1. The Excel data set Pollution.data.xlsx contains measurements on 9 pollutants made daily at a

large airport. Different pre-defined groups in the data are included as the variable Cluster.

(a) Construct a PCA biplot of the Pollution data without showing alpha bags.

(b) Repeat (a) but instead of sample labels show the different groups (clusters) as 90% bags.

(c) Repeat (a) but give an optimal two-dimensional display of the correlations between the

variables.

(d) Give a detailed interpretation of the plots constructed in (a), (b), and (c).

(e) Construct a CVA biplot of the Pollution data with 90% bags added. Interpret and discuss

the use of this biplot.

2. Consider again the Pollution data set introduced in Question 1

(a) Compute the following dissimilarity/distance matrices for Cluster 4:

i) Euclidean distances.

ii) Canberra distances (Hint: Study the help file of function dist() ).

(b) Perform separate classical scalings on both of the dissimilarity/distance matrices

computed in (a).

(c) Interpret and discuss the classical scalings in (b). Include in your answer a comparison

with the biplots constructed in Question 1.

3. The Excel data set ‘Brands.data.xlsx’ contains data of 10 features associated with 23 brands of a certain product.

(a) Obtain a Euclidean distance matrix of the Brands.

(b) Give a brief description of the use of a stress function associated with metric least

squares scaling.

(c) Use the function optim() to perform a metric least squares scaling on the Brands data

set.

(d) Discuss the output (graphical and statistical) of the analysis in (c).

(e) Transform the Brands.data into the R object Brands.data.ord where each feature is an

ordered categorical variable.

(f) Inspect the help file of function daisy() in the R package cluster. Use daisy() to

construct an ordinal dissimilarity matrix for the different brands considered in

Brands.data.ord based on Gower's coefficient.

(g) Perform a nonmetric MDS on the Brand dataset by uitilising the R package smacof.

Give a detailed interpretation of your nonmetric MDS. Refer also to your findings in (d).

(h) Repeat the nonmetric MDS graphical representation of the Brands in (g) but with the

values of Feature 5 replacing the different Brands. Interpret your graph.

4. Perform a detailed Procrustes analysis on the metric and non-metric configurations obtained in Question 3 by using as target configuration the metric least squares MDS configuration.

Discuss in detail the results of the Procrustes analysis. (20)

5. Ninety-two subjects took part in an opinion survey consisting of five questions. The subjects

were recruited in three different districts. The data are given in the file

OpinionSurvey.data.xlsx.

(a) Ensure that the answer to each question is a categorical variable.

(b) Construct an MCA biplot on the associated indicator matrix of the questions. Do not

colour the sample points but label them using their IDs as labels. Represent each

categorical variable in a different colour. Add a suitable legend to the MCA biplot.

(c) Repeat (b) but this time colour all CLPs in the same colour while distinguishing the

samples from the different districts using colour coding.

(d) Discuss in detail the conclusions to be drawn from the MCA biplots in (b) and (c) and

the associated optimal scores. (20)

6. Six features of products produced by three manufactures were scored as ordered categorical

variables and saved in Excel format as Manufacture.data.xlsx. Feature 6 was scored in

reversed order as the other features.

(a) Ensure that all features are ordered categorical variables in your imported data set.

(b) Use function CATPCAbipl as given in package UBbipl40 to carry out a Categorical

Principal Components Analysis on the Manufacture data. (Hint: due to many similar

response patterns argument jitter.bags should be set to TRUE when requesting alpha

bags)

(c) Give a brief discussion of the aim(s) of a Categorical Principal Components Analysis

followed by a detailed discussion/interpretation of the result of the analysis in (b). (20)

7. A research group aimed to construct a questionnaire for screening candidates for having an

underlying disease. They started with 50 binary questions (symptoms) given to a carefully

selected sample consisting of 5000 subjects (respondents or persons). The data are given in the file symptoms.data.xlsx where a 0 denotes no indication of the underlying disease and a 1 is an indication that the underlying disease is present.

(a) First, inspect the contents of R package CTT and then perform an Item Analysis on the

Symptoms data.

(b) Obtain the ‘person’ scores and transform these scores to a scale having a mean of 100

and a standard deviation of 15.

(c) Represent the transformed person scores in the form of a unidimensional scaling graph.

Explain how to use this graph in practice.

(d) Construct a unidimensional scale (in table and graph form) for the items and explain

how to interpret the scale.

(e) Motivate in detail which 20 items from the original 50 items would you recommend for

a final screening test. (15)

8. Obtain a new data set (new.data) consisting of the 20 recommended items in Question 7(e) for a systematic sample of persons chosen by taking the first and thereafter every 50th person in the Symptoms data set.

(a) Before attempting (b) study Unit 5 Slides 15-24. If you have not already installed

package ltm then install it from (https://cran.r-project.org › web › packages › ltm ).

Next, work through the R script file: IRT.Examples.R.

(b) Use item response theory and the data set, new.data created above to

(i) fit a Rasch model to the data;

(ii) find disease scores and express them in unidimensional scaling format;

(iii) obtain item characteristic curves;

(iv) obtain item information curves and

(v) write a report on your findings