SOLUTIONS TO THE EXAM QUESTIONS IN STK 4040/9040, DECEMBER 13/12-11

PROBLEM 1

a) See page 681 in J & W.

b) The table gives the following distance matrix (H = humans, G = gorillas, C = chimpanzees, O = orangutans)

$$\begin{array}{c} \\ H \\ G \\ C \\ O \end{array} \begin{array}{c} \begin{array}{cccc} H & G & C & O \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 1 & 2 & 0 & \\ 9 & 6 & 8 & 0 \end{bmatrix} \end{array}$$

The smallest distance is between humans and chimpanzees, so we cluster these two species. The clustering is at distance 1.

When computing the new distance matrix, we use complete linkage. Thus the distance between (H, C) and G is 3 and the distance between (H, C) and O is 9. This gives the new distance matrix

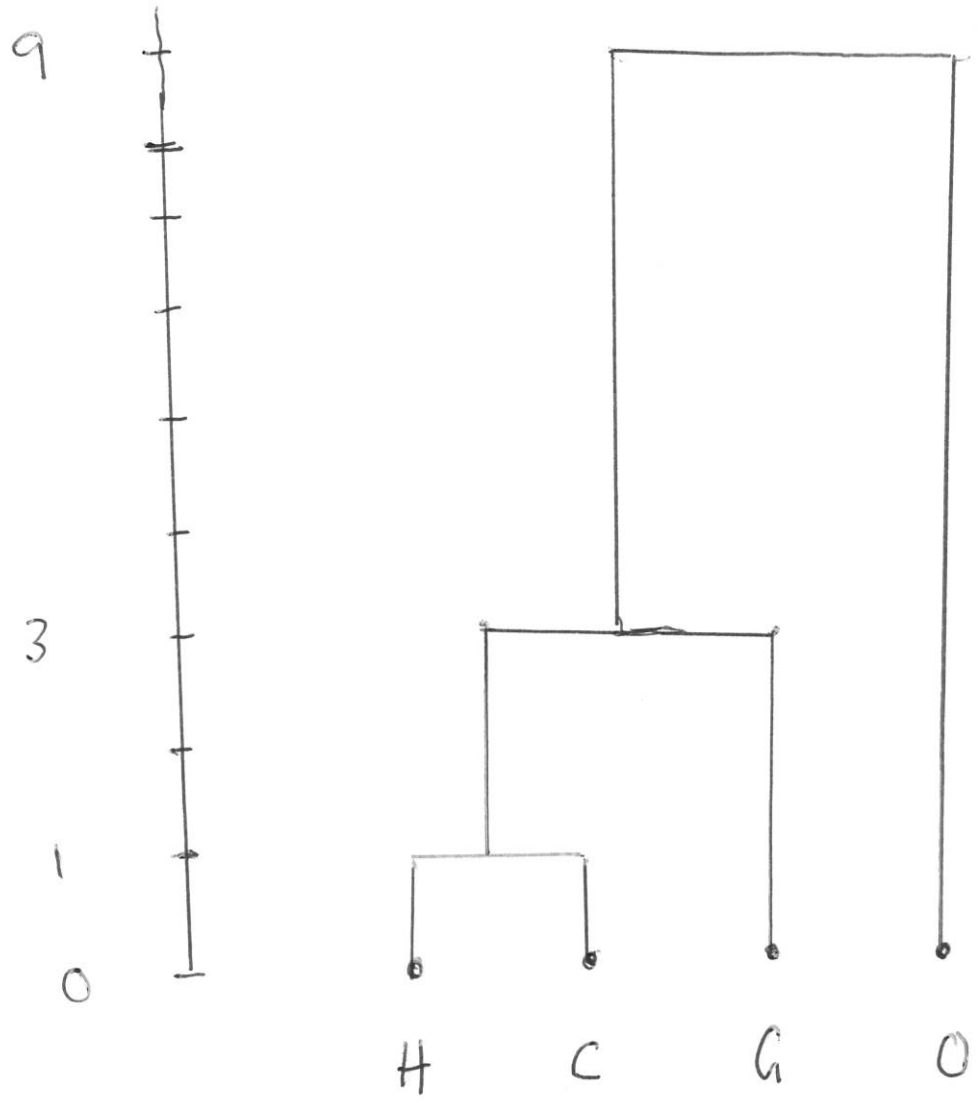
$$\begin{array}{c} \\ (H,C) \\ G \\ O \end{array} \begin{array}{c} \begin{array}{ccc} (H,C) & G & O \end{array} \\ \begin{bmatrix} 0 & & \\ 3 & 0 & \\ 9 & 6 & 0 \end{bmatrix} \end{array}$$

The smallest distance is now between (H, C) and G so we merge these two clusters at distance 3.

Now the distance between (H, C, G) and O is 9

So we finally cluster (H, C, G) and O at distance 9.

The dendrogram shows how the clusters are created and the distances at which the mergers occur:



From the dendrogram we see that humans and chimpanzees have the closest relation, and that they are fairly close to gorillas. Orangutans are less related with the other three species.

PROBLEM 2

a) For $j = 1, 2, \dots, n$ we introduce $Z_j = a' X_j$. Then the Z_j's are independent and normally distributed with means $\mu_Z = a'\mu$ and variances $\sigma_Z^2 = a'\Sigma a$. By standard results from earlier courses we know that

$$T_a = \frac{\bar{Z} - \mu_Z}{S_Z / \sqrt{n}}$$

is t-distributed with $n-1$

degrees of freedom. Here

$$S_z^2 = \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \bar{z})^2.$$

Now

$$\bar{z} = \frac{1}{n} \sum_{j=1}^{n} a' x_j = a' \bar{x}$$

and

$$S_z^2 = \frac{1}{n-1} \sum_{j=1}^{n} \{ (a' x_j - a' \bar{x}) (a' x_i - a' \bar{x})' \}$$

$$= a' \left(\frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(x_i - \bar{x})' \right) a$$

$$= a' S a$$

Hence

$$T_a = \frac{a' \bar{x} - a' \mu}{\sqrt{a' S a} / \sqrt{n}} = \sqrt{n} \, \frac{a' \bar{x} - a' \mu}{\sqrt{a' S a}}$$

is t-distributed with $n-1$ d.f.

b) It is known that $\bar{X}$ and S are independent and that

$$\bar{X} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$$

$$(n-1)S \sim W_{p,n-1}(\Sigma)$$

cf. D.2 in the collection of formulas.

Thus

$$Z = \sqrt{n}(\bar{X}-\mu) \sim N_p(0, \Sigma)$$

$$W = (n-1)S \sim W_{p,n-1}(\Sigma)$$

By C.9 in the collection of formulas we then have that

$$Z'\left(\frac{W}{n-1}\right)^{-1} Z$$

$$= n(\bar{X}-\mu)' S^{-1} (\bar{X}-\mu)$$

is distributed as

$$\frac{(n-1)p}{(n-1)-p+1} F_{p,(n-1)-p+1}$$

i.e as $\frac{(n-1)p}{n-p}$ times a

F-distributed random variable with p and $n-p$ degrees of freedom.

c) By question (b) and the result given just before it, we have that ($c>0$)

$$P(-c \le T_a \le c \text{ for all } a)$$

$$= P(T_a^2 \le c^2 \text{ for all } a)$$

$$= P(\max_a T_a^2 \le c^2)$$

$$= P\{n(\bar{X}-\mu)' S^{-1} (\bar{X}-\mu) \le c^2\}$$

$$= 1-\alpha$$

if we choose

$$c^2 = \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$

Now we have that

$$-c \leq T_a \leq c$$

if and only if

$$-c \leq \sqrt{n}\, \frac{a'\bar{x} - a'\mu}{\sqrt{a'Sa}} \leq c$$

which is equivalent to

$$a'\bar{x} - c\sqrt{\frac{a'Sa}{n}} \leq a'\mu \leq a'\bar{x} + c\sqrt{\frac{a'Sa}{n}}$$

Thus intervals of the form

$$a'\bar{x} \pm c\sqrt{\frac{a'Sa}{n}}$$

simultaneously contain $a'\mu$

for all vectors a. Here

$$c = \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)}$$

d) For the data on the boys we have $n = 6$ and $p = 3$. Using the F-table with $\alpha = 0.05$, we then have

$$c = \sqrt{\frac{5 \cdot 3}{6-3} F_{3,3}(0.05)}$$

$$= \sqrt{\frac{5 \cdot 3}{3} \cdot 9.28} = 6.81$$

Thus the 95% simultaneous confidence intervals for $a'\mu$ are given as

$$a'\bar{X} \pm 6.81 \cdot \sqrt{\frac{a'Sa}{6}}$$

For $a_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ we have that

$a_3' \bar{X} = \bar{X}_3$ (i.e mean MVAC value)

and $a_3' S a_3 = S_{33}$ (i.e. Sample variance of the MVAC values)

This gives the confidence interval

$$14.5 \pm 6.81 \sqrt{\frac{1.90}{6}}$$

i.e

$$14.5 \pm 3.8$$

This gives

$$[10.7, 18.3]$$

For $a_4 = \begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix}$ we have that

$$a_4' \bar{X} = -\bar{X}_1 + 2\bar{X}_2 - \bar{X}_3$$

$$= -82.0 + 2 \cdot 60.2 - 14.5 = 23.9$$

and from the R-output

$a_4' S a_4 = 9.788$. This gives the confidence interval

$$23.9 \pm 6.81 \sqrt{\frac{9.788}{6}}$$

i.e

$$23.9 \pm 8.7$$

This gives

$$[15.2, 32.6]$$

e) We see that the simultaneous confidence intervals for $a_1'\mu = \mu_1$, $a_2'\mu = \mu_2$ and $a_3'\mu = \mu_3$ contain the expected values of height, breast circumference and MUAC for boys from the low countries

However the confidence interval for $a_4'\mu = -\mu_1 + 2\mu_2 - \mu_3$ does not contain the corresponding value for boys from the low countries, which is

$$-90 + 2 \cdot 58 - 16 = 10$$

We know that Hotelling's test rejects the null hypothesis if and only if there exists a vector a such that the simultaneous confidence interval for $a'\mu$ does not contain the corresponding null hypothesis value $a'\mu_0$.

In our case a_4 is such a vector, so the null hypothesis will be rejected at the 5% level.

We see that the boys from the mountain area tend to be smaller and have lower MVAL values than those from the low countries, but their breast circumference tends to be larger. And this is exactly what is captured by the contrast $a_4'\mu$.

PROBLEM 3

a) We make a misclassification if an observation from Π_1 is classified as being from Π_2 or the other way around. Thus

$$TPM = P(\text{classify } X \text{ from } \Pi_1 \text{ as } \Pi_2)$$

$$+ P(\text{classify } X \text{ from } \pi_2 \text{ as } \pi_1)$$

$$= P(\pi_1)\, P(X \in R_2 \mid \pi_1) + P(\pi_2)\, P(X \in R_1 \mid \pi_2)$$

$$= \frac{1}{2} \int_{R_2} f_1(x)\,dx + \frac{1}{2} \int_{R_1} f_2(x)\,dx$$

We now use that

$$1 = \int f_1(x)\,dx = \int_{R_1} f_1(x)\,dx + \int_{R_2} f_1(x)\,dx$$

and obtain

$$TRM = \frac{1}{2}\left(1 - \int_{R_1} f_1(x)\,dx\right) + \frac{1}{2} \int_{R_1} f_2(x)\,dx$$

$$= \frac{1}{2} \int_{R_1} \{f_2(x) - f_1(x)\}\,dx + \frac{1}{2}$$

b) The integral

$$\int_{R_1} \{f_2(x) - f_1(x)\}\, dx$$

obtains it minimal value if we let

$$R_1 = \{x : f_2(x) - f_1(x) \leq 0\}$$

Thus TPM is minimized by choosing

$$R_1 = \{x : f_1(x) \geq f_2(x)\}$$

$$= \{x : f_1(x)/f_2(x) \geq 1\}$$

c) We now assume that $(i = 1, 2)$

$$f_i(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-(x-\mu_i)' \Sigma^{-1} (x-\mu_i)/2}$$

Then

$$\frac{f_1(x)}{f_2(x)} = \exp\{-(x-\mu_1)'\Sigma^{-1}(x-\mu_1)/2 + (x-\mu_2)\Sigma^{-1}(x-\mu_2)/2\}$$

$$= \exp\{(\mu_1-\mu_2)'\Sigma^{-1}x - \tfrac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)\}$$

Taking logs on both sides of this equality, we obtain the optimal classification rule

$$R_1 = \{x : (\mu_1-\mu_2)'\Sigma^{-1}x - \tfrac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) \geq 0\}$$

d) We have $X \sim N(\mu_i, \Sigma)$ when X comes from population π_i; $i=1,2$

By known results for linear transformations of a multivariate

normal vector (cf. C.4 in collection of formulas), we have that

$$Z \sim N_p(\mu_{zi}, \Sigma_z)$$

when X comes from population π_i.

Here

$$\mu_{zi} = V^{-1/2}(\mu_i - \bar{\mu})$$

$$\Sigma_z = V^{-1/2} \Sigma V^{-1/2}$$

Note that

$$\mu_{z1} = V^{-1/2}(\mu_1 - \mu_2)/2$$

$$\mu_{z2} = V^{-1/2}(\mu_2 - \mu_1)/2$$

e) From the result in (c) we have that the optimal rule classifies an observation $z_0 = V^{-1/2}(x_0 - \bar{\mu})$ as coming from Π_1 provided that

$$(\mu_{z1} - \mu_{z2})' \Sigma_z^{-1} z_0 - \frac{1}{2}(\mu_{z1} - \mu_{z2})' \Sigma_z^{-1} (\mu_{z1} + \mu_{z2}) \geq 0$$

Now $\mu_{z1} + \mu_{z2} = 0$, and hence the rule becomes

$$(\mu_{z1} - \mu_{z2})' \Sigma_z^{-1} z_0 \geq 0$$

Inserting the expressions for μ_{zi} $(i = 1, 2)$, Σ_z and z_0 in this inequality, we obtain

$$[V^{-1/2}(\mu_1 - \mu_2)]' [V^{-1/2} \Sigma V^{-1/2}]^{-1} V^{-1/2}(x_0 - \bar{\mu}) \geq 0$$

i.e

$$(\mu_1 - \mu_2)' V^{-1/2} V^{1/2} \Sigma^{-1} V^{1/2} V^{-1/2}(x_0 - \bar{\mu}) \geq 0$$

Thus the optimal rule classifies the observation to population π_1 provided that

$$(\mu_1 - \mu_2)' \Sigma^{-1} (x_0 - \bar{\mu}) \geq 0$$

i.e. provided that

$$(\mu_1 - \mu_2)' \Sigma^{-1} x_0 - (\mu_1 - \mu_2)' \Sigma^{-1} \bar{\mu} \geq 0$$

And this is the same rule as we obtained in question c.

Thus the classification rule becomes the same for non-standardized and standardized variables.