Exam 2022 - 02409 Multivariate Statistics

4 hours written exam - All aids allowed, including internet access.

There are **7** problems with a total of **30** questions.
The questions in a specific problem should be read in order.

The questions are weighted equally

A correct answer gives 5 points, a wrong answer gives –1 point. Unanswered questions or a “don’t know” give 0 points. The total number of
points needed for a satisfactorily answered exam is determined at the final evaluation of the exam. Especially note that the grade 10 may
be given even if only one answer is wrong or unanswered.

For each problem there is a link to an enclosure or a dataset, if needed to solve that problem

Problem 1

We consider a dataset regarding breakfast products. The dataset is from https://www.kaggle.com (https://www.kaggle.com/datasets/crawford/80-cereals?resource=download)

However, we will use a reduced version where one manufacturer has been removed. You can find the version we will be using as a SAS file here (https://resources.mcq.eksamen.dtu.dk/v1/1ec71786-d67a-45b2-a3a6-e48cc79e7ceb) or a R-file here (https://resources.mcq.eksamen.dtu.dk/v1/fde14c4e-8b50-4c9d-be96-509badeb53a5) .

In the dataset we have the following variables:

Name	Name of cereal

mfr	• Manufacturer of cereal
	  ◦ G = General Mills
	  ◦ K = Kelloggs
	  ◦ N = Nabisco
	  ◦ P = Post
	  ◦ Q = Quaker Oats
	  ◦ R = Ralston Purina

type	cold / hot
calories	calories per serving
protein	grams of protein
fat	grams of fat
sodium	milligrams of sodium
fiber	grams of dietary fiber
carbo	grams of complex carbohydrates
sugars	grams of sugars
potass	milligrams of potassium
vitamins	vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
shelf	display shelf (1, 2, or 3, counting from the floor)
weight	weight in ounces of one serving
cups	number of cups in one serving
rating	a rating of the cereals

We now consider if there are any differences between the products by different manufactures by means of the following model:

$$\begin{bmatrix} calories & protein & fat & sodium & fiber & carbo & sugars & potass & rating \end{bmatrix} = \mu + mfr_i \qquad i = 1, \ldots, 6$$

i.e., a one-way MANOVA.

Question 1.1

We test that all manufacturers have the same mean against all alternatives. Wilks Lambda for this test is:

Vælg en svarmulighed

- 0.0001
- 0.1661
- 3.08
- 1.56582
- 0.81248174
- Don't Know

Question 1.2

As measured solely by the type III SS, we see the largest difference between the manufacturers in the following variable:

Vælg en svarmulighed

- calories
- Don't know
- protein
- sodium
- potass
- rating

Question 1.3

We consider the null-hypothesis, that all manufacturers have the same mean, against all alternatives. If the the null-hypothesis is true, the usual test-statistic will follow the distribution:

Vælg en svarmulighed

- U(9,8,40)
- U(6,8,68)
- U(9,5,70)
- Don't know
- F(45, 330)
- F(3.08, 45)

Question 1.4

Without consideration of the previous question: We now assume a distribution
$U(s, r, n\text{-}k)=U(1,2,n\text{-}2)$
where n is the number of observations. We further have an observed test-statistic $u=0.2$.
The lowest number of observations, for which the test-statistic u is significant at the 5%-level, is:

Vælg en svarmulighed

- Don't know
- 5
- 6
- 9
- 8
- 7

Problem 2

We still consider the data introduced in Problem 1.

We now consider whether we can discriminate between the different manufacturers by means of a *Linear Discriminant Analysis.*
We *only* consider the following variables: *calories protein fat sodium fiber carbo sugars potass .*

Question 2.1

The number of (resubstitution) misclassifications when performing a linear discriminant analysis with equal losses and equal prior probabilities are:

Vælg en svarmulighed

- Don't know
- 18
- 23
- 30
- 45
- 72

Question 2.2

Based on Mahalanobis Distance the two manufacturers that are most difficult to discriminate, are:

Vælg en svarmulighed

- K and P
- N and Q
- Don't know
- G and R
- K and R
- G and N

Question 2.3

We want to test if there is a significant difference between manufacturers K and R. The Hotelling's T^2 test for difference in mean values, yields a P-value, that falls in the following interval:

Vælg en svarmulighed

- $[0.35,\ 0.45[$
- $[0.15,\ 0.25[$
- $[0.05,\ 0.15[$
- $[0.45,\ 0.55[$
- $[0.25,\ 0.35[$
- Don't know

Question 2.4

We will now consider if some of the variables are contributing to the discrimination or can be discarded. We will again consider the manufacturers K and R.
The usual test for the hypothesis that *protein* and *potass* **do not** contribute to the discrimination against all alternatives, yields a p-value in the following interval:

Vælg en svarmulighed

- $[0.8,\ 1]$
- $[0.4,\ 0.6[$
- $[0.6,\ 0.8[$
- $[0.2,\ 0.4[$
- $[0,\ 0.2[$
- Don't know

Problem 3

We still consider the breakfast data introduced in problem 1. For convenience, the description of the data is repeated below.

In the dataset we have the following variables:

Name	Name of cereal

mfr	• Manufacturer of cereal
	  ◦ G = General Mills
	  ◦ K = Kelloggs
	  ◦ N = Nabisco
	  ◦ P = Post
	  ◦ Q = Quaker Oats
	  ◦ R = Ralston Purina

type	cold / hot
calories	calories per serving
protein	grams of protein
fat	grams of fat
sodium	milligrams of sodium
fiber	grams of dietary fiber
carbo	grams of complex carbohydrates
sugars	grams of sugars
potass	milligrams of potassium
vitamins	vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
shelf	display shelf (1, 2, or 3, counting from the floor)
weight	weight in ounces of one serving
cups	number of cups in one serving
rating	a rating of the cereals

We now consider the following variables: *calories protein fat sodium fiber carbo sugars potass*
We will perform a factor analysis with VARIMAX rotation. The factors should be estimated using the principal factor solution.

NOTE: In this problem you will be provided with the necessary outputs to answer each of the questions. The outputs will appear in the individual questions.

Question 3.1

If we were to perform a Principal Component Analysis on the variables listed above, i.e., *calories protein fat sodium fiber carbo sugars potass,* we should use the following matrix for the following reason:

Vælg en svarmulighed

- ○ The covariance matrix, as we retain the original scale of the variables
- ○ Don't know
- ○ The correlation matrix, as it compresses the data
- ○ The correlation matrix, since the variables are on a very different scale.
- ○ It does not matter what matrix we use. PCA automatically rescales the data .
- ○ The covariance matrix since the data are on a very different scale. That mean we would lose a lot of information, if we use the correlation matrix.

Question 3.2

Eigenvalues of the Correlation Matrix:
Number of observations 76

	Eigenvalue	Difference	Proportion	Cumulative
1	2.65959651	0.67008608	0.3324	0.3324
2	1.98951043	0.57006752	0.2487	0.5811
3	1.41944291	0.53162008	0.1774	0.7586
4	0.88782283	0.34813403	0.1110	0.8695
5	0.53968880	0.15970917	0.0675	0.9370
6	0.37997963	0.31107246	0.0475	0.9845
7	0.06890717	0.01385543	0.0086	0.9931
8	0.05505174		0.0069	1.0000

We consider how many factors to retain. Regardless of the previous question, we consider the correlation matrix. The test-statistic for the last two eigenvalues of the correlation matrix being equal, against all alternatives is:

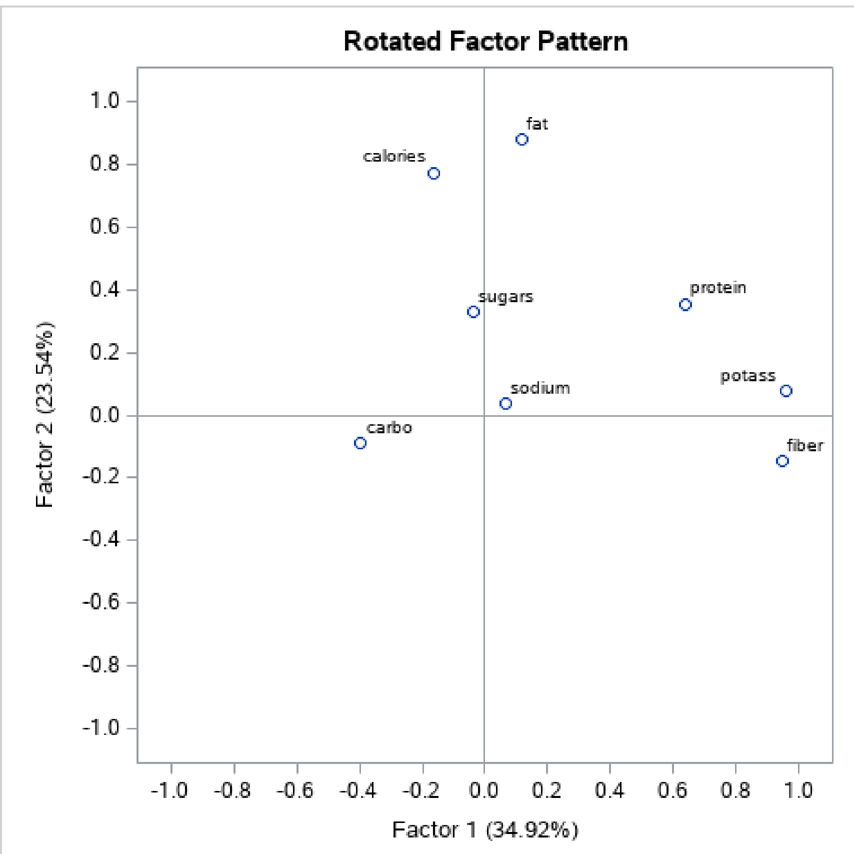
Vælg en svarmulighed

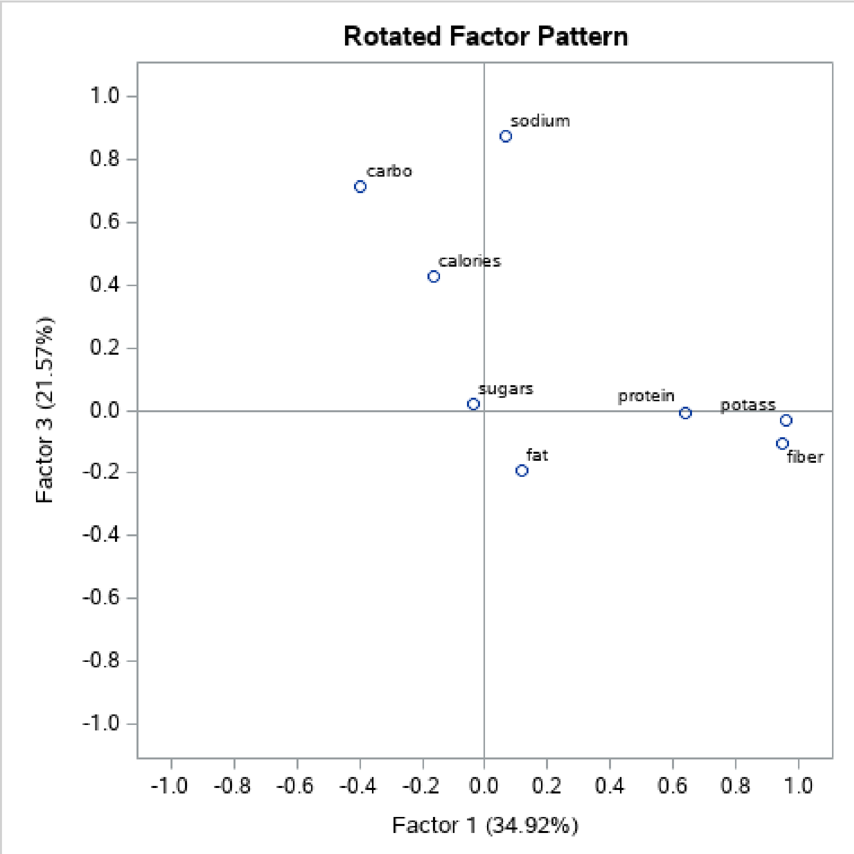
- 0.9555
- 0.01385543
- 264.9871
- 0.9931
- Don't know
- 0.8675

Question 3.3

Rotated Factor Pattern

	Factor1	Factor2	Factor3	Factor4
calories	-0.16104	0.77323	0.42776	0.28348
protein	0.63903	0.34998	-0.01125	-0.54056
fat	0.11557	0.87899	-0.19110	0.08489
sodium	0.06443	0.03574	0.87364	0.13687
fiber	0.94843	-0.14699	-0.10602	-0.06173
carbo	-0.39805	-0.08910	0.71036	-0.40915
sugars	-0.03852	0.32858	0.02119	0.90440
potass	0.95801	0.07735	-0.03346	0.04029





We consider a factor analysis - as described in the problem - *with 4 VARIMAX rotated factors.* The VARIMAX rotated factor 1 and factor 3 fits the following description:

Vælg en svarmulighed

- Rotated Factor 1 is mainly a weighting of *calories* and *fat,* and to a lesser degree sugars and protein.
  Rotated Factor 3 is mainly a weighting of *sodium*, *carbo*, and *calories.*
- Rotated Factor 1 is mainly a contrast between *fiber*, *potass*, and *sodium* vs. the remaining variables
  Rotated Factor 3 is mainly a weighting of *calories* and *fat,* and to a lesser degree sugars and protein.
- Rotated Factor 1 is mainly a contrast between *fiber*, *potass*, and *protein* vs. *carbo.*
  Rotated Factor 3 is mainly a weighting of *sodium*, *carbo*, and *calories.*
- Rotated Factor 1 is mainly an average of all variables
  Rotated Factor 3 is mainly a contrast between *calories* and *fat* vs. *carbo* and *fiber*
- Rotated Factor 1 is mainly a contrast between *fiber*, *potass*, and *protein* vs. *carbo.*
  Rotated Factor 3 is mainly a contrast between *sugars* vs. *carbo* and *fat*
- Don't know

Question 3.4

Rotated Factor Pattern

	Factor1	Factor2	Factor3	Factor4
calories	-0.16104	0.77323	0.42776	0.28348
protein	0.63903	0.34998	-0.01125	-0.54056
fat	0.11557	0.87899	-0.19110	0.08489
sodium	0.06443	0.03574	0.87364	0.13687
fiber	0.94843	-0.14699	-0.10602	-0.06173
carbo	-0.39805	-0.08910	0.71036	-0.40915
sugars	-0.03852	0.32858	0.02119	0.90440
potass	0.95801	0.07735	-0.03346	0.04029

Final Communality Estimates

calories	protein	fat	sodium	fiber	carbo	sugars	potass
0.88715595	0.82316829	0.82971237	0.78740839	0.93618199	0.83840423	0.92783529	0.92650616

We again consider a factor analysis - as described in the problem - *with 4 VARIMAX rotated factors*. The uniqueness of *protein* is:

Vælg en svarmulighed

- $\frac{0.8232}{8}$
- 6.956
- $0.63903^2 + 0.34998^2 + 0.01125^2 + 0.54056^2$
- 0.8232
- Don't know
- $1 - 0.63903^2 - 0.34998^2 - 0.01125^2 - 0.54056^2$

Question 3.5

Rotated Factor Pattern

	Factor1	Factor2	Factor3	Factor4
calories	-0.16104	0.77323	0.42776	0.28348
protein	0.63903	0.34998	-0.01125	-0.54056
fat	0.11557	0.87899	-0.19110	0.08489
sodium	0.06443	0.03574	0.87364	0.13687
fiber	0.94843	-0.14699	-0.10602	-0.06173
carbo	-0.39805	-0.08910	0.71036	-0.40915
sugars	-0.03852	0.32858	0.02119	0.90440
potass	0.95801	0.07735	-0.03346	0.04029

Final Communality Estimates

calories	protein	fat	sodium	fiber	carbo	sugars	potass
0.88715595	0.82316829	0.82971237	0.78740839	0.93618199	0.83840423	0.92783529	0.92650616

We again consider a factor analysis - as described in the problem - *with 4 VARIMAX rotated factors*. The VARIMAX rotated factor 3 and 4 together explains the following fraction of the variance in *carbo*:

Vælg en svarmulighed

- 0.6720
- Don't know
- 0.3372
- 0.8384
- 0.3012
- 1.1195

Problem 4

We stil consider the breakfast data introduced in Problem 1.

We now consider a regression task, where we want to predict rating by means of the following model:
$rating = \mu + \beta_1 \cdot calories + \beta_2 \cdot protein + \beta_3 \cdot fat + \beta_4 \cdot sodium + \beta_5 \cdot fiber + \beta_6 \cdot carbo + \beta_7 \cdot sugars + \beta_8 \cdot potass + \epsilon$
where μ is an intercept, the eight β's are coefficients for the different variables, and ϵ is the residual.

Question 4.1

Given the model stated in the problem, we explain the following fraction of variation in the *rating* variable:

Vælg en svarmulighed

- $1 - \frac{79.97979}{14766}$
- 0.9094
- $\frac{79.9798}{14766}$
- 0.7997
- Don't know
- $1 - \frac{2.57045}{42.50537}$

Question 4.2

The test statistic for testing the hypothesis, that the fraction of variation of the *rating* variable explained by the model is zero, will - if true - follow the following distribution:

Vælg en svarmulighed

- ○ Don't know
- ○ t(74)
- ○ F(8,67)
- ○ $\chi^2(65)$
- ○ t(65)
- ○ F(8,75)

Question 4.3

Given the model stated in the problem, we consider if we have problems with multicolinearity.

Vælg en svarmulighed

- ○ The lowest tolerance found is 0.111 and the largest VIF is 8.97. We are thus within the rules of thumb and multicolinearity does seem to be a problem.
- ○ VIF and tolerance gives complementary information and we need to consider both. Based on that multicolinearity does not seem to be a problem.
- ○ Since three variables have a tolerance lower than 0.1 and three observations have a VIF larger than 10 multicolinearity does not seem to be a problem
- ○ The lowest tolerance found is 0.111 and the largest VIF is 8.97. We are thus outside the rules of thumb and multicolinearity does seem to be a problem.
- ○ Since three variables have a tolerance lower than 0.1 and three observations have a VIF larger than 10 multicolinearity seem to be a problem
- ○ Don't know

Question 4.4

Given the model stated in the problem, the most influential observation, as measured by both Cook's D and DFFITS is:

Vælg en svarmulighed

- 57
- 53
- 69
- Don't know
- 40
- 70

Question 4.5

Given the design matrix (**X**-matrix) stated in the problem, the observation with the largest *potential* to influence the estimates of the model parameters is:

Vælg en svarmulighed

- Don't know
- 70
- 67
- 4
- 69
- 57

Question 4.6

Given the model stated in the problem, when performing a backward selection, the first variable that might be removed based on the F-value is:

Vælg en svarmulighed

- *potass*
- *sodium*
- *sugars*
- *fat*
- *calories*
- Don't know

Question 4.7

Given the model stated in the problem, we consider what constitutes a good breakfast. The highest rating as judged by the parameter estimates is achieved when:

Vælg en svarmulighed

- ○ We have a breakfast with a large amount of sugars and fat and minimal amount of fiber.
- ○ We cannot tell how to get a good rating based on the parameter estimates, due to the model explaining rating poorly.
- ○ We have a breakfast with a large intercept, as that is the largest parameter numerically.
- ○ We have a breakfast with a large amount of fiber and protein but no sodium or carbo. As we can see from the t-value, sodium has a large negative influence on the rating.
- ○ Don't know
- ○ We have a breakfast with fiber, protein and carbohydrates. The remaining variables should be minimised.

Problem 5

We now leave the exciting world of breakfast products and consider something else entirely.

We consider independent random variables $Y_i \sim N(\mu_i, \sigma^2)$, organized in a two-way layout with expected values as presented in the table below.

$$
\begin{array}{lll}
 & \text{columns} & \\
\text{rows} & E\left(Y_1\right) = \mu + \alpha & E\left(Y_2\right) = \mu - \alpha \\
 & \mathrm{E}\left(Y_3\right) = \mu - \alpha & E\left(Y_4\right) = \mu + \alpha + \beta
\end{array}
$$

In the sequel, you may find the following expressions useful

$$
\begin{bmatrix} 4 & 0 & 1 \\ 0 & 4 & 1 \\ 1 & 1 & 1 \end{bmatrix}^{-1} = \frac{1}{8}\begin{bmatrix} 3 & 1 & -4 \\ 1 & 3 & -4 \\ -4 & -4 & 16 \end{bmatrix}
$$

$$
\frac{1}{8}\begin{bmatrix} 3 & 1 & -4 \\ 1 & 3 & -4 \\ -4 & -4 & 16 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 2 & 1 & 1 & 0 \\ 2 & -1 & -1 & 0 \\ -4 & 0 & 0 & 4 \end{bmatrix}
$$

Question 5.1

The ordinary least squares estimator $\hat{\mu}$ of μ is of the form $aY_1 + bY_2 + cY_3 + dY_4$, where $\begin{bmatrix} a & b & c & d \end{bmatrix}$ is:

Vælg en svarmulighed

- $\frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$
- $\frac{1}{2}\begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix}$
- $\frac{1}{4}\begin{bmatrix} 2 & -1 & -1 & 0 \end{bmatrix}$
- $\frac{1}{4}\begin{bmatrix} 2 & 1 & 1 & 0 \end{bmatrix}$
- $\frac{1}{2}\begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}$
- Don't know

Question 5.2

The variance of $\hat{\mu}$ is:

Vælg en svarmulighed

- $\frac{1}{4}\sigma^2$
- $\frac{1}{2}\sigma^2$
- $\frac{1}{16}\sigma^2$
- $\frac{3}{8}\sigma^2$
- $\frac{1}{8}\sigma^2$
- Don't know

Question 5.3

The ordinary least squares estimator $\hat{\beta}$ of β is of the form $eY_1 + fY_2 + gY_3 + hY_4$, where $\begin{bmatrix} e & f & g & h \end{bmatrix}$ is:

Vælg en svarmulighed

- $\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix}$
- $\begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}$
- $\begin{bmatrix} -1 & 0 & 0 & 1 \end{bmatrix}$
- $\frac{1}{2}\begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix}$
- $\frac{1}{2}\begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}$
- Don't know

Question 5.4

The correlation between $\hat{\mu}$ and $\hat{\beta}$ is:

Vælg en svarmulighed

- Don't know
- $-\frac{\sqrt{3}}{3}$
- $-\frac{1}{2}$
- $\frac{1}{8}$
- $\frac{3}{8}$
- 0

Question 5.5

We now assume that $\beta = 0$.The ordinary least squares estimator for μ under this assumption is called $\hat{\mu}$. It is of the form $kY_1 + lY_2 + mY_3 + nY_4$, where $\begin{bmatrix} k & l & m & n \end{bmatrix}$ is:

Vælg en svarmulighed

- $\frac{1}{2}\begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix}$
- $\frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$
- $\frac{1}{4}\begin{bmatrix} 2 & 1 & 1 & 0 \end{bmatrix}$
- Don't know
- $\frac{1}{2}\begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}$
- $\frac{1}{4}\begin{bmatrix} 2 & -1 & -1 & 0 \end{bmatrix}$

Question 5.6

The correlation between the previously found two estimators for μ, i.e. $Corr(\hat{\mu}, \hat{\mu}) = Corr(aY_1 + bY_2 + cY_3 + dY_4,\ kY_1 + lY_2 + mY_3 + nY_4)$, is:

Vælg en svarmulighed

- $\frac{1}{16}$
- $\frac{3}{8}$
- Don't know
- $\frac{\sqrt{6}}{3}$
- $\frac{\sqrt{3}}{3}$
- $\frac{1}{4}$

Problem 6

We consider a three dimensional multivariate normal distribution, for which we know that

$$D(\boldsymbol{Y}_i) = \begin{bmatrix} \sigma & \gamma & \rho \\ \gamma & \sigma & \eta \\ \rho & \eta & \sigma \end{bmatrix}$$

We have 2 observations

Then $\mathrm{D}(\mathrm{vc}(\boldsymbol{Y}))$ is:

Vælg en svarmulighed

○

$$\begin{bmatrix} \sigma & \gamma & \rho & \sigma & \gamma & \rho \\ \gamma & \sigma & \eta & \gamma & \sigma & \eta \\ \rho & \eta & \sigma & \rho & \eta & \sigma \\ \sigma & \gamma & \rho & \sigma & \gamma & \rho \\ \gamma & \sigma & \eta & \gamma & \sigma & \eta \\ \rho & \eta & \sigma & \rho & \eta & \sigma \end{bmatrix}$$

○

$$\begin{bmatrix} \sigma & \sigma & \gamma & \gamma & \rho & \rho \\ \sigma & \sigma & \gamma & \gamma & \rho & \rho \\ \gamma & \gamma & \sigma & \sigma & \eta & \eta \\ \gamma & \gamma & \sigma & \sigma & \eta & \eta \\ \rho & \rho & \eta & \eta & \sigma & \sigma \\ \rho & \rho & \eta & \eta & \sigma & \sigma \end{bmatrix}$$

○ Don't know

○

$$\begin{bmatrix} \sigma & \gamma & \rho & \eta & 0 & 0 \\ \gamma & \sigma & \gamma & \rho & \eta & 0 \\ \rho & \gamma & \sigma & \gamma & \rho & \eta \\ \eta & \rho & \gamma & \sigma & \gamma & \rho \\ 0 & \eta & \rho & \gamma & \sigma & \gamma \\ 0 & 0 & \eta & \rho & \gamma & \sigma \end{bmatrix}$$

○

$$\begin{bmatrix} \sigma & \gamma & \rho & 0 & 0 & 0 \\ \gamma & \sigma & \eta & 0 & 0 & 0 \\ \rho & \eta & \sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma & \gamma & \rho \\ 0 & 0 & 0 & \gamma & \sigma & \eta \\ 0 & 0 & 0 & \rho & \eta & \sigma \end{bmatrix}$$

○

$$\begin{bmatrix} \sigma & 0 & \gamma & 0 & \rho & 0 \\ 0 & \sigma & 0 & \gamma & 0 & \rho \\ \gamma & 0 & \sigma & 0 & \eta & 0 \\ 0 & \gamma & 0 & \sigma & 0 & \eta \\ \rho & 0 & \eta & 0 & \sigma & 0 \\ 0 & \rho & 0 & \eta & 0 & \sigma \end{bmatrix}$$

Problem 7

As part of her research for her bachelor project "*Data-Driven Investigation of Endotypes in Knee Osteoarthritis Patients*" , Zofia Lisowska-Petersen looked at the relation between a clinical dataset with 6 variables describing knee functionality, and a bio-marker dataset with 11 variables. There were 528 observations in both datasets.

A Canonical Correlation Analysis yielded the following correlations:

	CV1	CV2	CV3	CV4	CV5	CV6
Canonical Correlation	0.3083	0.2767	0.2022	0.1362	0.1007	0.0485

Question 7.1

How much of the variation in the first biomarker canonical variable is explained by the first clinical canonical variable:

Vælg en svarmulighed

- 0.3083 + 0.2767 + 0.2022 +0.1362 + 0.1007 + 0.0485
- 0.3083
- $\frac{0.3083}{6+11}$
- Cannot be answered with the information provided
- Don't know
- 0.0950

Question 7.2

We want to test if the two datasets are independent against all alternatives. The usual test-statistic for this test, follows - if the null hypothesis is true - the following distribution:

Vælg en svarmulighed

- U(11,6,521)
- U(6,11,516)
- F(17,528)
- Don't know
- U(6,11,50)
- F(6,11)

Question 7.3

If the mean is unknown, the minimum number of observations needed to get a full rank estimate of Σ_{xx}, i.e. the dataset of bio-markers, is:

Vælg en svarmulighed

- Don't know
- 10
- 132
- 121
- 7
- 12