# Solution for Exam 2011 Problem 3

ANYM, 20191111

## Q 3.1

We calculate the Hotellings $T^2$ using Mahalanobis's Distance. Section 4.1.2 page 283:

$$T^2 = \frac{nm}{n+m}(\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \mathbf{S}^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}).$$

Mahalanobis's distance is the last part, see page 333

$$D^2 = \| \widehat{\mu}_1 - \widehat{\mu}_2 \|^2_{\widehat{\Sigma}^{-1}} = (\widehat{\mu}_1 - \widehat{\mu}_2)^T \widehat{\mathbf{\Sigma}}^{-1} (\widehat{\mu}_1 - \widehat{\mu}_2)$$

Mahalanobis's Distance is denoted squared distance in the SAS-output:

| Squared Distance to mask | | | | |
|---|---|---|---|---|
| **From mask** | **10** | **11** | **12** | **13** |
| **10** | 0 | 35.99468 | 4.32708 | 74.35995 |
| **11** | 35.99468 | 0 | 29.06112 | 36.86522 |
| **12** | 4.32708 | 29.06112 | 0 | 82.37583 |
| **13** | 74.35995 | 36.86522 | 82.37583 | 0 |

We further need the frequency, i.e. the number of observations:

| Class Level Information | | | | | |
|---|---|---|---|---|---|
| **mask** | **Variable Name** | **Frequency** | **Weight** | **Proportion** | **Prior Probability** |
| **10** | 10 | 16 | 16.0000 | 0.122137 | 0.250000 |
| **11** | 11 | 11 | 11.0000 | 0.083969 | 0.250000 |
| **12** | 12 | 88 | 88.0000 | 0.671756 | 0.250000 |
| **13** | 13 | 16 | 16.0000 | 0.122137 | 0.250000 |

We get

$$T^2 = \frac{16 \cdot 16}{16 + 16} 74.35995 = 8 \cdot 74.35995$$

Answer 2.

## Q 3.2

We consider the hypothesis that all for geological units have the same mean value. This a one-sided MANOVA, see section 4.3.1. Theorem 4.25 on page 304 gives us:

**Theorem 4.25**

The ratio test for the test of the hypothesis $H_0$ against $H_1$ is given by the critical region

$$\{y_{11},\ldots,y_{kn_k} \,|\, \frac{\det(\mathbf{w})}{\det(\mathbf{t})} \leq U(p,k-1,n-k)_\alpha\}.$$

Where $p$=6 is the dimension/number of variables in our observations, $k$=4 is the number of groups, and $n$=131 is the number observations.

$$U(p,k-1,n-k) = U(6,4-1,131-4) = U(6,3,127)$$

Answer 1.

## Q 3.3

The usual test statistic for additional information (section 5.4.1) of band 4-6 between unit 10 and 13. We use Theorem 5.21 on page 356.

**Theorem 5.21**

The critical region for testing the hypothesis that the last $p-q$ variables do not contribute to the discrimination between the populations $\pi_1$ and $\pi_2$, i.e. the hypothesis that $\Delta^2_{(2|1)} = 0$ against all alternatives is

$$\left\{x_{11},\cdots,x_{2n_2} \,\middle|\, \frac{n_1+n_2-p-1}{p-q}\frac{d^2-d_1^2}{(n_1+n_2)(n_1+n_2-2)/(n_1n_2)+d_1^2} > F(p-q,n_1+n_2-p-1)_{1-\alpha}\right\}$$

Here $d^2$ and $d_1^2$ are the observed values of $D^2$ and $D_1^2$.

We already have $d^2$=*74.35995* from before. We find $d_1^2$=*55.29380*. Inserting:

$$\frac{16+16-6-1}{3}\cdot\frac{74.35995-55.29380}{\frac{(16+16)(16+16-2)}{16\cdot 16}+55.29380} = \frac{25}{3}\cdot\frac{256\cdot(74.35995-55.29380)}{32\cdot 30+256\cdot 55.29380}$$

The answer is 1.

## Q 3.4

The test above follow – if true – the following F-distribution:

$$F(p-q,n_1+n_2-p-1) = F(3,16+16-6-1) = F(3,25)$$

Answer 3

## Q 3.5

The number of misclassifications increase by how much by omitting the 3 variables.

We look at the confusion tables.

Before

| Number of Observations and Percent Classified into mask | | | | | |
|---|---|---|---|---|---|
| From mask | 10 | 11 | 12 | 13 | Total |
| 10 | 16<br>100.00 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 16<br>100.00 |
| 11 | 0<br>0.00 | 10<br>90.91 | 0<br>0.00 | 1<br>9.09 | 11<br>100.00 |
| 12 | 8<br>9.09 | 0<br>0.00 | 80<br>90.91 | 0<br>0.00 | 88<br>100.00 |
| 13 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 16<br>100.00 | 16<br>100.00 |
| Total | 24<br>18.32 | 10<br>7.63 | 80<br>61.07 | 17<br>12.98 | 131<br>100.00 |
| Priors | 0.25 | 0.25 | 0.25 | 0.25 | |

After:

| Number of Observations and Percent Classified into mask | | | | | |
|---|---|---|---|---|---|
| From mask | 10 | 11 | 12 | 13 | Total |
| 10 | 15<br>93.75 | 0<br>0.00 | 1<br>6.25 | 0<br>0.00 | 16<br>100.00 |
| 11 | 0<br>0.00 | 9<br>81.82 | 1<br>9.09 | 1<br>9.09 | 11<br>100.00 |
| 12 | 17<br>19.32 | 0<br>0.00 | 71<br>80.68 | 0<br>0.00 | 88<br>100.00 |
| 13 | 0<br>0.00 | 0<br>0.00 | 0<br>0.00 | 16<br>100.00 | 16<br>100.00 |
| Total | 32<br>24.43 | 9<br>6.87 | 73<br>55.73 | 17<br>12.98 | 131<br>100.00 |
| Priors | 0.25 | 0.25 | 0.25 | 0.25 | |

We see the increase is from 9 to 20, i.e. an increase of 11. Answer 4.

## Q 3.5

The generalized variance in given in Definition 1.61 page 58.

> **Definition 1.59**
>
> Let the $p$-dimensional vector $\boldsymbol{X}$ have the variance-covariance matrix $\boldsymbol{\Sigma}$. By the term ***the generalized variance*** of $\boldsymbol{X}$ we mean the determinant of the variance-covariance matrix, i.e.
>
> $$\text{gen.var.}(\boldsymbol{X}) = \det(\boldsymbol{\Sigma}).$$

It is simple the determinant of the dispersion matrix. In the output:

| Pooled Covariance Matrix Information | |
|---|---|
| Covariance Matrix Rank | Natural Log of the Determinant of the Covariance Matrix |
| 6 | 15.30586 |

SAS has taken the log to that quantity and we simply need to raise it in e, to get the correct answer: 2