Selected solutions to Exercise Sheet 2

Question 4

```r
library(mvtnorm)

# Variables
muT = c(20,5)
muB = c(30,2)
Sigma = matrix(c(100, 15, 15, 9), 2, 2)

# Density of multivariate normal
fT <- function(x) dmvnorm(x, mean = muT, sigma = Sigma)
fB <- function(x) dmvnorm(x, mean = muB, sigma = Sigma)

# Priors
piT <- 0.4
piB <- 0.2

# Posterior for given individual i is proportional to
etaTi <- fT(c(25, 7)) * piT
etaBi <- fB(c(20, 3)) * piB
```

The density function for a multivariate normal distribution with mean μ_T and covariance matrix Σ is given by:

$$f(x|\mu_T, \Sigma) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu_T)'\Sigma^{-1}(x-\mu_T)\right) \tag{1}$$

Similarly, for a multivariate normal distribution with mean μ_B and covariance matrix Σ, we have:

$$f(x|\mu_B, \Sigma) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu_B)'\Sigma^{-1}(x-\mu_B)\right) \tag{2}$$

We get the posterior probabilities using Bayes rule:

$$\eta_k(x) := P(Y = k|X = x; \mu_k, \Sigma) \propto f_k(x|\mu_k, \Sigma)\pi_k$$

where $\pi_k = P(Y = k)$ is the prior probability.

To get the Bayes Decision Boundary note that we are indifferent for all $x \in \mathbb{R}^2$ if and only if:

$$\delta_B(x) = \delta_T(x) \iff \eta_B(x) = \eta_T(x).$$

Plugging in, taking logs, and ignoring the constant in front of the exponential (which does not depend on k) we get:

$$\log(\pi_T) - \frac{1}{2}(x-\mu_T)^T\Sigma^{-1}(x-\mu_T) = \log(\pi_B) - \frac{1}{2}(x-\mu_B)^T\Sigma^{-1}(x-\mu_B),$$

which simplifies to:

$$\log\left(\frac{\pi_T}{\pi_B}\right) - \frac{1}{2}(\mu_T^T\Sigma^{-1}\mu_T - \mu_B^T\Sigma^{-1}\mu_B) + (\mu_T - \mu_B)^T\Sigma^{-1}x = 0, \tag{3}$$

where we have made use of the fact that the first term of

$$(x - \mu_k)^T\Sigma^{-1}(x - \mu_k) = x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}\mu_k$$

does not depend on k.

The constants in equation (3) can be evaluated as follows

```
    (
        log(piT / piB)
        -0.5 * (t(muT) %*% solve(Sigma) %*% muT -
                t(muB) %*% solve(Sigma) %*% muB)
    )
```

```
##           [,1]
## [1,] 3.359814
```

```
    (
        t(muT - muB) %*% solve(Sigma)
    )
```

```
##      [,1]      [,2]
## [1,] -0.2 0.6666667
```

where we use `%*%` for matrix multiplication and `solve` for inverting a matrix. We obtain

$$3.36 - 0.2x_1 + 0.67x_2 = 0 \tag{4}$$

which we can solve for x_2 as a function of x_1:

$$x_2 = -5 + 0.3x_1 \tag{5}$$

This defines the decision boundary between Bus and Train.

Question 7

- Correlation: Positively correlated through (omitted) distance
- Normal distribution would put positive probability to negative time and cost. We had to choose a small variance to mitigate this.
- Linear vs. quadratic discriminant: There is an underlying individual choice problem associated with this statistical problem. People have utility over leisure (free time) and consumption (money). If they are not perfect substitutes, this would create non-linear discriminants. (Note: this is advanced – we will come back to this in more detail).