Solution exam 2015

Problem 1

Enclosure A with SAS program and SAS output belongs to this problem. The data are taken from "Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado (2015): Modeling and Analysis of Compositional Data, xvii + 247 pp., *John Wiley & Sons, Ltd.*". The data give percentages of protein consumption in 25 European countries in 1980-1989 provided by each of nine food categories:

1. Red meats (pork, beef, veal),
2. White meats (poultry(chicken, turkey)),
3. Eggs
4. Milk products
5. Fish (sea and freshwater)
6. Cereals
7. Starch sources (potatoes, rice),
8. Nuts
9. Vegetables (including fruit and orchard products).

The abbreviations used for the 25 countries considered are given below. Furthermore we give a 'geopolitical' code where the last letter (E or W) indicates whether the country belonged to the eastern or western block, and the prefix (Nor, Cen, or Sou) indicates the geographical location within Europe.

Country	Abbrev.	GeoPol
Albania	ALB	SouE
Austria	AUS	CenW
Belgium	BEL	CenW
Bulgaria	BUL	SouE
Czech Republic	CZE	CenE
Denmark	DEN	NorW
German Fed. Rep.	BRD	CenW
Finland	FIN	NorW
France	FRA	CenW
Greece	GRE	NorW
Hungary	HUN	CenE
Ireland	IRE	NorW
Italy	ITA	SouW

Country	Abbrev.	GeoPol
Netherlands	HOL	CenW
Norway	NOR	NorW
Poland	POL	CenE
Portugal	POR	SouW
Romania	ROM	SouE
Spain	SPA	SouW
Sweden	SWE	NorW
Switzerland	SCH	CenW
United Kingdom	UK	NorW
USSR	USSR	CenE
German Dem. Rep.	DDR	CenE
Yugoslavia	YUG	SouE

The problem continues on the next page

Question 1.1 - 5

How many principal components should be included in the analysis in order to describe at least 90% of the total variation?

Eigenvalues of the Correlation Matrix: Total = 9 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.00643757	2.37143813	0.4452	0.4452
2	1.63499945	0.50707994	0.1817	0.6268
3	1.12791950	0.17325554	0.1253	0.7522
4	0.95466396	0.49082557	0.1061	0.8582
5	0.46383840	0.13870742	0.0515	0.9098
6	0.32513097	0.05352464	0.0361	0.9459
7	0.27160633	0.15531443	0.0302	0.9761
8	0.11629190	0.01718000	0.0129	0.9890
9	0.09911190		0.0110	1.0000

Examining the output generated by 'proc factor', we particularly interested in the "Eigenvalues" section, which provides information about the variance explained by each factor (component). Looking at the cumulative we can find the sum of the proportion of the variance explained by each factor. We need 90% of the total variation, so 5 components are needed to explain this variation.

The answer is 5.

Question 1.2 - 1

The distribution of the usual test statistic for testing the hypothesis that the smallest 5 eigenvalues (in the correlation matrix) are equal against all alternatives is (under the hypothesis).

To find the distribution of the usual test statistic for testing the hypothesis that the smallest 5 eigenvalues are equal against all the alternatives we can use the Theorem 6.8, page 375 of the Book.

$$\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n|z_2 > \chi^2(\frac{1}{2}(k-m+2)(k-m-1))_{1-\alpha}\}.$$

According to this, in a dataset with k=9 variables.

m=4 the first 4 eigenvalues.

Since we want to test under the null hypothesis that the smallest 5 eigenvalues are equal to zero $m = k - 5 = 4$. Then

$$\chi^2\left(\frac{1}{2}(9-4+2)(9-4-1)\right) = \chi^2(14)$$

The answer is 1.

Question 1.3 - 4

We now consider a factor analysis with three factors of the data considered above. Consider the following

statements on interpretation of an arbitrary factor. The factor basically represents

I. the mean level of all meat values.

II. a large correlation with fish, potatoes/rice and vegetables on one side, and a numerically large, negative correlation with cereals on the other side.

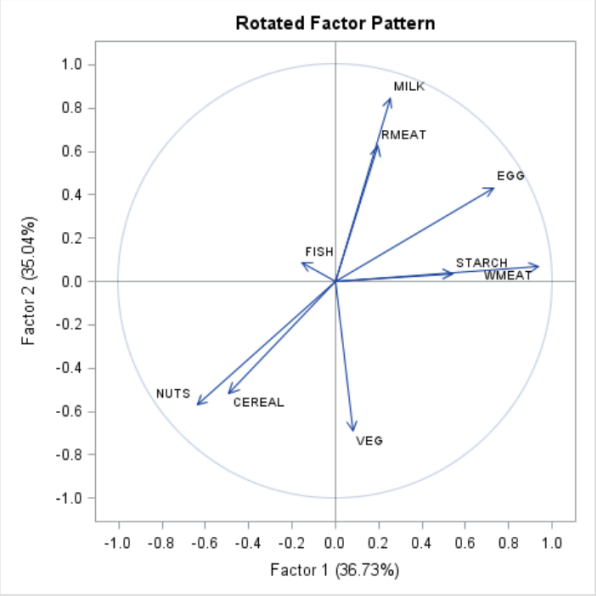
III. the mean level of animal protein sources

IV. a large correlation with poultry, eggs and potatoes/rice on one side, and a numerically large, negative correlation with nuts and cereals on the other side.

V. a large correlation with beef/pork/veal, eggs and milk products on one side, and a numerically large, negative correlation with cereals, nuts and vegetables on the other side.

For (VARIMAX rotated factor 1, VARIMAX rotated factor 2) the following characterization is adequate.

We are looking for the characterization of the set (VARIMAX rotated factor1, VARIMAX rotated factor2), so we can inspect the Rotated Factor Pattern Plots.



Factor 1 has a high correlation with egg, starch sources, and white meat, and a negative correlation with cereals and nuts which can be represented by factor IV.

Factor 2 has high correlation with milk, red meat and eggs, and negative correlation with cereal, nuts and vegetables which perfectly can be represented by factor V.

The answer is 4.

Question 1.4 - 1

The first 2 factor score values for a given country are denoted ($f1$, $f2$). We now consider the northernmost countries in the western block (NorW) and Central European countries in the western block (CenW). The distribution of the factor scores for the two areas satisfy.

From the Factor Score Plot it can be seen that f1 has positive values for Central European countries in the western block while f2 takes only intermediate values for these countries (see ◆). In contrast, f1 has intermediate values for northernmost countries in the western block, while f2 has positive large values (see▼).

The answer is 1.

Question 1.5 - 4

The variable that is explained best by the rotated factor model is:

From Factor analysis we are looking for the finale communality estimates.

Final Communality Estimates: Total = 6.769357								
RMEAT	WMEAT	EGG	MILK	FISH	CEREAL	STARCH	NUTS	VEG
0.471934	0.917164	0.768599	0.795097	0.874187	0.867220	0.624117	0.744777	0.706260

The highest value of communality 0.917164 for whit meat (close to 1) suggests that this variable is well-represented by the rotated factors. The factor analysis has successfully explained a significant portion of the variance in that variable.

The answer is 4.

Question 1.6 - 2

The reduction in the variance explained by the first factor by going from the unrotated model to the rotated model is:

We look at the Variance explained by each factor for both the initial and the rotated model and we subtract :

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
4.0064376	1.6349994	1.1279195

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
2.4863837	2.3723048	1.9106681

$4.0064376 - 2.4863837 = 1.5200539 \sim 1.5$

The answer is 2.

Problem 2

We consider a random variable

$$\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ X_1 \\ X_2 \end{bmatrix}$$

with expectation vector and dispersion matrix equal to

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 1 & 1/4 & 1/8 \\ 1 & 4 & 1 & 1 \\ 1/4 & 1 & 1 & 1 \\ 1/8 & 1 & 1 & 4 \end{bmatrix}$$

In the sequel you may find the following expressions useful

$$\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}^{-1} = \frac{1}{3}\begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\frac{1}{3}\begin{bmatrix} 1/4 & 1/8 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 7/24 & -1/24 \\ 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 7/24 & -1/24 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 1/4 & 1 \\ 1/8 & 1 \end{bmatrix} = \begin{bmatrix} 13/192 & 1/4 \\ 1/4 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} - \begin{bmatrix} 13/192 & 1/4 \\ 1/4 & 1 \end{bmatrix} = \begin{bmatrix} 179/192 & 3/4 \\ 3/4 & 3 \end{bmatrix}$$

Question 2.1 - 2

The squared correlation between Y_1 and Y_2 is:

The correlation between the two variables is given by:

$p_{(Y_1,Y_2)} = Corr(Y_1, Y_2) = \frac{Cov(Y_1,Y_2)}{\sqrt{V(Y_1)V(Y_2)}}$,

where $Cov(Y_1, Y_2) = 1$, $V(Y_1) = 1$ *and* $V(Y_2) = 4$ are given from the dispersion matrix.

$p_{(Y_1,Y_2)}{}^2 = \frac{Cov(Y_1,Y_2)^2}{V(Y_1)V(Y_2)} = \frac{1^2}{1 \cdot 4} = \frac{1}{4}$

The answer is 2.

Question 2.2 - 2

The squared correlation between Y_2 and X_1 is:

The correlation between the two variables is given by:

$p_{(Y_2,X_1)} = Corr(Y_2, X_1) = \frac{Cov(Y_2,X_1)}{\sqrt{V(Y_2)V(X_1)}}$,

where $Cov(Y_2, X_1) = 1$, $V(Y_2) = 4$ and $V(X_1) = 1$ are given from the dispersion matrix.

$p_{(Y_2,X_1)}{}^2 = \frac{Cov(Y_2,X_1)^2}{V(Y_2)V(X_1)} = \frac{1^2}{4 \cdot 1} = \frac{1}{4}$

The answer is 2.

Question 2.3 - 4

The squared partial correlation $\boldsymbol{p_{(Y_1,Y_2|X_1,X_2)}{}^2}$ between Y1 and Y2 given $\boldsymbol{X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}}$ is:

$$p_{(Y_1,Y_2|X_1,X_2)} = \frac{p_{(Y_1,Y_2|X_1)} - p_{(Y_1,X_2|X_1)} \cdot p_{(Y_2,X_2|X_1)}}{\sqrt{(1 - p_{(Y_1,X_2|X_1)}{}^2)(1 - p_{(Y_2,X_2|X_1)}{}^2)}}$$

Where

$$p_{(Y_1,Y_2|X_1)} = \frac{p_{(Y_1,Y_2)} - p_{(Y_1,X_1)} \cdot p_{(Y_2,X_1)}}{\sqrt{(1 - p_{(Y_1,X_1)}{}^2)(1 - p_{(Y_2,X_1)}{}^2)}}$$

$$p_{(Y_1,X_2|X_1)} = \frac{p_{(Y_1,X_2)} - p_{(Y_1,X_1)} \cdot p_{(X_2,X_1)}}{\sqrt{(1 - p_{(Y_1,X_1)}{}^2)(1 - p_{(X_2,X_1)}{}^2)}}$$

And

$$p_{(Y_2,X_2|X_1)} = \frac{p_{(Y_2,X_2)} - p_{(Y_2,X_1)} \cdot p_{(X_2,X_1)}}{\sqrt{(1 - p_{(Y_2,X_1)}{}^2)(1 - p_{(X_2,X_1)}{}^2)}}$$

From the dispersion matrix:

$$p_{(Y_1,Y_2)} = \frac{Cov(Y_1, Y_2)}{\sqrt{V(Y_1)V(Y_2)}} = \frac{1}{\sqrt{1 \cdot 4}} = \frac{1}{\sqrt{4}}$$

$$p_{(Y_1,X_2)} = \frac{Cov(Y_1, X_2)}{\sqrt{V(Y_1)V(X_2)}} = \frac{\frac{1}{8}}{\sqrt{1 \cdot 4}} = \frac{1}{8 \cdot \sqrt{4}}$$

$$p_{(Y_1,X_1)} = \frac{Cov(Y_1, X_1)}{\sqrt{V(Y_1)V(X_1)}} = \frac{1/4}{\sqrt{1 \cdot 1}} = \frac{1}{4}$$

$$p_{(Y_2,X_2)} = \frac{Cov(Y_2, X_2)}{\sqrt{V(Y_2)V(X_2)}} = \frac{1}{\sqrt{4 \cdot 4}} = \frac{1}{4}$$

$$p_{(Y_2,X_1)} = \frac{Cov(Y_2, X_1)}{\sqrt{V(Y_2)V(X_1)}} = \frac{1}{\sqrt{4 \cdot 1}} = \frac{1}{\sqrt{4}}$$

$$p_{(X_2,X_1)} = \frac{Cov(X_2,X_1)}{\sqrt{V(X_2)V(X_1)}} = \frac{1}{\sqrt{4\cdot 1}} = \frac{1}{\sqrt{4}}$$

By substitution :

$$p_{(Y_1,Y_2|X_1)} = \frac{\frac{1}{\sqrt{4}} - \frac{1}{4}\cdot\frac{1}{\sqrt{4}}}{\sqrt{(1-\frac{1}{4}^2)(1-\frac{1}{\sqrt{4}}^2)}} = 0.447$$

$$p_{(Y_1,X_2|X_1)} = \frac{\frac{1}{8}\cdot\frac{1}{\sqrt{4}} - \frac{1}{4}\cdot\frac{1}{\sqrt{4}}}{\sqrt{\left(1-\frac{1}{4}^2\right)\left(1-\frac{1}{\sqrt{4}}^2\right)}} = -0.069$$

$$p_{(Y_2,X_2|X_1)} = \frac{\frac{1}{4} - \frac{1}{\sqrt{4}}\cdot\frac{1}{\sqrt{4}}}{\sqrt{(1-\frac{1}{\sqrt{4}}^2)(1-\frac{1}{\sqrt{4}}^2)}} = 0$$

Finally :

$$p_{(Y_1,Y_2|X_1,X_2)}{}^2 = \frac{(0.447 + 0.069\cdot 0)^2}{(1-(-0.069)^2)(1-0^2)} = 0.2 \sim \frac{36}{179}$$

The answer is 4.

Question 2.4 - 1

The conditional mean $E(Y2|\ X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix})$ is:

According Theorem 1.27:

Theorem 1.27

If X_2 is regularly distributed, i.e. if Σ_{22} has full rank, then the distribution of X_1 conditioned on $X_2 = x_2$ is again a normal distribution, and the following holds

$$\begin{aligned} \mathrm{E}(X_1|X_2 = x_2) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \mathrm{D}(X_1|X_2 = x_2) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

If Σ_{22} does not have full rank then the conditional distribution is still normal and Σ_{22}^{-1} in the above equations should be substituted by a generalised inverse Σ_{22}^{-}.

Where $\mu_1 = 0, \Sigma_{12} = [1 \quad 1],\ \Sigma_{22} = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}, \mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$E\left(Y_2 \middle| X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right) = 0 + [1 \quad 1]\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}^{-1}\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = X_1$$

The answer is 1.

Question 2.5 - 2

The conditional mean $E(Y1|X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix})$ is:

Using the above equation Where $\mu_1 = 0, \Sigma_{12} = [\frac{1}{4} \quad \frac{1}{8}],\ \Sigma_{22} = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}, \mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$E\left(Y_2 \middle| X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right) = 0 + \begin{bmatrix} \frac{1}{4} & \frac{1}{8} \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}^{-1}\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \frac{7}{24}X_1 - \frac{1}{24}X_2$$

The answer is 2.

Question 2.6 - 3

The squared partial correlation $p_{(Y_2|X_1,X_2)}{}^2$ between Y2 and $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ is:

From Theorem 1.42:

Theorem 1.42

We consider the situation above. Let σ_i be the i'th column in $\mathbf{\Sigma}_{xy}$, i.e. σ_i^T is the i'th row in $\mathbf{\Sigma}_{yx}$. Further, let σ_{ii} denote the i'th diagonal element, i.e. the variance of Y_i

Then

$$\rho_{y_i|x} = \frac{\sqrt{\sigma_i^T \mathbf{\Sigma}_{xx}^{-1} \sigma_i}}{\sqrt{\sigma_{ii}}}.$$

If we let

$$\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_{ii} & \sigma_i^T \\ \sigma_i & \mathbf{\Sigma}_{xx} \end{bmatrix},$$

then

$$1 - \rho_{y_i|x}^2 = \frac{\det \mathbf{\Sigma}_i}{\sigma_{ii} \det \mathbf{\Sigma}_{xx}} = \frac{\mathrm{V}(Y_i|\boldsymbol{X})}{\mathrm{V}(Y_i)},$$

Where

$$\Sigma_i = \begin{bmatrix} 4 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 4 \end{bmatrix}, \Sigma_{XX} = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}, \sigma_i = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \sigma_{ii} = 4$$

$$p_{(Y_2|X_1,X_2)}{}^2 = 1 - \frac{\det(\Sigma_i)}{\sigma_{ii} \det(\Sigma_{XX})} = 1 - \frac{3}{4} = \frac{1}{4}$$

$\det(\Sigma_i) = 9, \det(\Sigma_{XX}) = 3,$

The answer is 3.

Question 2.7 - 5

The squared partial correlation $p_{(Y1|X_1,X_2)}{}^2$ between Y1 and $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ is:

Where

$$\Sigma_i = \begin{bmatrix} 1 & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{4} & 1 & 1 \\ \frac{1}{8} & 1 & 4 \end{bmatrix}, \Sigma_{XX} = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}, \sigma_i = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{8} \end{bmatrix}, \sigma_{ii} = 1$$

$$p_{(Y_1|X_1,X_2)}{}^2 = 1 - \frac{\det(\Sigma_i)}{\sigma_{ii} \det(\Sigma_{XX})} = 1 - \frac{179}{192} = \frac{13}{192}$$

$\det(\Sigma_i) = \frac{179}{64}, \det(\Sigma_{XX}) = 3,$

The answer is 5.

Problem 3

Enclosure B with SAS program and SAS output belongs to this problem. The data are taken from a study reported in "Camilla Himmelstrup Trinderup (2014): Spectral Imaging of Meat Quality - Color and Texture, PHD-201 4-358, DTU Compute, Lyngby". The problem investigated is the development over time of the diameters of salamis during fermentation. The data considered here comprise diameter measurements (in pixels) on salamis using two different starter cultures for the fermentation (type=1 or type=2) The measurements were taken 2, 3, 9, 14, 21, and 42 days after production.

We are now interested in to which extent the time development depends on the choice of starter culture. We consider the following model and hypotheses.

$$M: \quad E(Y_{tj}) = \alpha_j + \beta_j t + \gamma_j t^2 \quad j = 1,2$$

$$H_1: \quad E(Y_{tj}) = \alpha_j + \beta t + \gamma t^2 \quad j = 1,2$$

$$H_2: \quad E(Y_{tj}) = \alpha + \beta t + \gamma t^2 \quad j = 1,2$$

In all three cases we assume that the errors are independent and normally distributed with the same variance.

Question 3.1 - 2

What fraction of the total variation is explained by model H_2?

To find the information we want for H_2, we need to look at the results from the model:

width=day day*day with intercept

Model: width = day day*day with intercept
The GLM Procedure

Class Level Information		
Class	Levels	Values
type	2	1 2

Number of Observations Read	12
Number of Observations Used	12

Dependent Variable: width

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	41551.96036	20775.98018	307.37	<.0001
Error	9	608.33593	67.59288		
Corrected Total	11	42160.29629			

R-Square	Coeff Var	Root MSE	width Mean
0.985571	0.653594	8.221489	1257.889

R-Square gives the total variance explained by this model.

The answer is 2.

Question 3.2 - 1

If we assume the model $\mathcal{M}$ and want to test the hypothesis $\mathcal{H}1$, the usual test statistic becomes:

The usual test statistic is given by :

Test statistic for $H_0: E(\boldsymbol{Y}) \in H$ against $H_1: E(\boldsymbol{Y}) \in M \backslash H$:

$$\frac{\|p_M(Y) - p_H(Y)\|^2/(k-r)}{\|Y - p_M(Y)\|^2/(n-k)} = \frac{(SS_{res}(Hyp) - SS_{res}(Mod))/(DF_{res}(Hyp) - DF_{res}(Mod))}{SS_{res}(Mod)/DF_{res}(Mod)}$$

And the values of each terms can be found in the tables below:

For M model:

Dependent Variable: width

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	19029406.22	3171567.70	113568	<.0001
Error	6	167.56	27.93		
Uncorrected Total	12	19029573.78			

And for H1:

Dependent Variable: width

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	19029366.04	4757341.51	183201	<.0001
Error	8	207.74	25.97		
Uncorrected Total	12	19029573.78			

$$SSres(Mod) = 167.56, \quad SSres(Hyp) = 207.74$$

$$and\ DFres(Mod) = 6\ and\ DFres(Hyp) = 8$$

$$\frac{207.74 - 167.56/8 - 6}{167.56/6}$$

The answer is 1.

Question 3.3 - 4

The distribution of the test statistic under the hypothesis is?

Theorem 2.21

Let the situation be as above. Then the likelihood ratio test at level α of testing

$$H_0 : \boldsymbol{\mu} \in H \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \in M \backslash H,$$

is equivalent to the test given by the critical region

$$C_\alpha = \{(y_1, \ldots, y_n) | \frac{\|p_M(\boldsymbol{y}) - p_H(\boldsymbol{y})\|^2/(k-r)}{\|\boldsymbol{y} - p_M(\boldsymbol{y})\|^2/(n-k)} > \mathrm{F}(k-r, n-k)_{1-\alpha}\}.$$

Then we only need to know the degrees of freedom of model from hypothesis and the degrees of freedom of observations from model : $k - r = 2, n - k = 6$

Variation	SS	Degrees of freedom = dimension
Of model from hypothesis	$\|p_M(\boldsymbol{Y}) - p_H(\boldsymbol{Y})\|^2$	$k - r$
Of observations from model	$\|\boldsymbol{Y} - p_M(\boldsymbol{Y})\|^2$	$n - k$
Of observations from hypothesis	$\|\boldsymbol{Y} - p_H(\boldsymbol{Y})\|^2$	$n - r$

The answer is 4.

Question 3.4 - 4

The usual test statistic for testing H_2 assuming that H_1 is true becomes:

And the values of each term for the above equation can be found in the tables below:

For H2 model:

Dependent Variable: width

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	41551.96036	20775.98018	307.37	<.0001
Error	9	608.33593	67.59288		
Corrected Total	11	42160.29629			

And for H1:

Dependent Variable: width

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	19029366.04	4757341.51	183201	<.0001
Error	8	207.74	25.97		
Uncorrected Total	12	19029573.78			

$$SSres(Mod) = 207.74, \qquad SSres(Hyp) = 608.33593$$

$$and\ DFres(Mod) = 8\ and\ DFres(Hyp) = 9$$

$$\frac{608.33593 - 207.74/8 - 9}{207.74/8}$$

The answer is 4.

Question 3.5 - 5

We now consider model $\mathcal{H}2$. Which observation will cause the largest overall change in the estimated parameter vector if it is omitted from the estimation ?

High DFFITS value indicates that omitting that observation would significantly affect the predicted values. From the table observation 12 has highest DFFIT value.

Model: width = day day*day with intercept

Obs	day	type	width	PREDICTED	StdErrMPV	RESIDUAL	STUDENTR	RSTUDENT	CooksD	DFFITS	HATDiagH
1	2	1	1335.00	1336.84	4.10299	-1.8356	-0.25765	-0.24382	0.00734	-0.14042	0.24906
2	2	2	1341.67	1336.84	4.10299	4.8310	0.67809	0.65629	0.05083	0.37796	0.24906
3	3	1	1322.00	1326.25	3.69001	-4.2547	-0.57912	-0.55647	0.02820	-0.27949	0.20144
4	3	2	1326.00	1326.25	3.69001	-0.2547	-0.03467	-0.03269	0.00010	-0.01642	0.20144
5	9	1	1265.33	1270.10	2.76587	-4.7640	-0.61532	-0.59273	0.01611	-0.21175	0.11318
6	9	2	1277.33	1270.10	2.76587	7.2360	0.93461	0.92731	0.03716	0.33127	0.11318
7	14	1	1218.67	1232.90	3.42423	-14.2289	-1.90367	-2.32224	0.25353	-1.06387	0.17347
8	14	2	1246.67	1232.90	3.42423	13.7711	1.84242	2.20103	0.23748	1.00835	0.17347
9	21	1	1192.00	1195.47	4.25002	-3.4691	-0.49292	-0.47114	0.02954	-0.28451	0.26723
10	21	2	1198.33	1195.47	4.25002	2.8642	0.40698	0.38729	0.02013	0.23388	0.26723
11	42	1	1179.67	1185.78	5.78796	-6.1143	-1.04717	-1.05354	0.35917	-1.04436	0.49562
12	42	2	1192.00	1185.78	5.78796	6.2191	1.06511	1.07418	0.37159	1.06481	0.49562

The answer is 5.

Question 3.6 - 2

For model $\mathcal{H}2$ the usual 95% confidence interval for the mean value of an observation at day 42 using starter culture type 2 is:

According to the Theorem 2.15:

> **Theorem 2.15**
>
> Let the situation be as above. Then the $(1-\alpha)$-***confidence interval*** for the expected value of a new observation Y will be
>
> $$[u - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c},\quad u + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c}].$$

For a 95% confidence interval $a = 0.05 => 1 - a = 0.975$, and u is the predicted value of the observations of type 2 on day 42 $u = 1185.78$ from the table below.

Model: width = day day*day with intercept

Obs	day	type	width	PREDICTED	StdErrMPV	RESIDUAL	STUDENTR	RSTUDENT	CooksD	DFFITS	HATDiagH
1	2	1	1335.00	1336.84	4.10299	-1.8356	-0.25765	-0.24382	0.00734	-0.14042	0.24906
2	2	2	1341.67	1336.84	4.10299	4.8310	0.67809	0.65629	0.05083	0.37796	0.24906
3	3	1	1322.00	1326.25	3.69001	-4.2547	-0.57912	-0.55647	0.02820	-0.27949	0.20144
4	3	2	1326.00	1326.25	3.69001	-0.2547	-0.03467	-0.03269	0.00010	-0.01642	0.20144
5	9	1	1265.33	1270.10	2.76587	-4.7640	-0.61532	-0.59273	0.01611	-0.21175	0.11318
6	9	2	1277.33	1270.10	2.76587	7.2360	0.93461	0.92731	0.03716	0.33127	0.11318
7	14	1	1218.67	1232.90	3.42423	-14.2289	-1.90367	-2.32224	0.25353	-1.06387	0.17347
8	14	2	1246.67	1232.90	3.42423	13.7711	1.84242	2.20103	0.23748	1.00835	0.17347
9	21	1	1192.00	1195.47	4.25002	-3.4691	-0.49292	-0.47114	0.02954	-0.28451	0.26723
10	21	2	1198.33	1195.47	4.25002	2.8642	0.40698	0.38729	0.02013	0.23388	0.26723
11	42	1	1179.67	1185.78	5.78796	-6.1143	-1.04717	-1.05354	0.35917	-1.04436	0.49562
12	42	2	1192.00	1185.78	5.78796	6.2191	1.06511	1.07418	0.37159	1.06481	0.49562

It is show in question 3.3 that n-k for H_2 is 9 and

R-Square	Coeff Var	Root MSE	width Mean
0.985571	0.653594	8.221489	1257.889

s is Root MSE from the above table and c isgiven by the HATDiagH value of the above table. Do the answer is

$$1185.835 - t(9)_{0.975} \times 8.221489 \cdot \sqrt{0.49562}, 1185.835 + t(9)_{0.975} \times 8.221489 \cdot \sqrt{0.49562}$$

The answer is 2.

Question 3.7 - 4

For model H2 the usual 95% prediction interval for a new observation at day 42 using starter culture type 2 is:

According to Theorem 2.17

Theorem 2.17

Let us assume that a new observation taken at $(z_1, \ldots, z_k)$ has a variance $c_1\sigma^2$. Furthermore, it is independent of the earlier observations. In that case a $(1-\alpha)$-***prediction interval*** for the new observation equals the interval

$$[u - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c+c_1}, u + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c+c_1}].$$

We are now looking at the prediction interval of a new observation.

The answer is :

RootMSE = 8.221489 so MSE =67.5928

So the above equation becomes:

$$1185.835 - t(9)_{0.975} \times \sqrt{67.59288 \cdot (0.49562+1)}, 1185.835 + t(9)_{0.975} \times \sqrt{67.59288 \cdot (0.49562+1)}$$

The answer is 4.

Problem 4

Enclosure C with SAS program and SAS output belongs to this problem. The data are taken from a yet unpublished study at DTU Compute on the possible relation between some specific cell measurements using image analysis and possible heterogeneity in a donor population. The data used here comprise measurements on 20 cells from each of two donors (number 3 and number 4). The variables are

1. *areacell*, the logarithm of the cell area
2. *areanucl*, the logarithm of the nucleus fraction of the cell area
3. *avecytop*, the average cytoplasma intensity for the cell

We are now interested in how suitable the cell measurements are in distinguishing between the donor types represented by the two donors actually investigated.

Question 4.1 - 1

What is the absolute value $|t|$ of the usual t-test statistic for assessing whether the mean values for areacell are the same for the two donors:

For t^2 we can directly read the F-value for the *areacell.* So for the absolute value $|t|$:

$$\sqrt{F} = \sqrt{8.58}$$

The answer is 1.

Question 4.2 - 2

What is the F-test statistic for assessing whether the mean values for avecytop are the same for the two donors:

We could read directly the F-value for the avecytop, but this information is not given, so we will use the following matrices instead:

Dependent Variable: areacell

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.72621200	0.72621200	8.58	0.0057
Error	38	3.21492881	0.08460339		
Corrected Total	39	3.94114081			

E = Error SSCP Matrix

	areacell	areanucl	avecytop
areacell	3.2149288082	-1.695602673	53.002800162
areanucl	-1.695602673	3.9859687177	-65.35021642
avecytop	53.002800162	-65.35021642	1498.2752357

H = Type III SSCP Matrix for donor

	areacell	areanucl	avecytop
areacell	0.726212007	-0.869984451	19.161900055
areanucl	-0.869984451	1.0422203745	-22.95549377
avecytop	19.161900055	-22.95549377	505.60774572

We can see that the diagonal element for areacell is exactly what we had in the first table, i.e. the H table we find the donor SS and in the E table the error SS and we can just replace the values for areacell with those in the diagonal for avecytop and get :

$$F = \frac{505.6/1}{1498.3/38} = 12.82$$

The degrees of freedom is the same as for areacell since we are performing the same test, just using a different variable.

Test statistic for $H_0: E(\boldsymbol{Y}) \in H$ against $H_1: E(\boldsymbol{Y}) \in M \backslash H$:

$$\frac{\|p_M(Y) - p_H(Y)\|^2/(k-r)}{\|Y - p_M(Y)\|^2/(n-k)} = \frac{(SS_{res}(Hyp) - SS_{res}(Mod))/(DF_{res}(Hyp) - DF_{res}(Mod))}{SS_{res}(Mod)/DF_{res}(Mod)}$$

We can find the DF from page 129.

k is the rank of the full model (H1)

r is the rank of the simpler model (H0)

n is the number of observations.

We have in the full model two levels of donors, i.e. k=2

In the simple model we have only a common mean, since we are testing if the two donors are the same, i.e. r=1

The number of observations we can find them in the enclosure, n=40.

That gives us k-r=2-1=1DF in the numerator, n-k=40-2=38 DF in the denominator

The answer is 2.

Question 4.3 - 1

If we test whether the mean values for [areacell, areanucl, avecytop] are the same for the two donors, then the distribution of the MANOVA test statistic(assuming that the hypothesis is true) becomes:

Theorem 4.25 is used

Theorem 4.25

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

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

P=3 since we are testing the mean values of the 3 variables. And according to the above explanation,

k-1=1 and

n-k=38

The answer is 1.

Question 4.4 - 2

Proc Discrim computes the squared Mahalanobis distance between the two donors. The F-test statistic corresponding to this squared Mahalanobis distance between the two donors using all three variables is:

F-test is given from the table below:

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall donor Effect H = Type III SSCP Matrix for donor E = Error SSCP Matrix S=1 M=0.5 N=17					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.74349022	4.14	3	36	0.0128
Pillai's Trace	0.25650978	4.14	3	36	0.0128
Hotelling-Lawley Trace	0.34500760	4.14	3	36	0.0128
Roy's Greatest Root	0.34500760	4.14	3	36	0.0128

The answer is 2.

Question 4.5 - 4

If we assume that the prior probabilities are the same, we will classify a cell as belonging to donor (type) 3 if [areacell, areanucl, avecytop][a b c]' + d > 0, where a, b, c and d are constants. The first coefficient a is:

According to the Theorem 5.4:

Theorem 5.4

Let $\pi_1 \sim N(\mu_1, \Sigma)$ and $\pi_2 \sim N(\mu_2, \Sigma)$. Then we have

$$\frac{f_1(x)}{f_2(x)} \geq c \Leftrightarrow x^T\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 \geq \log c$$

$$\Leftrightarrow \left[x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right] - \left[x^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2\right] \geq \log c\,.$$

The value of a is given by subtraction the highlighted values:

Linear Discriminant Function for donor		
Variable	**3**	**4**
Constant	-1695	-1684
areacell	357.55281	356.61461
areanucl	-199.41315	-198.47936
avecytop	-20.93522	-21.04164

357.55281-356.61461=0.9382

The answer is 4.

Question 4.6 - 4

The fraction of correctly classified cells from donor 3 using resubstitution is:

The answes the fraction of the true positives for donor 3 to the total number of cells : 15/20

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CELLS
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into donor			
From donor	3	4	Total
3	15	5	20
	75.00	25.00	100.00
4	7	13	20
	35.00	65.00	100.00
Total	22	18	40
	55.00	45.00	100.00
Priors	0.5	0.5	

Error Count Estimates for donor			
	3	4	Total
Rate	0.2500	0.3500	0.3000
Priors	0.5000	0.5000	

The answer is 4.

Question 4.7 - 2

The fraction of correctly classified cells is reduced if we use cross validation instead of resubstitution. The reduction is:

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CELLS
Cross-validation Summary using Linear Discriminant Function

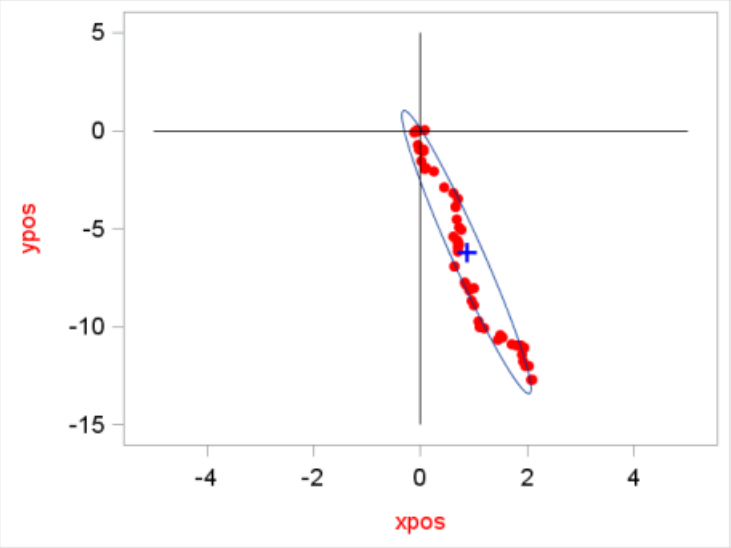
Number of Observations and Percent Classified into donor			
From donor	3	4	Total
3	14	6	20
	70.00	30.00	100.00
4	8	12	20
	40.00	60.00	100.00
Total	22	18	40
	55.00	45.00	100.00
Priors	0.5	0.5	

The fraction of the correctly classified cells when we use cross validation is 14/20. So it is reduced by 1/20=0.05

The answer is 2.

Problem 5

Enclosure D with SAS program and SAS output belongs to this problem. The data are taken from a study on the position of a certain phenomenum in microscopy images of materials samples. In the figure below is shown 58 such positions and the first 5 observations are given in the SAS output in Enclosure D.



Question 5.1 - 5

What is the value of the first principal component corresponding to observation no 1?

The 5 First Observations in the Position Data Set

Obs	num	xpos	ypos
1	14	2.07786	-12.7075
2	15	2.07041	-12.6801
3	29	2.01725	-12.0257
4	38	1.97279	-12.0029
5	42	1.96867	-11.9945

Eigenvectors		
	Prin1	Prin2
xpos	-.159745	0.987158
ypos	0.987158	0.159745

Simple Statistics		
	xpos	ypos
Mean	0.8598490508	-6.235902582
StD	0.7044660470	4.207708857

To Find the answer we need to use all the above Tables.

The value of the principal component 1 is given by the sum of the multiplication of the eigenvectors of Prin1 with the corresponding xpos and ypos reduced by the mean values.:

$$-0.159745 \cdot (2.07786 - 0.8598490508) + 0.987158 \cdot (-12.7075 - (-6.235902582))$$

The answer is 5.

Question 5.2 - 3

The value of the second principal component corresponding to observation no 1 lies in the interval:

The same way for the principal component 2 we calculate:

$$0.987158 \cdot (2.07786 - 0.8598490508) + 0.159745 \cdot \left(-12.7075 - (-6.235902582)\right) = 0.16889508$$

The answer is 3.

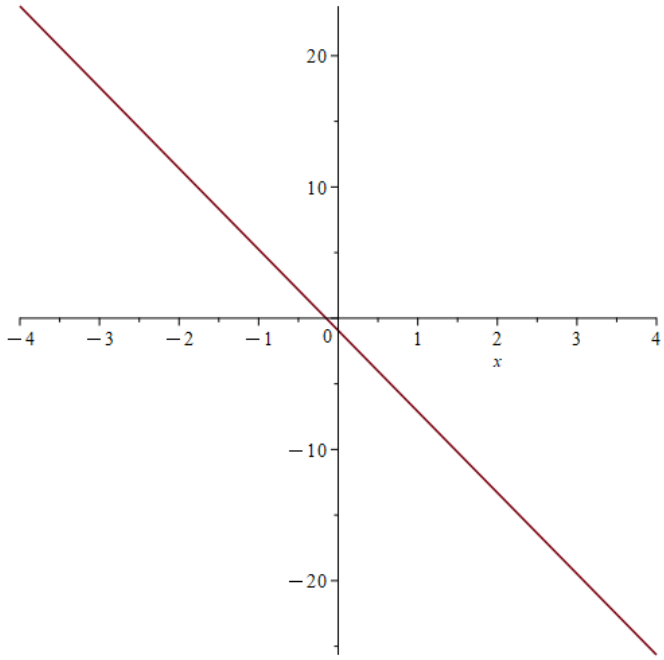
Question 5.3 - 5

If we assume that the data are normally distributed, then the major axis of the contour ellipse for the estimated probability density function lies on the line with the equation:

From the two eigenvectors: $\begin{matrix} 0.99 \\ 0.16 \end{matrix}$, $\begin{matrix} -0.16 \\ 0.99 \end{matrix}$ we can find the two axis respectively:

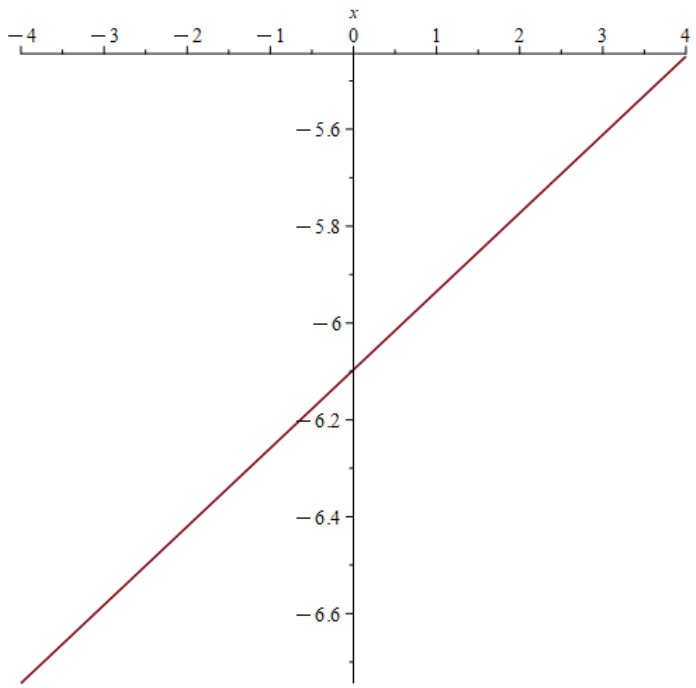
$0.987158 \cdot (xpos - 0.8598490508) + 0.159745 \cdot \left(ypos - (-6.235902582)\right)$=0 =>

$0.987158 \cdot (xpos) + 0.159745 \cdot (ypos) + 0.1473473887$=0



$-0.159745 \cdot (xpos - 0.8598490508) + 0.987158 \cdot \left(ypos - (-6.235902582)\right) = 0$=>

$-0.159745 \cdot (xpos) + 0.987158 \cdot (ypos) + 6.018464534 = 0$:



From the statement of the problem it can be seen that the major axis is the one that is given from the first equation.

The answer is 5.