DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

EXAM ASSIGNMENT

Study Programme and level	**MSc Business Intelligence**						
Term	Summer 23r						
Course name and exam code(s)	Customer Analytics						460202E007
Exam form and duration	WOA: On-site written exam submitted digitally in WISEflow, use of the internet NOT allowed during the exam, own PC required.						4 hours
Date and time	11 August 2023						09.00-13.00
Supplementary material/aids	All	X	No		Specified		
Hand-in of hand-written material allowed	Yes		No	X			
Hand-in of extra material (appendix) in WISEflow allowed	Yes		No	X			
Anonymous exam	Yes	X	No		Comments: Please do **not** write your name or student ID number anywhere. Use your flow-id number (find this on the cover sheet).		
Other relevant information	**Avoid being suspected of exam cheating** Remember to state references and use quotation marks, if you copy text from other sources or re-use parts of a previously submitted exam paper (plagiarism and self-plagiarism). Students must answer the exam assignment **individually**. All submitted exam papers are checked for plagiarism, so cheating and collaboration between students will be detected. A dataset is uploaded to WISEflow as appendix.						
Number of pages (incl. front page)	5 pages						

PART A. Lecturer: Ana Alina Tudoran (50 pts in total)

Exercise 1

Consider a survey scale that assesses customers' *product involvement*, using the survey shown in Table 1. This survey scale reflects a model in which product involvement (PIE) is a hierarchical construct comprising three factors: *general* involvement with a product category (Gnr), involvement with the *choices and features* of the product (Ftr), and involvement with the category in terms of personal *image* (Img).

Table 1. The hierarchical product involvement (PIES) scale showing the subscales (factors) and items. The survey is given for a specific product category by letting the informant fill in the blanks with a descriptive phrase such as "digital cameras" or "diet soda" (Chapman and Feit, 2015).

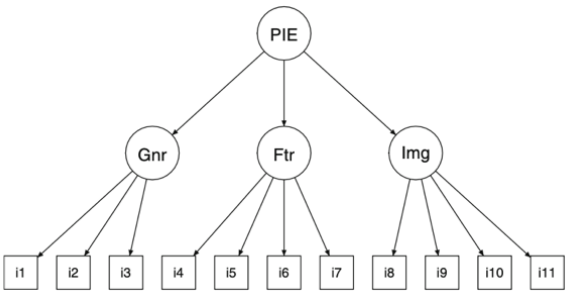
Item	Text	Reversed?
General scale		
i1	____ are not very important to me.	Yes
i2	I never think about ____.	Yes
i3	I am very interested in ____.	
Feature scale		
i4	In choosing a ____ I would look for some specific features or options.	
i5	If I chose a new ____ I would investigate the available choices in depth.	
i6	Some ____ are clearly better than others.	
i7	If I were choosing a ____, I would wish to learn about the available options in detail.	
Image scale		
i8	When people see someone's ____, they form an opinion of that person.	
i9	A ____ expresses a lot about the person who owns it.	
i10	You can learn a lot about a person by seeing the person's ____.	
i11	It is important to choose a ____ that matches one's image.	

1. Below, we present the results of a factor analysis model. Interpret the output. (5 pts)

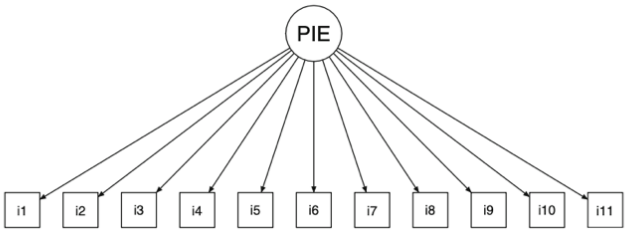
```
> factanal(piesSimData, factors=3)
...
Loadings:
    Factor1 Factor2 Factor3
i1  0.138   0.119   0.675
i2                  0.614
i3  0.277   0.362   0.476
i4  0.151   0.608
i5  0.126   0.715   0.102
i6          0.519
i7  0.133   0.678   0.154
i8  0.665   0.137   0.128
i9  0.706   0.138   0.130
i10 0.655   0.117   0.145
i11 0.632   0.126
...
```

2. Consider the three CFA models represented below with their path diagram. The output of the models fit, and comparison of the three models is presented afterwards. Make a thorough interpretation of the output and conclude about the best model among the alternatives. (15 pts)

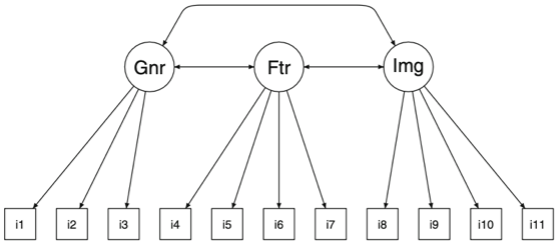
pies.fit



pies.fit.NH1



pies.fit.NH3




```
> library(semTools)
> compareFit(pies.fit.NH1, pies.fit.NH3, pies.fit)
################## Nested Model Comparison #########################
                              chi df      p delta.cfi
pies.fit - pies.fit.NH3      222.43  3  <.001    0.0222
pies.fit.NH3 - pies.fit.NH1 2774.50  0  <.001    0.2812

################### Fit Indices Summaries ##########################
               chisq df pvalue   cfi   tli         aic         bic rmsea  srmr
pies.fit.NH1 3284.581 44   .000† .672  .589  108812.709  108948.860  .143  .102
pies.fit.NH3  510.078 44   .000† .953  .941  106038.205  106174.356  .054  .078
pies.fit      287.649 41   .000† .975† .966† 105821.776† 105976.494† .041† .030†
```

Exercise 2

Imagine you are conducting a survey to gather opinions about a new product. You are interested in finding out the percentage of people that are likely to purchase the product, but you are concerned that people may not be completely honest in their responses. What can you do to ensure that you are getting accurate information? (10 pts)

Exercise 3

Suppose you are conducting an A/B test to determine whether a new webpage design will lead to a higher conversion rate compared to the existing design. Last month, the website received 5,000 visitors out of which there were 250 conversions. The manager will only change the website for a minimum detectable effect of 10%. Assuming a two-tailed hypothesis test with a significance level of 0.05 and a statistical power of 0.95, calculate the total sample size and per group. [Note: you must show your calculations. A number without explicit calculations will not be assessed]. (10 pts)

Exercise 4

In relation to the Bayesian Network Applications, what is the advantage of using probabilistic cluster learning versus a simple probabilistic learning algorithm for building a system to make product recommendations? (10 pts)

PART B. Lecturer: Morten Berg Jensen (35 pts in total)

The influencer model is described on pages 70-73 in your curriculum textbook (Hair Jr et al. 2021[1]). An online survey was carried out in 2023, building on this model. Compared to the description of the data collection in the textbook, the respondents in the 2023 online survey only saw the "real" influencer and used a different scale (rate how much you agree with each statement on a scale from 0 = strongly disagree to 100 = strongly agree). However, those responsible for the online survey forgot to include the "sic_7" question when they developed the survey. Thus, only "sic_1" to "sic_6" are available (and they used "global_sic" as the variable name for the global single item for redundancy analysis). See Tables 3.9 and 3.10 in the textbook for an overview of the variables. In order to solve this assignment, you must use the dataset in the file called "*PLS_data_exam.csv*". It has previously been established, as part of the assessment of the reflective measurement model, that "pl_4" should be removed from the measurement model. Thus, the "PL" construct is measured via "pl_1", "pl_2", and "pl_3".

1. Specify and run the influencer model. Next, evaluate the formative measurement model. Be specific: report your implementation, relevant results, and comment on the output. (15 pts)

2. A cluster analysis of the respondents based on the answers to variables "sic_1" to "sic_6" using hierarchical cluster analysis was performed below. Interpret the output. (10 pts)

```
> influencer_data <- read_csv(file = "PLS_data_exam.csv")
Rows: 504 Columns: 21

-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (21): sic_1, sic_2, sic_3, sic_4, sic_5, sic_6, global_sic, pl_1, pl_2, pl_3, pl_4, pq_1,..

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
> xlist <- influencer_data[,c("sic_1","sic_2","sic_3","sic_4","sic_5","sic_6")]
> nobs <- nrow(xlist)
> xlist$id <- seq(1,nobs)
> apply(xlist,2,sd)
    sic_1     sic_2     sic_3     sic_4     sic_5     sic_6        id
 15.96949  16.28558  16.03033  16.47900  15.85859  17.01446 145.63653
> cor(xlist)
            sic_1       sic_2        sic_3       sic_4        sic_5       sic_6           id
sic_1  1.00000000  0.24761659  0.324229426 -0.22637651 -0.246893484  0.32918811 -0.035469150
sic_2  0.24761659  1.00000000  0.222604830 -0.19327634 -0.234059313  0.22745780  0.058698929
sic_3  0.32422943  0.22260483  1.000000000 -0.25949360 -0.309976860  0.27820239  0.008506736
sic_4 -0.22637651 -0.19327634 -0.259493605  1.00000000  0.251438173 -0.25891471 -0.061151590
sic_5 -0.24689348 -0.23405931 -0.309976860  0.25143817  1.000000000 -0.28476755 -0.008833856
sic_6  0.32918811  0.22745780  0.278202394 -0.25891471 -0.284767548  1.00000000  0.027143430
id    -0.03546915  0.05869893  0.008506736 -0.06115159 -0.008833856  0.02714343  1.000000000
> dev <- t(t(xlist)-apply(xlist,2,mean))
> dev2 <- dev^2
> sumdev2 <- rowSums(dev2)
> tail(sort(sqrt(sumdev2)),n=30)
 [1] 239.5438 240.6906 240.8897 242.2757 242.6424 242.8164 244.0502 244.5330 244.8134 245.3713
[11] 245.6496 246.4568 246.9499 247.3092 247.8333 248.1604 248.4402 249.2875 249.3642 251.0146
[21] 251.0894 252.7583 252.8877 253.3598 254.1139 254.3013 254.4710 254.9024 256.6957 257.3319
> dist <- dist(xlist,method="euclidean")
> dist2 <- dist^2
> H.fit <- hclust(dist2,method="ward.D")
> denominator <- cumsum(H.fit[[2]])
> length(denominator) <- nobs-2
> denominator <- c(1,denominator)
> pct <- H.fit[[2]]/denominator
> tail(pct,n=10)
 [1] 0.03872693 0.04377437 0.04764945 0.06020092 0.07279324 0.11993677 0.28790181 0.28117180
 [9] 0.74050821 2.02980372
```

[1] Hair Jr, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook.

3. How would you determine the robustness of the results? (5 pts)

4. Suggest changes/extensions to the above cluster analysis to make it possible to relate the outcome of the cluster analysis (cluster membership) to demographic variables (assuming these variables were collected as well). (5 pts)

PART C. Lecturer: Surabhi Verma (15 pts in total)

1. A grocery store chain is experiencing a decline in sales and wants to improve its offerings to customers by implementing a recommendation system. Which of the following would be a better recommendation system and why? (7.5 pts)

   a) User-user collaborative filtering
   b) Item-item collaborative filtering
   c) User-item collaborative filtering
   d) Content-based recommendation

2. Discuss a scenario in which content-based recommendations will not perform as well as collaborative filtering. (7.5 pts)