Written examination, date: 9th of December 2019 *Page 1 of 27 pages* *Enclosure: XX pages*
Course name: Multivariate Statistics
Course number: 02409
Aids allowed: All
Exam duration: 4 hours
Weighting: The questions are given equal weight

This exam is answered by:

_____ _____ _____
(name) (signature) (study no.)

There is a total of 30 questions for the 6 problems. The answers to the 30 questions must be written into the table below.

Problem	1	1	1	1	1	1	1	1	2	2
Question	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	2.1	2.2
Answer	2	1	5	3	3	2	1	4	5	2

Problem	2	3	3	3	3	3	3	3	3	4
Question	2.3	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	4.1
Answer	4	1	1	2	4	4	3	2	4	4

Problem	4	4	5	5	5	6	6	6	6	6
Question	4.2	4.3	5.1	5.2	5.3	6.1	6.2	6.3	6.4	6.5
Answer	2	(3)	4	2	2	2	2	5	1	5

The possible answers for each question are numbered from 1 to 6. If you enter a wrong number, you may correct it by crossing the wrong number in the table and writing the correct answer immediately below. If there is any doubt about the meaning of a correction then the question will be considered not answered.

Only the front page must be returned. The front page must be returned even if you do not answer any of the questions or if you leave the exam prematurely. Drafts and/or comments are not considered, only the numbers entered above are registered.

A correct answer gives 5 points, a wrong answer gives – 1 point. Unanswered questions or a 6 (corresponding to "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.
Remember to write your name, signature, and study number on the front page.

Please note, that there is one and only one correct answer to each question. Furthermore, some of the possible alternative answers may not make sense. When the text refers to SAS-output, the values may be rounded to fewer decimal places than in the output itself. The enclosures do not necessarily contain all the output generated by the given SAS programs. Please check that all pages of the exam paper and the enclosures are present.

Problem 1.

You are encouraged to use statistical software in this problem.

We consider the following model

$$\begin{bmatrix} Y_1 & Z_1 & V_1 & W_1 \\ Y_2 & Z_2 & V_2 & W_2 \\ Y_3 & Z_3 & V_3 & W_3 \\ Y_4 & Z_4 & V_4 & W_4 \\ Y_5 & Z_5 & V_5 & W_5 \end{bmatrix} = \begin{bmatrix} 1 & -2 & -4 \\ 1 & -1 & -1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} \alpha_y & \alpha_z & \alpha_v & \alpha_w \\ \beta_y & \beta_z & \beta_v & \beta_w \\ \gamma_y & \gamma_z & \gamma_v & \gamma_w \end{bmatrix} + \begin{bmatrix} \delta_1 & \varepsilon_1 & \epsilon_1 & \vartheta_1 \\ \delta_2 & \varepsilon_2 & \epsilon_2 & \vartheta_2 \\ \delta_3 & \varepsilon_3 & \epsilon_3 & \vartheta_3 \\ \delta_4 & \varepsilon_4 & \epsilon_4 & \vartheta_4 \\ \delta_5 & \varepsilon_5 & \epsilon_5 & \vartheta_5 \end{bmatrix}$$

Where the error terms $[\delta_i \quad \varepsilon_i \quad \epsilon_i \quad \vartheta_i], for\ i = 1, \dots, 5$ are independent and normally distributed $N_4(\mathbf{0}, \mathbf{\Sigma})$, and where $\mathbf{\Sigma}$ is the unknown dispersion matrix.

We have obtained the following observations

$$\begin{bmatrix} 1 & 8 & 2 & 9 \\ 0 & 9 & 4 & 6 \\ 2 & 4 & 4 & 2 \\ 1 & 5 & 9 & 5 \\ 1 & 2 & 8 & 7 \end{bmatrix}$$

We can further calculate $(\boldsymbol{x}^T\boldsymbol{x})^{-1} = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 2.125 & -1.125 \\ 0 & -1.125 & 0.625 \end{bmatrix}$

Question 1.1.

The maximum likelihood estimate for the parameters $[\alpha_y \quad \beta_y \quad \gamma_y]$ *are*

This can be done by hand – or easier – using SAS. We input the data:

```
data MGLM;
input y z v w X1 X2 X3;
datalines;
1 8 2 9 1 -2 -4
0 9 4 6 1 -1 -1
2 4 4 2 1 0 0
1 5 9 5 1 1 1
1 2 8 7 1 2 4
;
```

We can either use

```
proc reg data=mglm ;
model y z v w = X1 X2 X3 /noint influence covb;
run;

proc glm data=mglm plots=all;
model y z v w = X1 X2 X3 /noint inverse ;
run;
```

with yields:
ANSWER 2: $[\mathbf{1} \quad \mathbf{1} \quad -\mathbf{0.5}]$

Question 1.2.

The covariance between the maximum likelihood estimates for α_x and β_x is

We can do immediately see from the given $(x^T x)^{-1} = \begin{bmatrix} 0.2 & 0 & 0 \\ 0 & 2.125 & -1.125 \\ 0 & -1.125 & 0.625 \end{bmatrix}$, that it is zero. We can also use

```
proc reg data=mglm ;
model x y z w = X1 X2 X3 /noint influence covb;
run;
```

which confirms the results
ANSWER 1: 0

Question 1.3.

The covariance between the maximum likelihood estimates for β_x and γ_x is

We already have $(x^T x)^{-1}$ given, we need to estimate σ_x^2, which can be done by theorem

Theorem 4.18

We consider the situation from theorem 4.14. Then the maximum likelihood estimate for Σ equals

$$\begin{aligned} \hat{\Sigma}^* &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\theta}^T x_i)(Y_i - \hat{\theta}^T x_i)^T \\ &= \frac{1}{n}(\mathbf{Y} - \mathbf{x}\hat{\theta})^T(\mathbf{Y} - \mathbf{x}\hat{\theta}) \\ &= \frac{1}{n}[\mathbf{Y}^T\mathbf{Y} - (\mathbf{x}\hat{\theta})^T(\mathbf{x}\hat{\theta})]. \end{aligned}$$

The (i, j)'th element can also be written

$$\hat{\sigma}_{ij}^* = \frac{1}{n}(\mathbf{Y}_{i|} - \mathbf{x}\hat{\theta}_{i|})^T(\mathbf{Y}_{j|} - \mathbf{x}\hat{\theta}_{j|}).$$

Alternatively, we use

```
proc reg data=mglm ;
```

```
model y z v w = X1 X2 X3 /noint influence covb;
run;
```
and get

Covariance of Estimates			
Variable	X1	X2	X3
X1	0.15	0	0
X2	0	1.59375	-0.84375
X3	0	-0.84375	0.46875

ANSWER 5: -0.84375

Question 1.4.

The variance of the maximum likelihood estimates for α_x is

We again need to estimate σ_x^2.
Alternatively we use

```
proc reg data=mglm ;
model y z v w = X1 X2 X3 /noint influence covb;
run;
```

Covariance of Estimates			
Variable	X1	X2	X3
X1	0.15	0	0
X2	0	1.59375	-0.84375
X3	0	-0.84375	0.46875

If we have used

```
proc glm data=mglm plots=all;
model y z v w = X1 X2 X3 /noint inverse ;
run;
```

or have forgotten the `covb` option, we can instead use the parameter output

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
X1	1	1.00000	0.38730	2.58	0.1229
X2	1	1.00000	1.26244	0.79	0.5113
X3	1	-0.50000	0.68465	-0.73	0.5412

$$0.38730^2 = 0.15$$

ANSWER 3: 0.15

Question 1.5.

The observation with the lowest leverage is

Can quickly be computed in Maple, Matlab etc, or we can use

```
proc reg data=mglm ;
model y z v w = X1 X2 X3 /noint influence covb;
run;
```

Output Statistics								
						DFBETAS		
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	X1	X2	X3
1	2.22E-16	3.31E-16	0.7000	26.6667	0.0000	0.0000	0.0000	-0.0000
2	-0.5000	-1.1180	0.7000	2.3411	-1.7078	-0.9129	1.4003	-1.2910
3	1.0000	2.2361	0.2000	0.0463	1.1180	1.1180	0.0000	0.0000
4	-0.5000	-1.1180	0.7000	2.3411	-1.7078	-0.9129	-1.4003	1.2910
5	0	0	0.7000	26.6667	0.0000	0.0000	0.0000	0.0000

ANSWER 3: 3

Question 1.6.

The dependent variable with the lowest MSE is:

Can quickly be computed in Maple, Matlab etc, or we can use either of

```
proc reg data=mglm ;
model y z v w = X1 X2 X3 /noint influence covb;
run;

proc glm data=mglm plots=all;
model y z v w = X1 X2 X3 /noint inverse ;
run;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5.50000	1.83333	2.44	0.3035
Error	2	1.50000	0.75000		
Uncorrected Total	5	7.00000			

Y

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	182.80000	60.93333	16.93	0.0563
Error	2	7.20000	3.60000		
Uncorrected Total	5	190.00000			

Z

V

Analysis of Variance					
Sou	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	176.30000	58.76667	25.01	0.0387
Error	2	4.70000	2.35000		
Uncorrected Total	5	181.00000			

W

Analysis of Variance					
Sourc	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	170.70000	56.90000	4.68	0.1810
Error	2	24.30000	12.15000		
Uncorrected Total	5	195.00000			

ANSWER 2: Y

We now test whether $[\beta_x \quad \beta_y \quad \beta_z \quad \beta_w]$ *are all equal to 0 with the following model*

$$H_0: \boldsymbol{A}\begin{bmatrix}\alpha_y & \alpha_z & \alpha_v & \alpha_w\\ \beta_y & \beta_z & \beta_v & \beta_w\\ \gamma_y & \gamma_z & \gamma_v & \gamma_w\end{bmatrix}\boldsymbol{B}^T = \boldsymbol{C} \quad vs. \quad H_1: \boldsymbol{A}\begin{bmatrix}\alpha_y & \alpha_z & \alpha_v & \alpha_w\\ \beta_y & \beta_z & \beta_v & \beta_w\\ \gamma_y & \gamma_z & \gamma_v & \gamma_w\end{bmatrix}\boldsymbol{B}^T \neq \boldsymbol{C}$$

Question 1.7.

In the above model A is equal to:

A selects the rows, and we need to select the second row.

ANSWER 1: [0 1 0]

Question 1.8.

The usual test-statistic for the above model has – under the null-hypothesis – the following distribution:

We find the relevant theorem

Theorem 4.21

We consider the above mentioned situation including the assumption of the normality of the observations. Furthermore we consider the hypothesis

$$H_0 : \mathbf{A}\,\boldsymbol{\theta}\,\mathbf{B}^T = \mathbf{C} \quad \text{against} \quad H_1 : \mathbf{A}\,\boldsymbol{\theta}\,\mathbf{B}^T \neq \mathbf{C},$$

where $\mathbf{A}(r \times k)$, $\mathbf{B}(s \times p)$ and $\mathbf{C}(r \times s)$ are given (known) matrices. We introduce

$$\begin{aligned}\boldsymbol{\Delta} &= \mathbf{A}\,\hat{\boldsymbol{\theta}}\,\mathbf{B}^T - \mathbf{C}\\ \mathbf{R} &= n\hat{\boldsymbol{\Sigma}}^* = (\mathbf{Y} - \mathbf{x}\,\hat{\boldsymbol{\theta}})^T(\mathbf{Y} - \mathbf{x}\,\hat{\boldsymbol{\theta}}) = \mathbf{Y}^T\mathbf{Y} - \hat{\boldsymbol{\theta}}^T(\mathbf{x}^T\mathbf{x})\hat{\boldsymbol{\theta}}\end{aligned}$$

and

$$\begin{aligned}\mathbf{E} &= \mathbf{B}\,\mathbf{R}\,\mathbf{B}^T\\ \mathbf{H} &= \boldsymbol{\Delta}^T[\mathbf{A}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{A}^T]^{-1}\boldsymbol{\Delta}.\end{aligned}$$

The likelihood ratio test for testing H_0 against H_1 is equivalent to the test given by the critical region

$$\{\mathbf{y}|\frac{\det(\mathbf{e})}{\det(\mathbf{e}+\mathbf{h})} \leq \mathrm{U}(s,r,n-k)_\alpha\},$$

where $\mathrm{U}(s,r,n-k)_\alpha$ is the α quantile in the null-hypothesis distribution of the test statistic (see below).

U(s, r, n-k)

We know C(1 x 4)
As we need the entire row B must be the identity matrix B(4 x 4)
And A is A(1 x 3)
Finally we have 5 observations

ANSWER 4: $U(4, 1, 5 - 3) = U(4, 1, 2)$

Problem 2.

Enclosure A with SAS program and SAS output belongs to this problem. We consider data for the 98 municipalities (kommuner) in Denmark. We have the rates (pr. 1000 capita) of different types of library use, e.g. book loan, music loan, etc. Further, we consider the educational levels as the fraction of the population with that educational level, e.g. a H1 at 0.25 means that 25 % of the population in a given municipality has primary school has the highest education. (Source http://www.statistikbanken.dk)

We shall now investigate the relations between the use of libraries and the educational level by means of a Canonical Correlation Analysis.

We consider the following variables for library use

SAS-name	Meaning
U1	Books
U2	Serial publications
U3	Audio books
U4	Music
U5	Live images (movies)
U6	Multi media material
U7	Other material

And for educational level

SAS-name	Meaning
H1	Primary school
H2	High school
H3	Vocational school
H4	Short further educations
H5	Medium further education
H6	Bachelor level
H7	Master level
H8	Ph.D. level

We shall now investigate the relations between the use of libraries and the educational level by means of a Canonical Correlation Analysis.

Question 2.1.

The first canonical correlation describes which fraction of the variation between V1 and W1

We find the squared canonical correlation in the output

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.7648	0.7188	0.0426	0.5849	1.4090	0.6740	0.4823	0.4823	0.1222	3.7700	56	441.51	<.0001
2	0.6509	0.5898	0.0591	0.4236	0.7350	0.2585	0.2516	0.7338	0.2943	2.7500	42	388.07	<.0001
3	0.5681	0.5318	0.0695	0.3227	0.4765	0.2950	0.1631	0.8969	0.5106	2.0400	30	334.00	0.0014
4	0.3919	0.3138	0.0868	0.1536	0.1815	0.0975	0.0621	0.9590	0.7538	1.2400	20	279.55	0.2182
5	0.2783	0.2110	0.0947	0.0774	0.0839	0.0487	0.0287	0.9877	0.8906	0.8400	12	225.18	0.6092
6	0.1845	0.1636	0.0991	0.0340	0.0352	0.0346	0.0121	0.9998	0.9654	0.5100	6	172.00	0.8004
7	0.0248	-0.1740	0.1025	0.0006	0.0006		0.0002	1.0000	0.9994	0.0300	2	87.00	0.9736

ANSWER 5: 0.5849

Question 2.2.

How much of the variance in U1 is explained by V1

We find the correlation between the two

Correlations Between the VAR Variables and Their Canonical Variables

	V1	V2	V3	V4	V5	V6	V7
U1	-0.7532	-0.4078	0.0131	0.4316	0.1035	0.2455	-0.0950
U2	0.2246	-0.5347	0.0256	0.3848	0.6136	0.3431	-0.1436
U3	-0.4422	0.2017	0.5650	0.3946	0.4631	0.1152	-0.2472
U4	-0.5616	0.4018	-0.3959	0.3622	0.4362	-0.0245	-0.2108
U5	-0.3860	0.1324	0.2111	0.2110	0.3682	0.6565	-0.4216

Correlations Between the VAR Variables and Their Canonical Variables							
	V1	V2	V3	V4	V5	V6	V7
U6	0.0958	0.4143	0.1493	0.6380	0.0214	0.6237	-0.0204
U7	-0.4215	0.2987	-0.0629	-0.0146	0.3268	0.2678	0.7419

We square it $(-0.7532)^2 = 0.5673$
ANSWER 2: 0.5673

Question 2.3.

The first canonical variate W1 can be interpreted as

We find it in the output

Correlations Between the WITH Variables and Their Canonical Variables							
	W1	W2	W3	W4	W5	W6	W7
H1	0.8011	0.2581	0.1093	0.1871	0.2245	-0.0680	0.1485
H2	-0.5932	0.4694	-0.3295	-0.0183	-0.3781	0.1419	-0.1080
H3	0.8678	-0.0786	0.2586	0.0179	0.3341	-0.2469	0.0214
H4	-0.4394	0.1912	0.3919	-0.1344	-0.3453	-0.4744	-0.4129
H5	-0.7879	0.0105	-0.3570	-0.1090	0.1755	-0.4051	-0.1169
H6	-0.5541	0.2595	-0.4377	-0.3858	-0.3719	0.2241	0.0428
H7	-0.8623	-0.0016	-0.1228	-0.2225	-0.2708	0.3133	0.0311
H8	-0.8735	-0.0707	-0.0324	-0.0798	-0.2877	0.2479	0.2460

We see that it is a contrast between primary and vocational school against all other educations.

ANSWER 4: A contrast between primary and vocational school against all other educations.

Problem 3.

Enclosure B with SAS program and SAS output belongs to this problem. As in Problem 2, we consider library data, but now for the Capital region only. We want to predict the music loans (pr. 1000 capita) pr. municipality (the U4 variable in Problem 2) based on 5 financial variables (in 1.000 DKK pr. 1.000 capita) for the libraries in each municipality

SAS-name	Meaning
F1	Salary costs
F2	Material costs
F3	Other costs
F4	Income (e.g. late fees)
F5	Netto costs

We consider two models – all with an intercept:

- M1: All variables
- M2: which is the resulting model, after we have reduced the number of explanatory variables by stepwise model selection.

Question 3.1.

The reduction in variance explained when going from model M1 to model M2 is

We find in the enclosure for model M1

Root MSE	73.53864	R-Square	0.6151
Dependent Mean	169.58621	Adj R-Sq	0.5315
Coeff Var	43.36357		

And for model M2

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	F1		1	0.5289	0.5289	3.1545	30.31	<.0001
2	F4		2	0.0461	0.5750	2.3998	2.82	0.1051

The reduction is thus 0.6151-0.575 = 0.0401

ANSWER 1: 0.0401

Question 3.2.

If we performed backwards elimination from M1, the first variable to be excluded is

We find in the output for model M1

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-41.74991	72.35605	-0.58	0.5695	-191.42981	107.92999
F1	1	-22.42111	25.45826	-0.88	0.3876	-75.08553	30.24331
F2	1	-24.16698	25.64101	-0.94	0.3557	-77.20945	28.87549
F3	1	-23.23197	25.60095	-0.91	0.3736	-76.19156	29.72762
F4	1	23.84917	25.64438	0.93	0.3620	-29.20028	76.89862
F5	1	23.14720	25.53541	0.91	0.3741	-29.67682	75.97123

And see it is F1

ANSWER 1: F1

Question 3.3.

What is the usual test statstic for M1 vs M2:

We find the ANOVA tables in the output

The REG Procedure
Model: MODEL1
Dependent Variable: U4

Number of Observations Read	29
Number of Observations Used	29

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	198793	39759	7.35	0.0003
Error	23	124382	5407.93108		
Corrected Total	28	323175			

And for model M2

Stepwise Selection: Step 2

Variable F4 Entered: R-Square = 0.5750 and C(p) = 2.3998

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	185815	92907	17.59	<.0001
Error	26	137360	5283.08351		
Corrected Total	28	323175			

The relevant test is given in

Test statistic for $H_0: E(\boldsymbol{Y}) \in H_{i+1}$ against $H_1: E(\boldsymbol{Y}) \in H_i \backslash H_{i+1}$:

$$\frac{\left\|p_{H_i}(Y)-p_{H_{i+1}}(Y)\right\|^2/(r_i-r_{i+1})}{\left\|Y-p_{H_i}(Y)\right\|^2/(n-r_i)} = \frac{[SS_{res}(H_{i+1})-SS_{res}(H_i)]/[DF_{res}(H_{i+1})-DF_{res}(H_i)]}{SS_{res}(H_i)/DF_{res}(H_i)}$$

We insert:

$$\frac{(137360-124382)/(26-23)}{\frac{124382}{23}}$$

ANSWER 2

Question 3.4.

The distribution of the above statistic under the null hypothesis is

We find in the notes:

$$\left\{Y \,\middle|\, F > F(DF_{res}(H_{i+1}) - DF_{res}(H_i), DF_{res}(H_i))_{1-p}\right\}$$

F(26-23,23)

ANSWER 4: F(3,23)

We now only consider M2

Question 3.5.

The observation with the highest leverage is

We find the relevant table in the output

Dependent Variable: U4

Output Statistics										
					Hat			DFBETAS		
Obs	Dependent Variable	Predicted Value	Residual	RStudent	Diag H	Cov Ratio	DFFITS	Intercept	F1	F4
1	103	66.9892	36.0108	0.5196	0.1163	1.2326	0.1885	0.1776	-0.1581	0.0469
2	297	173.7829	123.2171	1.7979	0.0346	0.8090	0.3403	0.0783	0.0182	-0.0009
3	55	245.8493	-190.8493	-3.2262	0.0979	0.4389	-1.0629	0.5743	-0.8387	0.4138
4	254	262.8382	-8.8382	-0.1276	0.1257	1.2840	-0.0484	0.0294	-0.0405	0.0192
5	447	329.3542	117.6458	1.8943	0.1972	0.9370	0.9389	-0.7064	0.8379	-0.1010
6	136	236.8447	-100.8447	-1.4649	0.0635	0.9382	-0.3813	0.0933	-0.1382	-0.1657
7	53	108.1888	-55.1888	-0.7757	0.0565	1.1101	-0.1898	-0.1312	0.0757	0.0641
8	350	352.0847	-2.0847	-0.1112	0.9360	17.5583	-0.4255	0.0080	0.0546	-0.4112
9	372	273.4345	98.5655	1.4890	0.1318	1.0041	0.5801	-0.3719	0.4962	-0.1946
10	185	167.2975	17.7025	0.2439	0.0389	1.1621	0.0491	0.0084	0.0081	-0.0162
11	266	182.8171	83.1829	1.1784	0.0427	0.9992	0.2490	-0.0065	0.0882	-0.0883

Output Statistics										
					Hat			DFBETAS		
Obs	Dependent Variable	Predicted Value	Residual	RStudent	Diag H	Cov Ratio	DFFITS	Intercept	F1	F4
12	243	205.0262	37.9738	0.5312	0.0594	1.1563	0.1335	-0.0377	0.0786	-0.0581
13	99	129.0066	-30.0066	-0.4160	0.0464	1.1554	-0.0917	-0.0497	0.0202	0.0338
14	48	172.5641	-124.5641	-1.8210	0.0355	0.8026	-0.3492	-0.0658	-0.0424	0.0508
15	255	211.4022	43.5978	0.6074	0.0484	1.1312	0.1370	-0.0359	0.0734	-0.0230
16	144	148.6520	-4.6520	-0.0640	0.0386	1.1694	-0.0128	-0.0053	0.0011	0.0035
17	53	33.4659	19.5341	0.2850	0.1423	1.2987	0.1161	0.1125	-0.0960	-0.0010
18	110	134.6479	-24.6479	-0.3404	0.0414	1.1573	-0.0708	-0.0457	0.0271	0.0017
19	77	84.3695	-7.3695	-0.1035	0.0765	1.2164	-0.0298	-0.0267	0.0209	0.0005
20	101	164.5000	-63.5000	-0.8882	0.0404	1.0679	-0.1824	-0.0324	-0.0298	0.0694
21	116	108.9383	7.0617	0.0980	0.0555	1.1896	0.0238	0.0192	-0.0137	-0.0006
22	82	100.4049	-18.4049	-0.2567	0.0617	1.1894	-0.0658	-0.0554	0.0410	0.0023
23	52	140.7586	-88.7586	-1.2609	0.0408	0.9747	-0.2602	-0.1644	0.1026	-0.0292
24	105	85.8387	19.1613	0.2690	0.0740	1.2042	0.0760	0.0672	-0.0516	-0.0042
25	283	313.0220	-30.0220	-0.4464	0.1702	1.3236	-0.2021	0.1470	-0.1785	0.0281
26	154	122.5045	31.4955	0.4368	0.0464	1.1531	0.0964	0.0661	-0.0391	-0.0163
27	165	116.8929	48.1071	0.6743	0.0568	1.1298	0.1654	0.1334	-0.1036	0.0330
28	188	169.9855	18.0145	0.2480	0.0371	1.1596	0.0487	0.0085	0.0074	-0.0123
29	125	76.5392	48.4608	0.6909	0.0876	1.1648	0.2141	0.1979	-0.1621	0.0118

ANSWER 4: 8

Question 3.6.

The observation with the highest impact on the intercept is

We use the table above and look at the DFBETAS

ANSWER 3: 5

Question 3.7.

What is the 95% confidence interval for observation no. 3

We use

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

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

We have the predicted value from the table above, 245.8493, further we need the number of observation n=29, and the number of parameters in the model k=3. As $c = h_{ii} = 0.0979$ and from

Variable F4 Entered: R-Square = 0.5750 and C(p) = 2.3998

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	185815	92907	17.59	<.0001
Error	26	137360	5283.08351		
Corrected Total	28	323175			

We have the the MSE=5283.08, leading to s=RMSE=72.6848

$$s\sqrt{c} = 72.6848\sqrt{0.0979} = 22.7423$$

ANSWER 2 : $\mathbf{245.8493 \pm t(26)_{0.975} \times 22.7423}$

Question 3.8.

What is the variance of the model M2, if observation 8 is deleted?

We use

RSTUDENT is a so-called "studentised" residual, i.e.

$$\text{RSTUDENT}_i = \frac{r_i}{\hat{\sigma}(i)\sqrt{1-h_{ii}}},$$

where $\hat{\sigma}(i)^2$ is the estimate of variance corresponding to deletion of the i'th observation.

We isolate and insert

$$\hat{\sigma}(8) = \frac{r_8}{RSTUDENT_8\sqrt{1-h_{88}}} = \frac{-2.0847}{-0.1112\sqrt{1-0.9360}} = 74.1052$$

And the variance is thus 5491.58

ANSWER 4 : 5491.58

Problem 4.

Enclosure C with SAS program and SAS output belongs to this problem. We now consider a subset of municipalities, that are either high or low crime. There might be a link between crime levels and library use. We will test if we can classify these municipalities as a high or low crime municipality, based on the 7 use of library variables (U1 – U7) described in problem 2. (Source http://www.statistikbanken.dk)

Question 4.1.

The number of misclassifications by resubstitution when going from Linear Discriminant Analysis with all variables to Quadratic Discriminant Analysis with all variables is reduced with:

We find the relevant tables in the output

Quadratic Discriminant Analysis

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.KRIMBIB2
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into type			
From type	HighCrim	LowCrime	Total
HighCrim	20 95.24	1 4.76	21 100.00
LowCrime	0 0.00	12 100.00	12 100.00
Total	20 60.61	13 39.39	33 100.00
Priors	0.5	0.5	

Linear Discriminant Analysis

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.KRIMBIB2
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into type			
From type	HighCrim	LowCrime	Total
HighCrim	17 80.95	4 19.05	21 100.00
LowCrime	1 8.33	11 91.67	12 100.00
Total	18 54.55	15 45.45	33 100.00
Priors	0.5	0.5	

We go from 5 misclassification to 1, so

ANSWER 4 : 4

Question 4.2.

The Hotelling T^2 for the hypothesis of same mean in the two groups is

Here we have to be a bit careful, as we have both the table from the QDA and the LDA. We consider

Definition 5.15

Assuming that the hypothesis $H_0 : \Sigma_1 = \cdots = \Sigma_k$ is true, we define *the squared generalized distance from* $\widehat{\mu}_j$ *to population* π_i as

$$D_i^2(\widehat{\mu}_j) = \begin{cases} \left(\widehat{\mu}_i - \widehat{\mu}_j\right)^T \widehat{\Sigma}^{-1} \left(\widehat{\mu}_i - \widehat{\mu}_j\right) & \text{if the priors are equal} \\ \left(\widehat{\mu}_i - \widehat{\mu}_j\right)^T \widehat{\Sigma}^{-1} \left(\widehat{\mu}_i - \widehat{\mu}_j\right) - 2\log p_i & \text{if the priors are not all equal} \end{cases}$$

If the hypothesis is *not* true, we define *the squared generalized distance from* $\widehat{\mu}_j$ *to population* π_i as

$$D_i^2(\widehat{\mu}_j) = \begin{cases} \left(\widehat{\mu}_i - \widehat{\mu}_j\right)^T \widehat{\Sigma}_i^{-1} \left(\widehat{\mu}_i - \widehat{\mu}_j\right) & \text{if the priors are equal} \\ \left(\widehat{\mu}_i - \widehat{\mu}_j\right)^T \widehat{\Sigma}_i^{-1} \left(\widehat{\mu}_i - \widehat{\mu}_j\right) + \log\det\widehat{\Sigma}_i - 2\log p_i & \text{if the priors are not all equal} \end{cases}$$

And see that we should use the one from LDA. We use

Theorem 4.9

We use the same notation as given above. Now, let

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

Then the critical region for a test of H_0 against H_1 at level α is equal to

$$C = \{x_1, \ldots, x_n, y_1, \ldots, y_m | \frac{n+m-p-1}{(n+m-2)p} t^2 > \mathrm{F}(p, n+m-p-1)_{1-\alpha}\}$$

Here t^2 is the observed value of T^2.

Linear Discriminant Analysis

The DISCRIM Procedure

Generalized Squared Distance to type		
From type	HighCrim	LowCrime
HighCrim	0	3.62001
LowCrime	3.62001	0

We further have

The DISCRIM Procedure

Class Level Information					
type	Variable Name	Frequency	Weight	Proportion	Prior Probability
HighCrim	HighCrim	21	21.0000	0.636364	0.500000
LowCrime	LowCrime	12	12.0000	0.363636	0.500000

Thus n=21 and m=12

And the answer is $T^2 = \frac{n \cdot m}{n+m} D^2 = \frac{21 \cdot 12}{21+} 3.62001 = 27.6437$

ANSWER 2 : 27.6437

Question 4.3.

We now test if U1, U2, U3, U5, and U8 contribute to the discrimination between the groups using Linear Discriminant Analysis. The usual test statistic is given by

We use

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 π_1 and π_2, i.e. the hypothesis that $\Delta^2_{(2|1)} = 0$ against all alternatives is

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

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

We find the tables in the previous question, and for the reduced model

Linear Discriminant Analysis - Reduced Model

The DISCRIM Procedure

Generalized Squared Distance to type		
From type	HighCrim	LowCrime
HighCrim	0	1.52415
LowCrime	1.52415	0

We can then insert

ANSWER 3 : $\frac{21+12-8-1}{8-3} \cdot \frac{3.62001-1.52415}{(21+12)(21+12-2)/(21 \cdot 12)+1.52415}$

Problem 5.

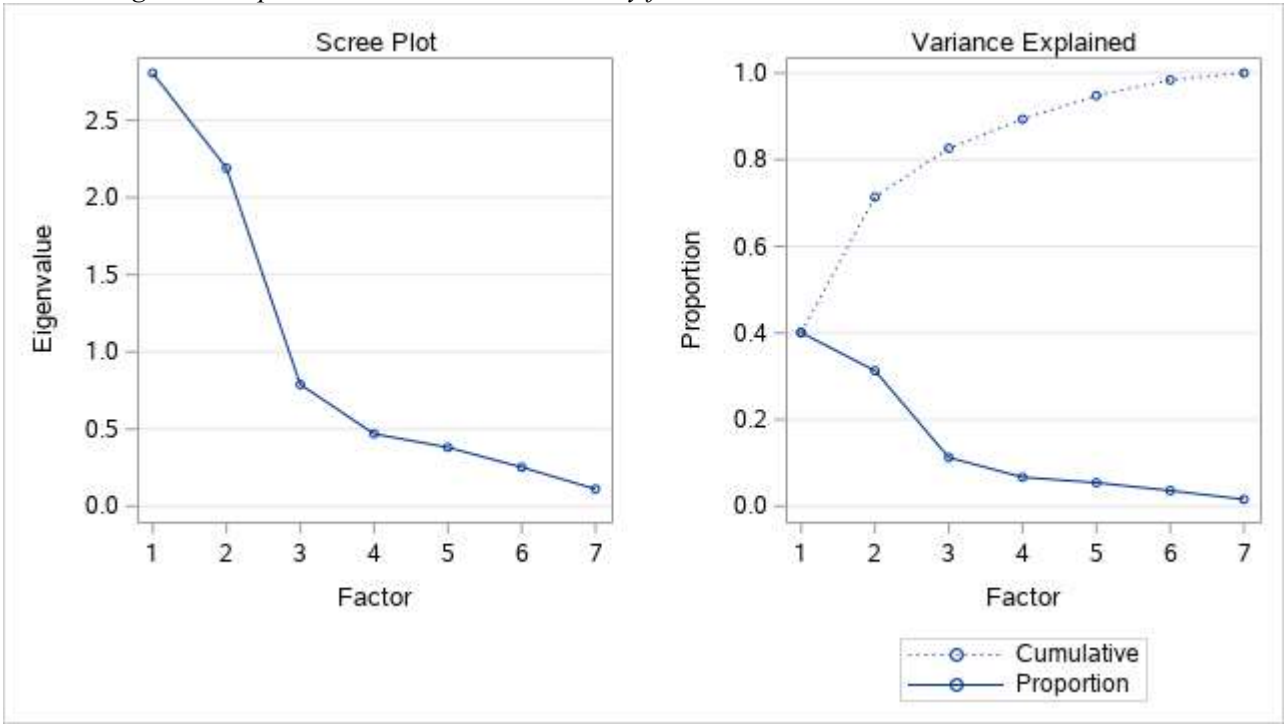
Enclosure D with SAS program and SAS output belongs to this problem. We now consider the library contents among the 98 municipalities in Denmark given in items pr. 1000 capita. We will analyse if there are any patterns or trends by means of a factor analysis. (Source http://www.statistikbanken.dk)

We consider the following variables for library contents

SAS-name	Meaning
B1	Books
B2	Audio books
B3	Music
B4	Live images (movies)
B5	Multi media material
B6	Other material
B7	Electronic resources

Question 5.1.

Considering the screeplot we should retain, how many factors in the model



There is a clear elbow after the first 2. We should thus include 2 factors

ANSWER 4: 2

Question 5.2.

Which of the original variables has the least amount of its variance described by the Factor Analysis

We look at the communalities

3 factors will be retained by the NFACTOR criterion.

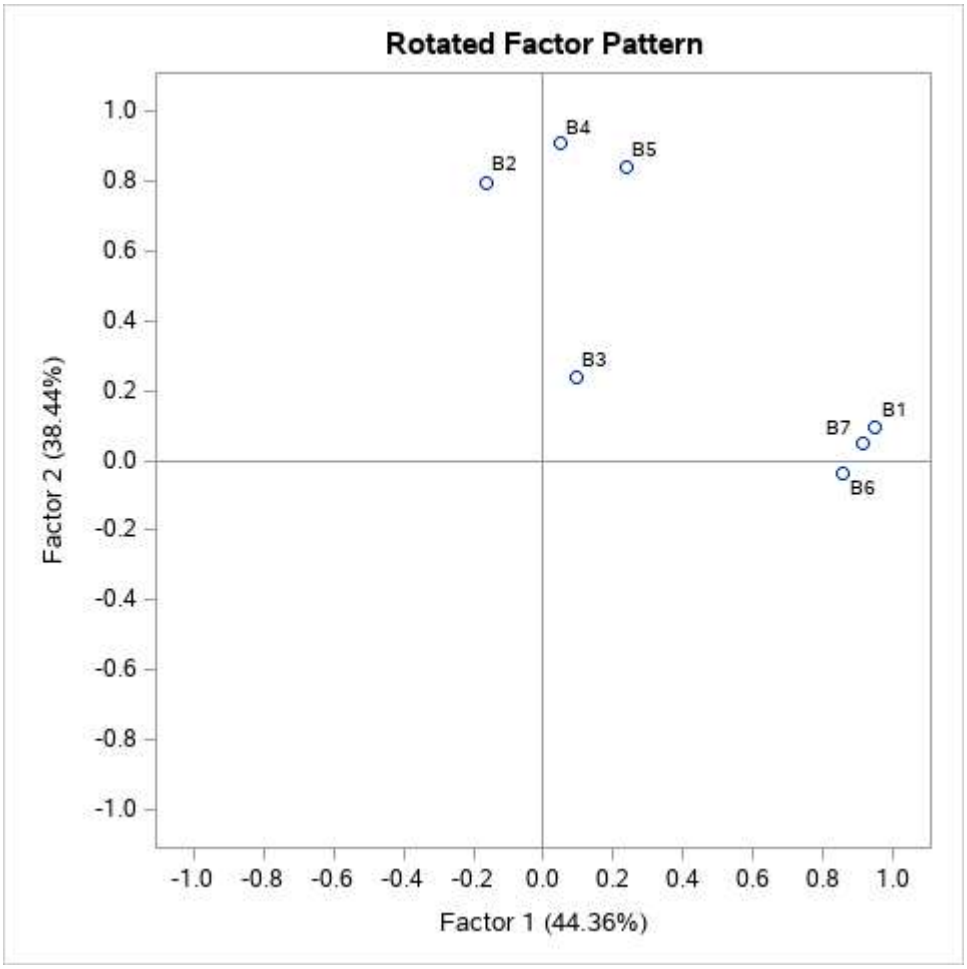
Final Communality Estimates: Total = 5.784919						
B1	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**
0.91159983	0.68926825	0.99103354	0.82877011	0.77181742	0.75630004	0.83613014

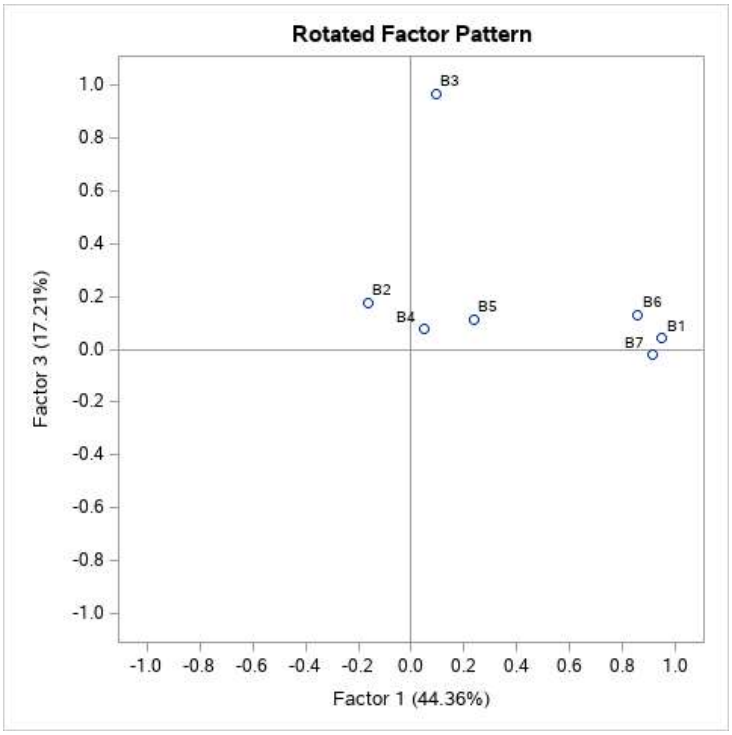
ANSWER 2 : B2

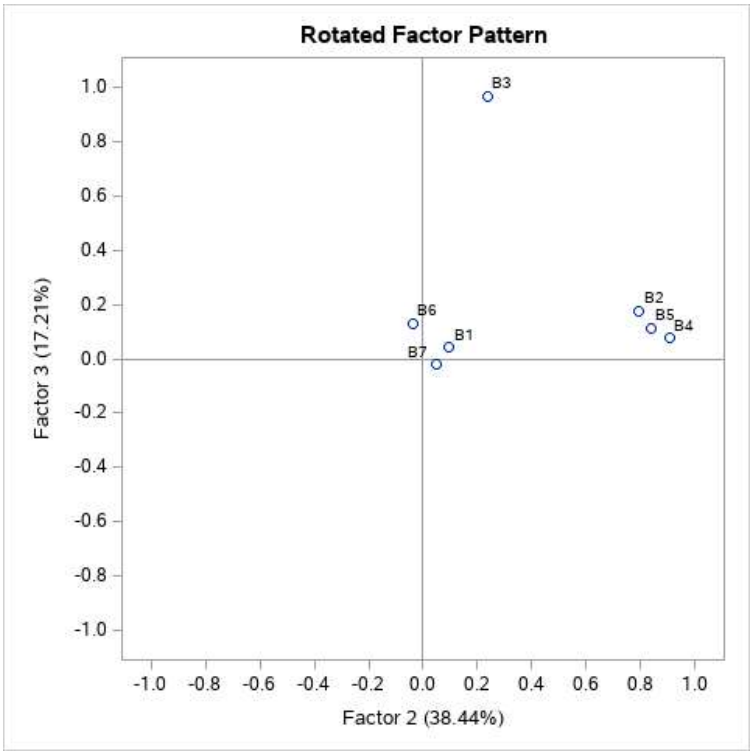
Question 5.3.

Looking at the score plots, we can conclude

First we will investigate the meaning of the different rotated factors.
We inspect the factor patterns.



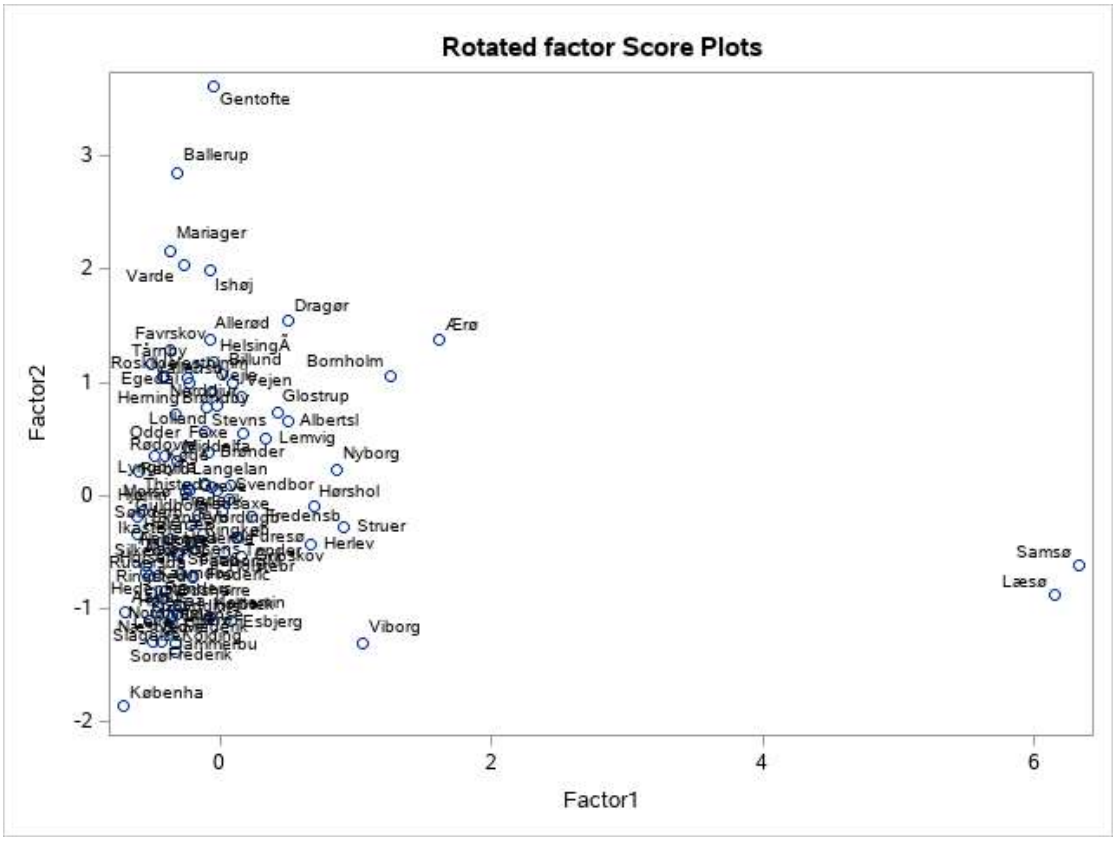


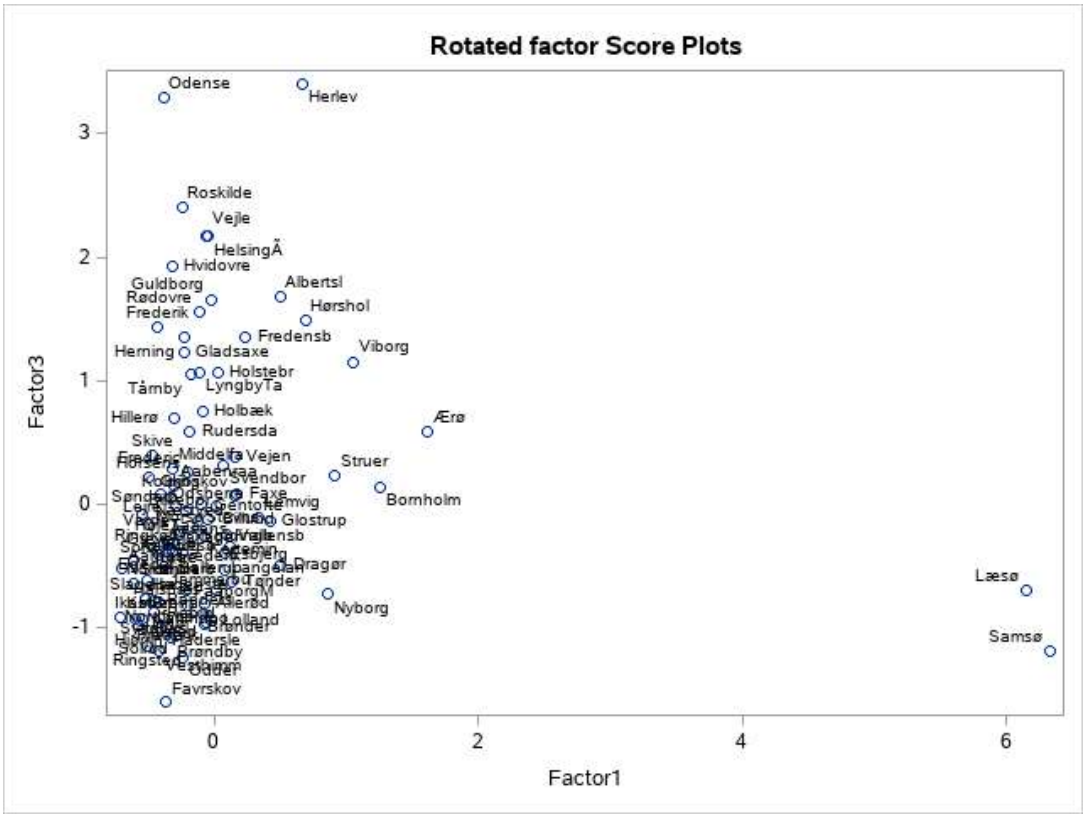


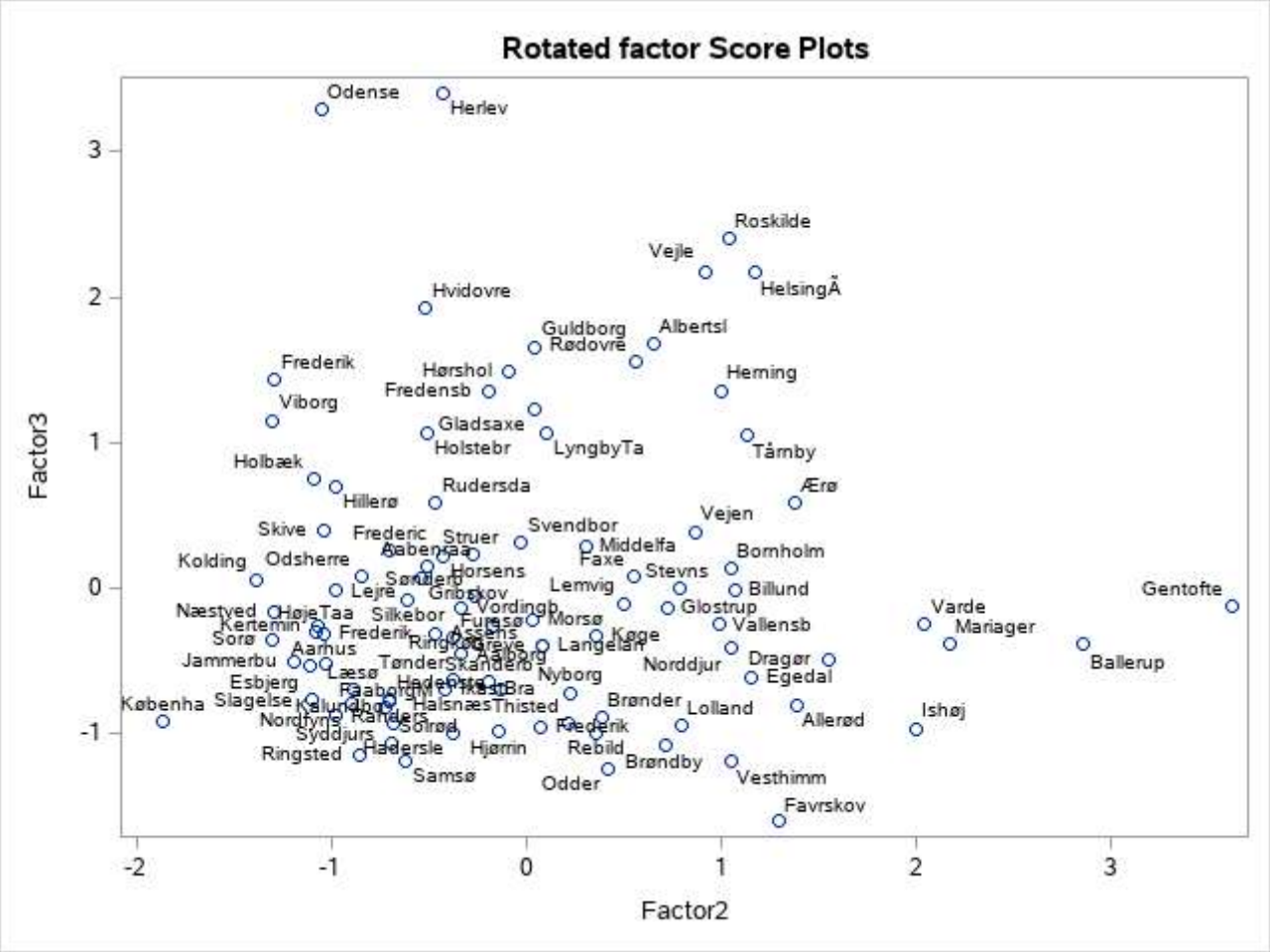
ROTATED FACTOR 1: B1, B6, B7 – Books, Other Material, Electronic Resources

ROTATED FACTOR 2: B2, B4, B5 – Audio Books, Live Images (Movies), Multi Media Material

ROTATED FACTOR 3: B3 - Music


Rotated factor Score Plots
Gentofte
Ballerup
Mariager
Varde
Ishøj
Dragør
Allerød
Ærø
Favrskov
HelsingÃ
Bornholm
Vejen
Glostrup
Albertsl
Lemvig
Nyborg
Hørshol
Struer
Herlev
Samsø
Læsø
Viborg
Frederik
København
Factor2
Factor1
3
2
1
0
-1
-2
0
2
4
6


Rotated factor Score Plots
Odense
Herlev
Roskilde
Vejle
HelsingÃ
Hvidovre
Guldborg
Albertsl
Rødovre
Hørshol
Frederik
Fredensb
Herning
Gladsaxe
Viborg
Holstebr
Tårnby
LyngbyTa
Holbæk
Hillerø
Ærø
Skive
Rudersda
Vejen
Struer
Svendbor
Faxe
Bornholm
Lemvig
Glostrup
Dragør
Nyborg
Læsø
Samsø
Favrskov
Factor3
Factor1
3
2
1
0
-1
0
2
4
6



Læsø and Samsø has a larger collection of books, Other Material and Electronic Resources than average. Københa(vn) has a smaller and Gentofte a larger Audio Books, Live Images (Movies), Multi Media Material collection. Odense and Herlev have a larger Music collection than average, while Favrskov has a smaller

ANSWER 2

Problem 6.

We consider a random variable

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

with mean value and dispersion matrix respectively equal to

$$\mu = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

Furthermore we consider the random variables

$$S = X - Y$$
$$T = Y - Z$$

Question 6.1.

The mean value of the two-dimensional random variable $\begin{bmatrix} S \\ T \end{bmatrix}$ *is*

We use

Remark 1.10 Rules for computing moments of simple functions

$$\mathrm{E}(a+bX) = a + b\,\mathrm{E}(X)$$
$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$$

$$\mathrm{V}(a+bX) = b^2\,\mathrm{V}(X)$$
$$\mathrm{V}(X+Y) = \mathrm{V}(X) + \mathrm{V}(Y) + 2\mathrm{Cov}(X,Y)$$
$$= \mathrm{V}(X) + \mathrm{V}(Y) \quad \text{iff } X, Y \text{ independent}$$

$$\mathrm{Cov}(X,X) = \mathrm{V}(X)$$
$$\mathrm{Cov}(aX,bY) = ab\mathrm{Cov}(X,Y)$$
$$\mathrm{Cov}(X+U,Y) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(U,Y)$$
$$\mathrm{Cov}(X,Y+V) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,V)$$

$$\mathrm{E}(\mathbf{A}+\mathbf{X}) = \mathbf{A} + \mathrm{E}(\mathbf{X})$$
$$\mathrm{E}(\mathbf{AX}) = \mathbf{A}\,\mathrm{E}(\mathbf{X})$$
$$\mathrm{E}(\mathbf{XB}) = \mathrm{E}(\mathbf{X})\mathbf{B}$$
$$\mathrm{E}(\mathbf{X}+\mathbf{Y}) = \mathrm{E}(\mathbf{X}) + \mathrm{E}(\mathbf{Y})$$
$$\mathrm{D}(\mathbf{b}+\mathbf{X}) = \mathrm{D}(\mathbf{X})$$
$$\mathrm{D}(\mathbf{AX}) = \mathbf{A}\,\mathrm{D}(\mathbf{X})\mathbf{A}^T$$
$$\mathrm{D}(\mathbf{X}+\mathbf{Y}) = \mathrm{D}(\mathbf{X}) + \mathrm{D}(\mathbf{Y}) + \mathrm{C}(\mathbf{X},\mathbf{Y}) + \mathrm{C}(\mathbf{Y},\mathbf{X})$$
$$= \mathrm{D}(\mathbf{X}) + \mathrm{D}(\mathbf{Y}) \quad \text{iff } X, Y \text{ independent}$$
$$\mathrm{C}(\mathbf{X},\mathbf{X}) = \mathrm{D}(\mathbf{X})$$
$$\mathrm{C}(\mathbf{X},\mathbf{Y}) = \mathrm{C}(\mathbf{Y},\mathbf{X})^T$$
$$\mathrm{C}(\mathbf{AX},\mathbf{BY}) = \mathbf{A}\,\mathrm{C}(\mathbf{X},\mathbf{Y})\mathbf{B}^T$$
$$\mathrm{C}(\mathbf{X}+\mathbf{U},\mathbf{Y}) = \mathrm{C}(\mathbf{X},\mathbf{Y}) + \mathrm{C}(\mathbf{U},\mathbf{Y})$$
$$\mathrm{C}(\mathbf{X},\mathbf{Y}+\mathbf{V}) = \mathrm{C}(\mathbf{X},\mathbf{Y}) + \mathrm{C}(\mathbf{X},\mathbf{V})$$

$$\begin{bmatrix} S \\ T \end{bmatrix} = \begin{bmatrix} X - Y \\ Y - Z \end{bmatrix} = \begin{bmatrix} 1 - 2 \\ 2 - 3 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

ANSWER 2 : $\begin{bmatrix} \mathbf{-1} \\ \mathbf{-1} \end{bmatrix}$

Question 6.2.

The dispersion matrix for the two-dimensional random variable $\begin{bmatrix} S \\ T \end{bmatrix}$ *is*

We again use Remark 1.10

$$V(S) = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = (1 - \rho) - (\rho - 1) = 2 - 2\rho$$

$$V(T) = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = (1 - \rho) - (\rho - 1) = 2 - 2\rho$$

$$C(S,T) = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = (\rho - \rho^2) - (1 - \rho) = 2\rho - \rho^2 - 1$$

Collecting

$$D\left(\begin{bmatrix} S \\ T \end{bmatrix}\right) = \begin{bmatrix} 2 - 2\rho & 2\rho - \rho^2 - 1 \\ 2\rho - \rho^2 - 1 & 2 - 2\rho \end{bmatrix} = \begin{bmatrix} 2(1-\rho) & -(1-\rho)^2 \\ -(1-\rho)^2 & 2(1-\rho) \end{bmatrix} = (1-\rho)\begin{bmatrix} 2 & \rho - 1 \\ \rho - 1 & 2 \end{bmatrix}$$

ANSWER 2 : $(1-\rho)\begin{bmatrix} 2 & \rho - 1 \\ \rho - 1 & 2 \end{bmatrix}$

Question 6.3.

The covariance between X and S is:

$$C(X,S) = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = 1 - \rho$$

ANSWER 5: $\mathbf{1 - \rho}$

Question 6.4.

The conditional mean $\mathrm{E}(X|Y)$ is

We use

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}^{-}.

$$\mathrm{E}(X|Y) = 1 + \rho \cdot 1 \cdot (y - 2)$$

ANSWER 1: $\boldsymbol{\rho(Y - 2) + 1}$

Question 6.5.

The conditional dispersion matrix $\mathrm{D}\left(\begin{bmatrix} X \\ Z \end{bmatrix} \middle| Y\right)$ *is*

We again use Theorem 1.27

$$\mathrm{D}\left(\begin{bmatrix} X \\ Z \end{bmatrix} \middle| Y\right) = \begin{bmatrix} 1 & \rho^2 \\ \rho^2 & 1 \end{bmatrix} - \begin{bmatrix} \rho \\ \rho \end{bmatrix} \cdot 1 \cdot [\rho \quad \rho] = \begin{bmatrix} 1 & \rho^2 \\ \rho^2 & 1 \end{bmatrix} - \begin{bmatrix} \rho^2 & \rho^2 \\ \rho^2 & \rho^2 \end{bmatrix} = \begin{bmatrix} 1-\rho^2 & 0 \\ 0 & 1-\rho^2 \end{bmatrix}$$

ANSWER 5 : $\begin{bmatrix} \boldsymbol{1-\rho^2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1-\rho^2} \end{bmatrix}$

LAST PAGE:
END OF THE EXAM SET