Solutions

1. (i) The sample covariance matrix is $\mathbf{S} := (s_{jk})$ and the sample correlation matrix is $\mathbf{R} := (r_{jk})$, where

$$s_{jj} := s_j^2 = \frac{\sum_{i=1}^n (y_{ij} - \bar{y}_j)^2}{n-1} = \frac{\sum_{i=1}^n y_{ij}^2 - n\bar{y}_j^2}{n-1},$$

where $\bar{y}_j$ is the mean of the j^{th} variable, and the sample covariance of the j^{th} and k^{th} variables is

$$s_{jk} = \frac{\sum_{i=1}^n (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)}{n-1} = \frac{\sum_{i=1}^n y_{ij}y_{ik} - n\bar{y}_j\bar{y}_k}{n-1}$$

and the sample correlation between the variables j^{th} and k^{th} is

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{s_{jk}}{\sqrt{s_j^2}\sqrt{s_k^2}}.$$

In our case

$$\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, \bar{y}_3)' = (27.17, 4.88, 3.22)',$$

and*

$$\begin{aligned}
s_1^2 &= s_{11} = \frac{33^2 + 20^2 + 35^2 + 10^2 + 30^2 + 35^2 - 6 \cdot 27.17^2}{5} = 101.95, \\
s_2^2 &= s_{22} = \frac{4.5^2 + 2.9^2 + 10.1^2 + 2.3^2 + 1.5^2 + 8.0^2 - 6 \cdot 4.88^2}{5} = 11.87, \\
s_3^2 &= s_{33} = \frac{2.7^2 + 2.8^2 + 3.3^2 + 3.2^2 + 2.9^2 + 4.4^2 - 6 \cdot 3.22^2}{5} = 0.36
\end{aligned}$$

and

$$\begin{aligned}
s_{12} &= s_{21} = \frac{33 \cdot 4.5 + 20 \cdot 2.9 + 35 \cdot 10.1 + 10 \cdot 2.3 + 30 \cdot 1.5 + 35 \cdot 8.0 - 6 \cdot 27.17 \cdot 4.88}{5} = 22.49, \\
s_{13} &= s_{31} = \frac{33 \cdot 2.7 + 20 \cdot 2.8 + 35 \cdot 3.3 + 10 \cdot 3.2 + 30 \cdot 2.9 + 35 \cdot 4.4 - 6 \cdot 27.17 \cdot 3.22}{5} = 1.74, \\
s_{23} &= s_{23} = \frac{4.5 \cdot 2.7 + 2.9 \cdot 2.8 + 10.1 \cdot 3.3 + 2.3 \cdot 3.2 + 1.5 \cdot 2.9 + 8.0 \cdot 4.4 - 6 \cdot 4.88 \cdot 3.22}{5} = 1.25.
\end{aligned}$$

Hence

$$\mathbf{S} = \begin{pmatrix} 101.95 & 22.49 & 1.74 \\ 22.49 & 11.87 & 1.25 \\ 1.74 & 1.25 & 0.36 \end{pmatrix}$$

and

$$\begin{aligned}
r_{11} &= \frac{s_{11}}{\sqrt{s_1^2}\sqrt{s_1^2}} = 1, \quad r_{22} = \frac{s_{22}}{\sqrt{s_2^2}\sqrt{s_2^2}} = 1, \quad r_{33} = \frac{s_{33}}{\sqrt{s_3^2}\sqrt{s_3^2}} = 1, \\
r_{12} &= \frac{s_{12}}{\sqrt{s_1^2}\sqrt{s_2^2}} = \frac{22.49}{\sqrt{101.95}\sqrt{11.87}} = 0.65, \\
r_{13} &= \frac{s_{13}}{\sqrt{s_1^2}\sqrt{s_3^2}} = \frac{1.74}{\sqrt{101.95}\sqrt{0.36}} = 0.29, \\
r_{23} &= \frac{s_{23}}{\sqrt{s_2^2}\sqrt{s_3^2}} = \frac{1.25}{\sqrt{11.87}\sqrt{0.36}} = 0.61.
\end{aligned}$$

[1] **It is not necessary to compute all the coefficents !**

and

$$\mathbf{R} = \begin{pmatrix} 1 & 0.65 & 0.29 \\ 0.65 & 1 & 0.61 \\ 0.29 & 0.61 & 1 \end{pmatrix}.$$

Also we have the possibility to use the formula (and to make some more complicated computations):

$$\mathbf{R} = \mathbf{D}^{-1} \cdot \mathbf{S} \cdot \mathbf{D}^{-1},$$

where

$$\mathbf{D} = \begin{pmatrix} \sqrt{s_1^2} & 0 & 0 \\ 0 & \sqrt{s_2^2} & 0 \\ 0 & 0 & \sqrt{s_3^2} \end{pmatrix} = \begin{pmatrix} \sqrt{101.95} & 0 & 0 \\ 0 & \sqrt{11.87} & 0 \\ 0 & 0 & \sqrt{0.36} \end{pmatrix} = \begin{pmatrix} 10.10 & 0 & 0 \\ 0 & 3.45 & 0 \\ 0 & 0 & 0.6 \end{pmatrix}.$$

Hence

$$\begin{aligned} \mathbf{R} &= \begin{pmatrix} 10.10 & 0 & 0 \\ 0 & 3.45 & 0 \\ 0 & 0 & 0.6 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 101.95 & 22.49 & 1.74 \\ 22.49 & 11.87 & 1.25 \\ 1.74 & 1.25 & 0.36 \end{pmatrix} \cdot \begin{pmatrix} 10.10 & 0 & 0 \\ 0 & 3.45 & 0 \\ 0 & 0 & 0.6 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0.999 & 0.645 & 0.287 \\ 0.645 & 0.997 & 0.603 \\ 0.287 & 0.603 & 1.0 \end{pmatrix} = \begin{pmatrix} 1 & 0.65 & 0.29 \\ 0.65 & 1 & 0.61 \\ 0.29 & 0.61 & 1 \end{pmatrix} \end{aligned}$$

(ii) The statistical distance of the p–dimensional point $\mathbf{y}$ to $\boldsymbol{\mu}$ is $(\mathbf{y}-\boldsymbol{\mu})' \mathbf{S}^{-1} (\mathbf{y}-\boldsymbol{\mu})$, where $\mathbf{S}$ is the covariance matrix.

If $(\mathbf{y}-\boldsymbol{\mu})' \mathbf{S}^{-1} (\mathbf{y}-\boldsymbol{\mu}) = (\text{constant}) = c^2$ and $\mathbf{S} = \begin{pmatrix} 2 & 2 \\ 2 & 3 \end{pmatrix}$ (hence $p = 2$), then we obtain

$$\begin{aligned} c^2 &= (\mathbf{y}-\boldsymbol{\mu})' \mathbf{S}^{-1} (\mathbf{y}-\boldsymbol{\mu}) = \begin{pmatrix} y_1-\mu_1 & y_2-\mu_2 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ 2 & 3 \end{pmatrix}^{-1} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix} \\ &= \begin{pmatrix} y_1-\mu_1 & y_2-\mu_2 \end{pmatrix} \begin{pmatrix} 3/2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix} \\ &= -(y_1-\mu_1)\left((y_2-\mu_2) - \frac{3}{2}(y_1-\mu_1)\right) - (y_2-\mu_2)(y_1-\mu_1-(y_2-\mu_2)) \\ &= \frac{3}{2}(y_1-\mu_1)^2 - 2(y_1-\mu_1)(y_2-\mu_2) + (y_2-\mu_2)^2 \end{aligned}$$

which represent an rotated ellipse (therefore, the contours of constant statistical distances are ellipses).

Indeed, the determinants associated to the conic

$$\frac{3}{2}(y_1-\mu_1)^2 - 2(y_1-\mu_1)(y_2-\mu_2) + (y_2-\mu_2)^2 - c^2 = 0$$

are

$$I := 3/2 + 1 = 5/2, \quad \delta := \begin{vmatrix} 3/2 & -1 \\ -1 & 1 \end{vmatrix} = 1/2, \quad \Delta := \begin{vmatrix} 3/2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & -c^2 \end{vmatrix} = -c^2/2.$$

Hence our conic is ellipse since $\Delta \neq 0$, $\delta > 0$ and $\dfrac{\Delta}{\delta} < 0$.

If $\mathbf{y}$ satisfies that $(\mathbf{y}-\boldsymbol{\mu})' \mathbf{S}^{-1} (\mathbf{y}-\boldsymbol{\mu}) \leq c^2$, then the point $\mathbf{y}$ is inside the ellipse or on the ellipse.

2. The pooled covariance matrix is

$$\begin{aligned}\mathbf{S}_{\text{pl}} \quad &: \quad = \frac{(30-1)\,\mathbf{S}_1 + (30-1)\,\mathbf{S}_2}{30+30-2} \\ &= \quad \frac{(30-1)\begin{pmatrix} 5 & -1 & -3 \\ -1 & 2 & 2 \\ -3 & 2 & 1 \end{pmatrix} + (30-1)\begin{pmatrix} 2 & 3 & 2 \\ 3 & 1 & 5 \\ 2 & 5 & 1 \end{pmatrix}}{30+30-2} = \begin{pmatrix} 7/2 & 1 & -1/2 \\ 1 & 3/2 & 7/2 \\ -1/2 & 7/2 & 1 \end{pmatrix}.\end{aligned}$$

We compute ($p = 2$ and $q = 1$)

$$\begin{aligned}T^2_{p+q} \quad &= \quad \frac{n_1 \cdot n_2}{n_1+n_2}\,(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \cdot \mathbf{S}^{-1}_{\text{pl}} \cdot (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \\ &= \quad \frac{30\cdot 30}{30+30}\left(\begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 \\ 4 \\ 3 \end{pmatrix}\right)' \begin{pmatrix} 7/2 & 1 & -1/2 \\ 1 & 3/2 & 7/2 \\ -1/2 & 7/2 & 1 \end{pmatrix}^{-1} \left(\begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 \\ 4 \\ 3 \end{pmatrix}\right) \\ &= \quad \frac{309}{34} = 9.0882,\end{aligned}$$

since

$$\begin{pmatrix} 7/2 & 1 & -1/2 \\ 1 & 3/2 & 7/2 \\ -1/2 & 7/2 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 43/170 & 11/170 & -1/10 \\ 11/170 & -13/170 & 3/10 \\ -1/10 & 3/10 & -1/10 \end{pmatrix} = \begin{pmatrix} 0.2529 & 0.0647 & -0.1 \\ 0.0647 & -0.0765 & 0.3 \\ -0.1 & 0.3 & -0.1 \end{pmatrix}.$$

We take $\alpha = 0.01$. The critical value $T^2_{\alpha,p+q,n_1+n_2-2} = T^2_{0.01,3,58}$ is given by the formula*:

$$T^2_{0.01,3,58} = \frac{58\cdot 3}{58-3+1} F_{0.01,3,58-3+1} = \frac{58\cdot 3}{58-3+1} F_{0.01,3,56} \simeq \frac{87}{28} F_{0.01,3,60} = \frac{87}{28} 4.13 = 12.833.$$

Therefore we accept the null hypothesis H_0 because $T^2_{p+q} = 9.0882 < 12.833$.

After that we compute

$$\begin{aligned}T^2_p \quad &= \quad \frac{30\cdot 30}{30+30}\left(\begin{pmatrix} 3 \\ 2 \end{pmatrix} - \begin{pmatrix} 4 \\ 3 \end{pmatrix}\right)' \begin{pmatrix} 3/2 & 7/2 \\ 7/2 & 1 \end{pmatrix}^{-1} \left(\begin{pmatrix} 3 \\ 2 \end{pmatrix} - \begin{pmatrix} 4 \\ 3 \end{pmatrix}\right) \\ &= \quad \frac{270}{43} = 6.2791,\end{aligned}$$

since

$$\begin{pmatrix} 3/2 & 7/2 \\ 7/2 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} -4/43 & 14/43 \\ 14/43 & -6/43 \end{pmatrix} = \begin{pmatrix} -0.0930 & 0.3256 \\ 0.3256 & -0.1395 \end{pmatrix}.$$

We have

$$T^2\left(y_1 | y_1, y_2\right) := (\nu - p)\,\frac{T^2_{p+q} - T^2_p}{\nu + T^2_p} = (58-2)\,\frac{9.0882 - 6.2791}{58 + 6.2791} = 2.4473,$$

since $\nu = 30+30-2 = 58$.

We take $\alpha = 0.01$. The critical value $T^2_{\alpha,q,\nu-p} = T^2_{0.01,1,56}$ is given by

$$T^2_{0.01,1,56} = \frac{56\cdot 1}{56-1+1} F_{0.01,1,56-1+1} = F_{0.01,1,56} \simeq F_{0.01,1,60} = 7.08.$$

Therefore we accept the null hypothesis H_0 which states that $\mathbf{x} = (y_1)$ is redundant because $T^2\left(y_1 | y_1, y_2\right) = 2.4473 < 7.08$. Hence the first variable doesn't add a significant amount of separation to the last two variables.

*The connection between the tables of the Fisher distribution and the Hotteling distribution is $T^2_{\alpha,a,b} = \frac{b\cdot a}{b-a+1} F_{\alpha,a,b-a+1}$.

3. We have $n_1 = 3$, $n_2 = 2$, $n_3 = 3$, $p = 2$, $k = 3$.

We test the null hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 \,.$$

We consider the significance level $\alpha = 0.05$.

In the matrix form, we write

$$\begin{pmatrix} \begin{bmatrix} 6 \\ 8 \end{bmatrix} & \begin{bmatrix} 4 \\ 6 \end{bmatrix} & \begin{bmatrix} 2 \\ 6 \end{bmatrix} \\ \begin{bmatrix} 3 \\ 8 \end{bmatrix} & \begin{bmatrix} -4 \\ 3 \end{bmatrix} & \\ \begin{bmatrix} -3 \\ 2 \end{bmatrix} & \begin{bmatrix} -4 \\ -5 \end{bmatrix} & \begin{bmatrix} 3 \\ -3 \end{bmatrix} \end{pmatrix}$$

Using the standard notations we obtain

$$\mathbf{y}_{1.} = (y_{1.1}, y_{1.2})' = \begin{bmatrix} 12 \\ 20 \end{bmatrix}, \quad \mathbf{y}_{2.} = (y_{2.1}, y_{2.2})' = \begin{bmatrix} -1 \\ 11 \end{bmatrix}, \quad \mathbf{y}_{3.} = (y_{3.1}, y_{3.2})' = \begin{bmatrix} -4 \\ -6 \end{bmatrix}$$

and

$$\begin{aligned} \bar{\mathbf{y}}_{1.} &= (\bar{y}_{1.1}, \bar{y}_{1.2})' = \frac{\mathbf{y}_{1.}}{3} = \begin{bmatrix} 12/3 \\ 20/3 \end{bmatrix} = \begin{bmatrix} 4 \\ 6.667 \end{bmatrix}, \\ \bar{\mathbf{y}}_{2.} &= (\bar{y}_{2.1}, \bar{y}_{2.2})' = \frac{\mathbf{y}_{2.}}{2} = \begin{bmatrix} -1/2 \\ 11/2 \end{bmatrix} = \begin{bmatrix} -0.500 \\ 5.500 \end{bmatrix}, \\ \bar{\mathbf{y}}_{3.} &= (\bar{y}_{3.1}, \bar{y}_{3.2})' = \frac{\mathbf{y}_{3.}}{3} = \begin{bmatrix} -4/3 \\ -6/3 \end{bmatrix} = \begin{bmatrix} -1.333 \\ -2 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} \mathbf{y}_{..} &= (y_{..1}, y_{..2})' = \sum_{i=1}^{3} \mathbf{y}_{i.} = \begin{bmatrix} 7 \\ 25 \end{bmatrix} \\ \bar{\mathbf{y}}_{..} &= (\bar{y}_{..1}, \bar{y}_{..2})' = \frac{\sum_{i=1}^{3} \mathbf{y}_{i.}}{8} = \begin{bmatrix} 7/8 \\ 25/8 \end{bmatrix} = \begin{bmatrix} 0.875 \\ 3.125 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} N &= 3 + 2 + 3 = 8 \\ \nu_H &= k - 1 = 2, \quad \nu_E = \sum_{i=1}^{k} n_i - k = N - k = 5. \end{aligned}$$

We have

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ij} \,, \quad i = \overline{1,k} \,, \quad j = \overline{1,n_i} \,,$$

where

$$\boldsymbol{\mu}_i \text{ is the mean of the } i^{th} \text{ population} \quad \text{and} \quad \boldsymbol{\alpha}_i := \boldsymbol{\mu}_i - \boldsymbol{\mu} \quad \text{and} \quad \boldsymbol{\epsilon}_{ij} := \mathbf{y}_{ij} - \boldsymbol{\mu}_i \,.$$

We obtain two (since $p = 2$) matrix equalities

$$\begin{pmatrix} 6 & 4 & 2 \\ 3 & -4 & \\ -3 & -4 & 3 \end{pmatrix} = \begin{pmatrix} \frac{7}{8} & \frac{7}{8} & \frac{7}{8} \\ \frac{7}{8} & \frac{7}{8} & \\ \frac{7}{8} & \frac{7}{8} & \frac{7}{8} \end{pmatrix} + \begin{pmatrix} \frac{25}{8} & \frac{25}{8} & \frac{25}{8} \\ -\frac{11}{8} & -\frac{11}{8} & \\ -\frac{53}{24} & -\frac{53}{24} & -\frac{53}{24} \end{pmatrix} + \begin{pmatrix} 2 & 0 & -2 \\ \frac{7}{2} & -\frac{7}{2} & \\ -\frac{5}{3} & -\frac{8}{3} & \frac{13}{3} \end{pmatrix}$$

$$\text{observation} = \text{mean} + \text{treatment effect} + \text{residual}$$

and

$$\begin{pmatrix} 8 & 6 & 6 \\ 8 & 3 & \\ 2 & -5 & -3 \end{pmatrix} = \begin{pmatrix} \frac{25}{8} & \frac{25}{8} & \frac{25}{8} \\ \frac{25}{8} & \frac{25}{8} & \\ \frac{25}{8} & \frac{25}{8} & \frac{25}{8} \end{pmatrix} + \begin{pmatrix} \frac{85}{24} & \frac{85}{24} & \frac{85}{24} \\ \frac{19}{8} & \frac{19}{8} & \\ -\frac{41}{8} & -\frac{41}{8} & -\frac{41}{8} \end{pmatrix} + \begin{pmatrix} \frac{4}{3} & -\frac{2}{3} & -\frac{2}{3} \\ \frac{5}{2} & -\frac{5}{2} & \\ 4 & -3 & -1 \end{pmatrix}$$

observation = mean + treatment effect + residual

Now SSH_{11} is the sum of the square of the elements of the third matrix from the first matrix equality:

$$\mathrm{SSH}_{11} = 3 \cdot (25/8)^2 + 2 \cdot (-11/8)^2 + 3 \cdot (-53/24)^2 = \frac{1145}{24} = 47.708$$

and SSH_{22} is the sum of the square of the elements of the third matrix from the second matrix equality:

$$\mathrm{SSH}_{22} = 3 \cdot (85/24)^2 + 2 \cdot (19/8)^2 + 3 \cdot (-41/8)^2 = \frac{3065}{24} = 127.71.$$

Now SSE_{11} is the sum of the square of the elements of the fourth matrix from the first matrix equality:

$$\mathrm{SSE}_{11} = 2^2 + 0^2 + (-2)^2 + (7/2)^2 + (-7/2)^2 + (-5/3)^2 + (-8/3)^2 + (13/3)^2 = \frac{367}{6} = 61.167$$

and SSE_{22} is the sum of the square of the elements of the third matrix from the second matrix equality:

$$\mathrm{SSE}_{22} = (4/3)^2 + (-2/3)^2 + (-2/3)^2 + (5/2)^2 + (-5/2)^2 + 4^2 + (-3)^2 + (-1)^2 = \frac{247}{6} = 41.167.$$

It remains to obtain the **cross products**:

$$\begin{aligned} \mathrm{SPH}_{12} &= \sum_{i=1}^{3} n_i \left(\bar{y}_{i.1} - \bar{y}_{..1}\right)\left(\bar{y}_{i.2} - \bar{y}_{..2}\right) = \sum_{i=1}^{3} \frac{y_{i.1} y_{i.2}}{n_i} - \frac{y_{..1} y_{..2}}{N} \\ \mathrm{SPE}_{12} &= \sum_{i=1}^{3} \sum_{j=1}^{n_i} \left(y_{ij1} - \bar{y}_{i.1}\right)\left(y_{ij2} - \bar{y}_{i.2}\right) = \sum_{i=1}^{3} \sum_{j=1}^{n_i} y_{ij1} y_{ij2} - \sum_{i=1}^{3} \frac{y_{i.1} y_{i.2}}{n_i}. \end{aligned}$$

Therefore

SPH_{12} is the sum of the products of the elements of the third matrix from the first matrix equality with their corresponding elements of the third matrix from the second matrix equality and

SPE_{12} is the sum of the products of the elements of the fourth matrix from the first matrix equality with their corresponding elements of the fourth matrix from the second matrix equality:

$$\begin{aligned} \mathrm{SPH}_{12} &= 3 \cdot 25/8 \cdot 85/24 + 2 \cdot (-11/8) \cdot 19/8 + 3 \cdot (-53/24) \cdot (-41/8) = 485/8 = 60.625 \\ \mathrm{SPE}_{12} &= 2 \cdot 4/3 + 0 \cdot (-2/3) + (-2) \cdot (-2/3) + 7/2 \cdot 5/2 + (-7/2) \cdot (-5/2) + (-5/3) \cdot 4 \\ &\quad + (-8/3) \cdot (-3) + 13/3 \cdot (-1) = \tfrac{37}{2} = 18.5 \end{aligned}$$

We obtain

$$\mathbf{H} = \begin{pmatrix} \mathrm{SSH}_{11} & \mathrm{SPH}_{12} \\ \mathrm{SPH}_{12} & \mathrm{SSH}_{22} \end{pmatrix} = \begin{pmatrix} 47.71 & 60.63 \\ 60.63 & 127.71 \end{pmatrix}$$

and

$$\mathbf{E} = \begin{pmatrix} \mathrm{SSE}_{11} & \mathrm{SPE}_{12} \\ \mathrm{SPE}_{12} & \mathrm{SSE}_{22} \end{pmatrix} = \begin{pmatrix} 61.17 & 18.5 \\ 18.5 & 41.17 \end{pmatrix},$$

and therefore

$$\mathbf{E} + \mathbf{H} = \begin{pmatrix} 108.88 & 79.13 \\ 79.13 & 168.88 \end{pmatrix}.$$

The Wilks' lambda test is

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|} = \frac{2176.1}{12126.0} = 0.1795,$$

the parameters are

$$p = 2, \quad \nu_H = k - 1 = 2, \quad \nu_E = N - k = 5.$$

Therefore we can apply formula

$$F = \frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}} \frac{\nu_E - 1}{\nu_H} = \frac{1 - \sqrt{0.1795}}{\sqrt{0.1795}} \frac{5 - 1}{2} = 2.721$$

which should be compared with

$$F_{\alpha, 2\nu_H, 2(\nu_E - 1)} = F_{0.05,4,8} = 3.84.$$

Since $2.721 < 7.01$, the null hypothesis H_0 is accepted at the 5% level of significance.

4. The multivariate linear regression model is $\mathbf{Y} = \mathbf{XB} + \mathbf{\Xi}$, where $\mathbf{Y} = \begin{pmatrix} 5 & -3 \\ 3 & -1 \\ 4 & -1 \\ 2 & 2 \\ 1 & 3 \end{pmatrix}$ and

$$\mathbf{X} = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{X}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{pmatrix}.$$

We estimate $\mathbf{B}$ by $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. We take $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$ and $\hat{\mathbf{\Xi}} = \mathbf{Y} - \hat{\mathbf{Y}}$.

Hence

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.1 \end{pmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 5 & -3 \\ 3 & -1 \\ 4 & -1 \\ 2 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 15 & 0 \\ -9 & 15 \end{pmatrix}.$$

Consequently,

$$\hat{\mathbf{B}} = \begin{pmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.1 \end{pmatrix} \begin{pmatrix} 15 & 0 \\ -9 & 15 \end{pmatrix} = \begin{pmatrix} 3.0 & 0 \\ -0.9 & 1.5 \end{pmatrix}$$

and we obtain the regression lines

$$\hat{y}_1 = 3 - 0.9x_1 \quad \text{and} \quad \hat{y}_2 = 0 + 1.5x_2.$$

We have

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}} = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 3.0 & 0 \\ -0.9 & 1.5 \end{pmatrix} = \begin{pmatrix} 4.8 & -3.0 \\ 3.9 & -1.5 \\ 3.0 & 0 \\ 2.1 & 1.5 \\ 1.2 & 3.0 \end{pmatrix}.$$

Hence

$$\hat{\mathbf{\Xi}} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{pmatrix} 5 & -3 \\ 3 & -1 \\ 4 & -1 \\ 2 & 2 \\ 1 & 3 \end{pmatrix} - \begin{pmatrix} 4.8 & -3.0 \\ 3.9 & -1.5 \\ 3.0 & 0 \\ 2.1 & 1.5 \\ 1.2 & 3.0 \end{pmatrix} = \begin{pmatrix} 0.2 & 0 \\ -0.9 & 0.5 \\ 1.0 & -1 \\ -0.1 & 0.5 \\ -0.2 & 0 \end{pmatrix}$$

Can be easily verified the formulas

$$\mathbf{X}'\hat{\mathbf{\Xi}} = \mathbf{0} \quad \text{and} \quad \hat{\mathbf{Y}}'\hat{\mathbf{\Xi}} = 0.$$

- The meaning of $\hat{\mathbf{B}}$:
  The matrix $\hat{\mathbf{B}}$, given by $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, is the least squares estimator since "minimizes" the matrix
  $$(\mathbf{Y} - \mathbf{X}\mathbf{B})'(\mathbf{Y} - \mathbf{X}\mathbf{B}).$$
- The geometrical interpretation of the least squares technique in the linear regression models:
  If we consider case $q = 1$, we obtain that the ordinary regression line of y on x minimizes the sum of squares of *vertical distances* from the points y_i to the line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. We can reverse the roles and we see that the ordinary regression line of x on y minimizes the sum of squares of *horizontal distances* from the points x_i to the line $\hat{x} = \hat{\beta}_0 + \hat{\beta}_1 y$.
- The connection with the geometrical interpretation of the first principal component:
  The first principal component line represent a perpendicular regression since this line is such that minimizes the total sum of squared *perpendicular distances* from the points to this line. The first principal component line lies between the other two regression lines.

5. (i) The eigenvalues of $\mathbf{S}$ are $\lambda_1 = 6 \quad \text{and} \quad \lambda_2 = 1$. The corresponding eigenvectors are

$$\mathbf{a}_1 = \begin{pmatrix} 2a & a \end{pmatrix}' \quad \text{and} \quad \mathbf{a}_2 = \begin{pmatrix} -\frac{1}{2}b & b \end{pmatrix}', \text{with } a, b \in \mathbb{R},$$

but we should consider the normalized version:

$$\mathbf{a}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.894 \\ 0.447 \end{pmatrix} \quad \text{and} \quad \mathbf{a}_2 = \frac{1}{\sqrt{5/4}} \begin{pmatrix} -0.5 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.447 \\ 0.894 \end{pmatrix}.$$

Of course

$$\mathbf{a}_1'\mathbf{a}_2 = 0 \quad \text{and} \quad \mathbf{a}_1'\mathbf{a}_1 = \mathbf{a}_2'\mathbf{a}_2 = 1.$$

We consider $\mathbf{z} = \mathbf{A}\mathbf{y}$ with

$$\mathbf{A} = \mathbf{C}' = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{pmatrix} = \begin{pmatrix} 0.894 & 0.447 \\ -0.447 & 0.894 \end{pmatrix}$$

hence the principal components are

$$\begin{aligned} \mathbf{z}_1 &= \mathbf{a}_1'(\mathbf{y} - \bar{\mathbf{y}}) = 0.894\,(y_1 - 1) + 0.447\,(y_2 - 2) \\ \mathbf{z}_2 &= \mathbf{a}_2'(\mathbf{y} - \bar{\mathbf{y}}) = -0.447\,(y_1 - 1) + 0.894\,(y_2 - 2)\ . \end{aligned}$$

The major axis is the line passing through $\bar{\mathbf{y}}$ in the direction of the vector

$$\mathbf{a}_1' = \begin{pmatrix} 0.894 & 0.447 \end{pmatrix}.$$

The length of the first semi axis is $\sqrt{\lambda_1} = \sqrt{6} = 2.4495$ and the length of the second semi axis is $\sqrt{\lambda_2} = \sqrt{1} = 1$.

Geometric interpretation: the direction of the first principal line (which correspond to the first principal component) is the direction along which the observations are maximally separated or spread out. In our case, the equation of the principal component is an equation of a "principal" line and is given by

$$z_2 = 0$$

or

$$-0.447\,(y_1 - 1) + 0.894\,(y_2 - 2) = 0 \Leftrightarrow \frac{y_1 - 1}{0.894} = \frac{y_2 - 2}{0.447}$$

which direction is, of course, exactly $a_1' = \begin{pmatrix} 0.894 & 0.447 \end{pmatrix}$.

Algebraic interpretation: the first PC is a linear combination of the variables with maximal variance. The second principal component is the linear combination with maximal variance in a direction orthogonal to the first principal component. In our case, we calculate

$$\begin{aligned} \operatorname{var}(z_1) &= \operatorname{var}(0.894y_1 + 0.447y_2) = (0.894)^2 \operatorname{var}(y_1) + (0.447)^2 \operatorname{var}(y_2) \\ &\quad +2 \cdot 0.894 \cdot 0.447 \operatorname{cov}(y_1, y_2) = \ldots = \lambda_1 \end{aligned}$$

and

$$\operatorname{cov}(z_1, z_2) = \operatorname{cov}(0.894y_1 + 0.447y_2, -0.447y_1 + 0.894y_2) = \ldots = 0$$

and $\operatorname{var}(z_2) = \lambda_2$.

If $\mathbf{z} = \mathbf{A}\mathbf{y}$, then $\mathbf{S}_z = \mathbf{A}\mathbf{S}\mathbf{A}'$ and, after the computations,

$$\mathbf{S}_z = \mathbf{A}\mathbf{S}\mathbf{A}' = \ldots = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

We consider the proportion of variance

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{6}{7} = 0.857\,,$$

hence 86% of the total variance $\sum_{j=1}^{2} s_{z_j}^2$ is due to the first components, then the first principal component can "replace" the original 2 variables without loss of information. We also remark that the largest coefficient in z_1 (which is 0.894) corresponds to the largest variances on the diagonal of $\mathbf{S}$, which is $s_{y_1}^2 = 5$. Hence the variable y_1 has a notable influence on the first principal component.

(ii) If $\mathbf{S} = \begin{pmatrix} 2 & 2r & 2r \\ 2r & 2 & 2r \\ 2r & 2r & 2 \end{pmatrix}$, then the correlation matrix becomes (it is very easy to apply formulas for sample correlation) $\mathbf{R} = \begin{pmatrix} 1 & r & r \\ r & 1 & r \\ r & r & 1 \end{pmatrix}$.

The eigenvalues and eigenvectors of $\mathbf{R}$ are

$$\begin{aligned}
&\lambda_1 = 1+2r, \quad \mathbf{a}_1 = \begin{pmatrix} a & a & a \end{pmatrix}', \\
&\lambda_2 = 1-r, \quad \mathbf{a}_2 = \begin{pmatrix} -b & b & 0 \end{pmatrix}' \quad \text{and} \\
&\lambda_3 = 1-r, \quad \mathbf{a}_3 = \begin{pmatrix} -c & 0 & c \end{pmatrix}', \quad \text{with } a,b,c \in \mathbb{R},
\end{aligned}$$

but should be normalized, hence

$$\mathbf{a}_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.577 \\ 0.577 \\ 0.577 \end{pmatrix} \quad \text{and} \quad \mathbf{a}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -0.707 \\ 0.707 \\ 0 \end{pmatrix}$$

$$\text{and} \quad \mathbf{a}_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.707 \\ 0 \\ 0.707 \end{pmatrix}.$$

We have

$$\mathbf{a}_i'\mathbf{a}_j = 0 \quad \text{and} \quad \mathbf{a}_i'\mathbf{a}_i = 1, \quad i \neq j.$$

We consider $\mathbf{z} = \mathbf{A}\mathbf{y}$ with

$$\mathbf{A} = \mathbf{C}' = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \mathbf{a}_3' \end{pmatrix} = \begin{pmatrix} 0.577 & 0.577 & 0.577 \\ -0.707 & 0.707 & 0 \\ -0.707 & 0 & 0.707 \end{pmatrix}$$

hence

$$\begin{aligned}
\mathbf{z}_1 &= \mathbf{a}_1'\mathbf{y} = 0.577 \cdot \frac{y_1 - \bar{y}_1}{\sqrt{s_1^2}} + 0.577 \cdot \frac{y_2 - \bar{y}_2}{\sqrt{s_2^2}} + 0.577 \cdot \frac{y_3 - \bar{y}_3}{\sqrt{s_3^2}}, \\
\mathbf{z}_2 &= \mathbf{a}_2'\mathbf{y} = -0.707 \cdot \frac{y_1 - \bar{y}_1}{\sqrt{s_1^2}} + 0.707 \cdot \frac{y_2 - \bar{y}_2}{\sqrt{s_2^2}}, \\
\mathbf{z}_3 &= \mathbf{a}_3'\mathbf{y} = -0.707 \cdot \frac{y_1 - \bar{y}_1}{\sqrt{s_1^2}} + 0.707 \cdot \frac{y_3 - \bar{y}_3}{\sqrt{s_3^2}}.
\end{aligned}$$

We remark that the principal components do not depend on r but the proportion of variance depends on r. For instance, if $r = 0.01$, the proportion of variance is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{1+2r}{(1+2r)+(1-r)+(1-r)} = \frac{1+2r}{3} = \frac{1.02}{3} = 0.34$$

and if $r = 0.99$, the proportion of variance is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1 + 2 \cdot 0.99}{3} = 0.9933,$$

hence to the same principal components corresponds different values for total variance and therefore there is no connection between the principal components and the values for total variance.

Another possible remarks:

- The principal components constitute a rotation of axes.
- Another important geometric property of the line formed by the first principal component (the direction of this line is given by the first eigenvector $\mathbf{a}_1'$) is that it minimizes the total sum of squared perpendicular distances from the points to the line. Hence, the line formed by the major axis can be considered to be a regression line; the perpendicular distance from the points to this line is minimized (rather that the distance to the semiminor axis).

- We recall the geometrical interpretation of the least squares technique in the linear regression models: If we consider case $q = 1$, we obtain that the ordinary regression line of y on x minimizes the sum of squares of *vertical distances* from the points y_i to the line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The first principal component line represent a perpendicular regression since this line is such that minimizes the total sum of squared *perpendicular distances* from the points to this line.
- The coefficients of the principal components obtained from $\mathbf{R}$ differ from those obtained from $\mathbf{S}$.
- The proportion of variance obtained starting from $\mathbf{R}$ differ from the proportion of variance obtained from $\mathbf{S}$.