1}{1 + (x_1 - \mu_1)^2} \frac{1}{1 + (x_2 - \mu_2)^2} > \frac{1}{1 + (x_1 - \mu_2)^2} \frac{1}{1 + (x_2 - \mu_1)^2}.$$

This is a quartic inequality in x_1, x_2, μ_1 and μ_2. However, it gives equality if either $x_1 = x_2$ or $\mu_1 = \mu_2$. Thus both $x_1 - x_2$ and $\mu_1 - \mu_2$ must factor out. When this is done, we obtain, under the assumption $x_1 < x_2$ and after some tedious algebra, the quadratic inequality,

$$(x_1 - \mu_1)(x_2 - \mu_2) + (x_1 - \mu_2)(x_2 - \mu_1) < 2.$$

In the special case $\mu_1 = -1$ and $\mu_2 = 1$, this inequality reduces to $x_1 x_2 < 2$. If $x_1 > x_2$, this inequality must be reversed. This gives as the region in the plane for which we rank the populations into the order (1,2) as $\{(x_1, x_2) : (x_1 < x_2 \text{ and } x_1 x_2 < 2) \text{ or } (x_1 > x_2 \text{ and } x_1 x_2 > 2)\}$.

6.2.5. We are given $\mu_1 < \mu_2 < \cdots < \mu_n$. We are to show that if $x_1 \leq x_2 \leq \cdots \leq x_n$, then

$$\prod_{i=1}^{n} f(x_i|\mu_i) \geq \prod_{i=1}^{n} f(x_i|\mu_{\nu_i}) \qquad (*)$$

for any permutation, $(\nu_1, \nu_2, \ldots, \nu_n)$ of $(1, 2, \ldots, n)$. We are given that the likelihood ratio, $f(x|\mu_2)/f(x|\mu_1)$ is nondecreasing in x when $\mu_1 < \mu_2$. This means that if $\mu_1 < \mu_2$ and $x_1 \leq x_2$, then

$$f(x_1|\mu_1) f(x_2|\mu_2) \geq f(x_1|\mu_2) f(x_2|\mu_1). \qquad (**)$$

Let us say that the observations (x_1, x_2) are discordant with the parameters (μ_1, μ_2) if we have $x_1 < x_2$ and $\mu_1 > \mu_2$ or $x_1 > x_2$ and $\mu_1 < \mu_2$. The inequality (**) says that if observations are discordant with their parameters, then the density cannot decrease if we switch parameters. Therefore, in the product on the right side of (*), if two observations are discordant with their parameters, we may switch them without decreasing the product. This may be continued until all observations are concordant with their parameters.

6.2.6. The parameter space and the action space are both the set of permutations of $(1, 2)$. For $\theta = \theta_1 = (1, 2)$, we have

$$f(x_1, x_2|\theta_1) = \left(\frac{1}{2\pi}\right)^m \exp\{-\frac{1}{2}\|x_1 - \mu_1\|^2 - \frac{1}{2}\|x_2 - \mu_2\|^2\}$$

and for $\theta = \theta_2 = (2, 1)$, we have

$$f(x_1, x_2|\theta_2) = \left(\frac{1}{2\pi}\right)^m \exp\{-\frac{1}{2}\|x_1 - \mu_2\|^2 - \frac{1}{2}\|x_2 - \mu_1\|^2\}$$

One easily finds

$$f(x_1, x_2|\theta_1)/f(x_1, x_2|\theta_2) = \exp\{(x_1 - x_2)^T(\mu_1 - \mu_2)\}.$$

We take action $a = \theta_1$ if this is greater than 1, or equivalently, if $(x_1 - x_2)^T(\mu_1 - \mu_2) > 0$. Geometrically speaking, we take $a = \theta_1$ if the angle between $x_1 - x_2$ and $\mu_1 - \mu_2$ is less than $90°$, and $a = \theta_2$ if this angle is greater than $90°$.

6.2.7. (a) The prior distribution has density $g(\nu, \theta_1, \ldots, \theta_k) \propto p_\nu \exp\{-\frac{1}{2}\sum_1^k \theta_i^T \Sigma^{-1}\theta_i\}$. The density of the observations given the parameters is

$$f_{X_1,\ldots,X_k,Y}(x_1, \ldots, x_k, y|\nu, \theta_1, \ldots, \theta_k) \propto \exp\{-\frac{1}{2}\sum_1^k (x_i - \theta_i)^T(x_i - \theta_i) - \frac{1}{2}(y - \theta_\nu)^T(y - \theta_\nu)\}$$

The posterior density is proportional to the product of these, namely

$$\begin{aligned}
&\propto p_\nu \exp\{-\frac{1}{2}[\sum_1^k \theta_i^T\Sigma^{-1}\theta_i + \sum_1^k \theta_i^T\theta_i - 2\sum_1^k x_i^T\theta_i + \theta_\nu^T\theta_\nu - 2y^T\theta_\nu]\} \\
&= p_\nu \exp\{-\frac{1}{2}[\sum_{i\neq\nu} \theta_i^T(\Sigma^{-1} + I)\theta_i - 2\sum_{i\neq\nu} x_i^T\theta_i + \theta_\nu^T(\Sigma^{-1} + 2I)\theta_\nu - 2(y + x_\nu)^T\theta_\nu]\} \\
&\propto p_\nu \exp\{-\frac{1}{2}[\sum_{i\neq\nu} (\theta_i - (\Sigma^{-1} + I)^{-1}x_i)^T(\Sigma^{-1} + I)(\theta_i - (\Sigma^{-1} + I)^{-1}x_i)]\} \\
&\quad \cdot \exp\{-\frac{1}{2}(\theta_\nu - (\Sigma^{-1} + 2I)^{-1}(y + x_\nu))^T(\Sigma^{-1} + 2I)(\theta_\nu - (\Sigma^{-1} + 2I)^{-1}(y + x_\nu))]\} \\
&\quad \cdot \exp\{\frac{1}{2}[(x_\nu + y)^T(\Sigma^{-1} + 2I)^{-1}(x_\nu + y) - x_\nu^T(\Sigma^{-1} + I)^{-1}x_\nu]\}
\end{aligned}$$

So the posterior distribution may be described as follows: ν is chosen with probability proportional to

$$p'_\nu \propto p_\nu \exp\{\frac{1}{2}[(x_\nu + y)^T(\Sigma^{-1} + 2I)^{-1}(x_\nu + y) - x_\nu^T(\Sigma^{-1} + I)^{-1}x_\nu]\},$$

and given ν, the parameters $\theta_1, \ldots, \theta_k$ are independent with $\theta_i \in \mathcal{N}((\Sigma^{-1} + I)^{-1}x_i, (\Sigma^{-1} + I)^{-1})$ for $i \neq \nu$ and $\theta_\nu \in \mathcal{N}((\Sigma^{-1} + 2I)^{-1}(y + x_\nu), (\Sigma^{-1} + 2I)^{-1})$. We now simplify the expression for p'_ν by writing the exponent as a quadratic form in y. Let $B = (\Sigma^{-1} + I)^{-1} - (\Sigma^{-1} + 2I)^{-1}$.

$$\begin{aligned}
p'_\nu &\propto p_\nu \exp\{\frac{1}{2}[x_\nu^T(\Sigma^{-1} + 2I)^{-1}x_\nu + 2y^T(\Sigma^{-1} + 2I)^{-1}x_\nu - x_\nu^T(\Sigma^{-1} + I)^{-1}x_\nu]\} \\
&= p_\nu \exp\{-\frac{1}{2}[x_\nu^T B x_\nu - 2y^T(\Sigma^{-1} + 2I)^{-1}B^{-1}Bx_\nu]\} \\
&\propto p_\nu \exp\{-\frac{1}{2}[(x_\nu - B^{-1}(\Sigma^{-1} + 2I)^{-1}y)^T B(x_\nu - B^{-1}(\Sigma^{-1} + 2I)^{-1}y)]\} \\
&= p_\nu \exp\{-\frac{1}{2}[(x_\nu - \hat{y})^T B(x_\nu - \hat{y})]\}
\end{aligned}$$

where $\hat{y} = B^{-1}(\Sigma^{-1} + 2I)^{-1}y = ((\Sigma^{-1} + I)^{-1}(\Sigma^{-1} + 2I) - I)^{-1}y$. This is a simpler form than given in the text.

(b) The Bayes rule chooses a as that integer ν for which p'_ν is the largest, or equivalently, for which $(x_\nu - \hat{y})^T B(x_\nu - \hat{y})$ is the smallest. When $\Sigma = \sigma^2 I$, then $\hat{y} = ((1 + \sigma^2)/\sigma^2)y$ and B is a constant so the Bayes rule chooses a equal to that ν for which x_ν is closest to $\hat{y}$.