Professor: Dr. Alex van Venrooij
Faculty of Social and Behavioural Sciences
Course: Measuring Meaning in Mixed Methods
Winter semester 2022/2023

UNIVERSITEIT VAN AMSTERDAM

Measuring Meaning in Mixed Methods

Individual Assignment 2

Helge Moes
11348801

Date: 03.02.2023

1) Multidimensional scaling and cluster analysis are widely used techniques in the measurement of meaning. The Jang and Barnett data can also be analyzed using both techniques (You can find the file in the data folder of session 3.1). It contains a two-mode data set with words in the rows and companies in the columns (companies labeled 1-18 are US companies and 19-35 are Japanese companies).

a) Apply cluster analysis to the Jang and Barnett data. Discuss the steps that you need to take to apply cluster analysis of the relations between companies. Describe the results and whether you can find evidence for cultural differences. Add the R script in an appendix. (200 words)

The result of cluster analysis portrays the groupings among all nodes that are represented by their relative closeness and height (Jang & Barnett, 1994). In order to come to the concluding cluster analysis, the data has to be pre-processed, this entails making a matrix and transposing the matrix to make a one mode out of a two mode matrix that only considers companies. Furthermore, a distance matrix with a euclidean method is run to observe the distance between the companies with the 'hclust()' function. The method of 'ward.D' allows for a cluster analysis that explores the minimum variances.

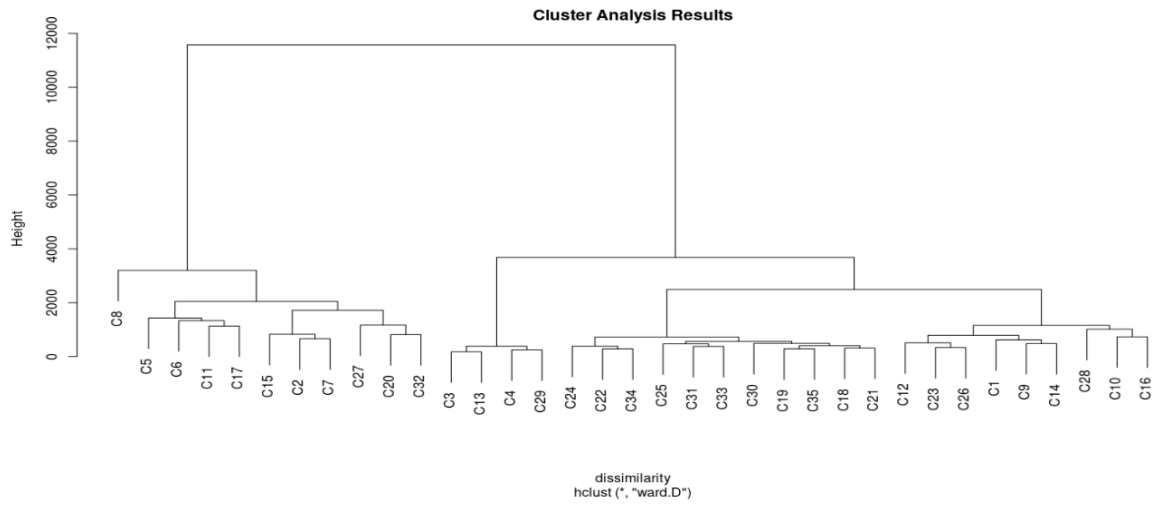

Figure 1: cluster analysis

Based on the cluster analysis, the one-mode dataset shows that American companies (1-18) and Japanese companies (19-35) are clustered separately and in some instances overlapping. The cultures of companies differed and were generally clustered into two groups: Japanese and American. It is apparent that there is a cultural difference between Japanese and American companies. Nevertheless, the results showed that some companies

that had the same business type were more tightly clustered, which shows an overlap between the cultures. These results indicate that the network structure of companies was influenced by the differences in national cultures as well as the companies' business types.

b) Apply multidimensional scaling to the Jang and Barnett data. Discuss the steps that you need to take to apply an MDS analysis of the relations between companies. Describe the results and whether you can find evidence for cultural differences. Add the R script in an appendix. (199 words)

A multidimensional scaling analysis reads in a square matrix of similarities or differences and produces a transformation of the data that locates the objects in a common space and emphasizes the similarities in the input matrix that are transformed into Euclidean distances (Mohr, 1998). In order to generate a multidimensional scaling of the 'Jang_Barnett' data, the preprocessing of the cluster analysis was taken into account. However, the 'cmdscale()' function is implemented to perform a multidimensional scaling. Two separate methods of multidimensional scaling were implemented to observe differences in the results. A general plot was made for exploration of the distribution and an ordinal scaling was made with the package 'smacof'.

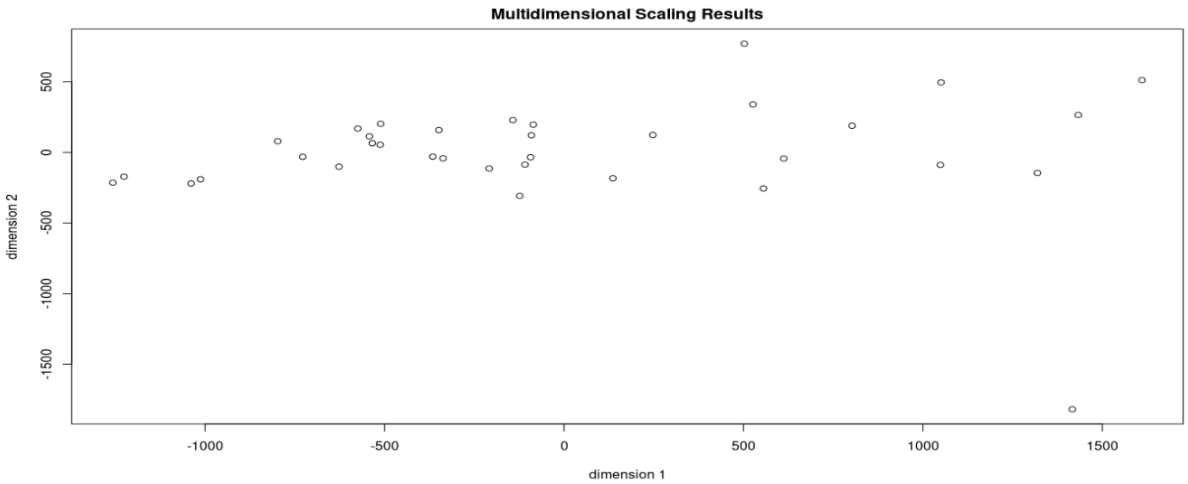

Figure 2: general plot of multidimensional scaling

The general plot shows a linear relation between the companies cultures, since the nodes seem to be distributed equally through the plot. Figure 2 portrays a cluster where companies that are considered to be similar in culture are shorter in distance. Consequently, one may conclude that, based on this multidimensional scaling, although the nationalities of the companies are different, the culture in the company are to some degree similar. Some

companies overlap, this might be due to cultural differences. Thus, there is a correlation between nationality and company culture.

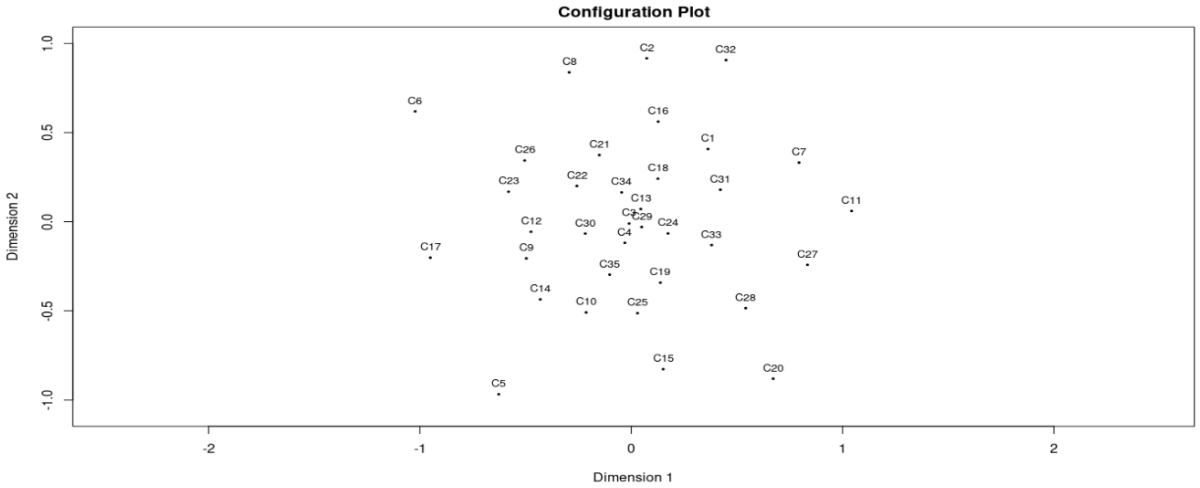

Figure 3: ordinal plot of the multidimensional scaling

2) In this course we have read various studies that have used formal analytical techniques, such as network analysis, for measuring meaning in archival texts.

a. Identify an archival data source (not yet discussed in the required readings or the lecture) that you could potentially use in a measuring meaning project of your own. Give a short description of the archival text(s) you would want to analyze and the reason why you choose it. (299 words)

For my own project, I refer to the Social Evolution experiment (Madan, 2008) to closely track the everyday life of an undergraduate dormitory at MIT university by examining personal archives as an archival data source to track health behavior and negative symptoms. Rather than personal papers, photographs and other documents that characterize an individual's life, I would conduct a measuring-meaning method on surveys and interviews, since most personal archives may contain valuable health information.

The Social Evolution experiment covers multiple files that contain the locations, proximities, and phone calls from October 2008 to May 2009 among others. The particular research was conducted with mobile phones (Madan, 2008), so that social scientists can validate their models against the spatio-temporal patterns and behavior-network co-evolution. This experiment was designed to examine the adoption of political opinions, exercise, diet, eating habits, depression and stress, interpersonal relationships, and privacy of students. More specifically, the research considered a population of 80 students in different phases of their

academic career. Madan and colleagues (2008) explained health symptoms by emphasizing how subjects moved, talked to others, when they talked and the weight of a subject by measuring changes in weight on a weekly basis.

My research would build on the dataset of this research. I shall expand on the research by conducting a measuring-meaning research at the current MIT dormitory and assess whether the health standards of the students are satisfactory. By surveying students at the same dormitory, a comparison of the findings with the results of 2008 is possible. It would grant new insights on the mental and health state of students in the dorm.

Hence, the datasets 'Health.csv' and 'FluSymptoms.csv' are used to research the following question: *"How do negative symptoms and health behavior influence each other among students at the MIT dormitory?"*

b. Describe how you would analyze this archival data, by moving through the three steps of structural analysis (i.e. identifying the units, defining relations of similarities or difference, and how to find the structure). Provide details on the most important choices that you would have to make in the process. (600 words)

Based on the steps of structural analysis by John Mohr (1998), the research consists of the following elements: the units, relations of similarities or differences and the structure of the data. The current data gathered on the students lack in certain aspects, which will be addressed in the follow up study. The Research project concerns the well-being of students in the dormitory.

The first csv that is implemented is 'Health.csv,' that consists of the units that are participants id (user.id). Other variables it takes into account are: the monthly self-reports of height (current_height), weight (current_weight), number of fruits per day (fruits_per_day), salads per week (salads_per_week), considered healthy meals (health_diet) and sports per week (sports_per_week), the days per week with 20 minutes aerobic exercise (aerobic_per_week) and smoking habit (current_smoking).

The second file is 'FluSymptoms.csv,' which has the daily self-reports of the students (user_id) as the units. Moreover, the symptoms comprises of sore throat / coughing (sore.throat.cough), runny nose / congestion / sneezing (runnynose.congestion.sneezing), fever (fever), nausea/vomiting/diarrhea (naussea.vomiting.diarrhea), sad / depressed

(sad.depressed) and stressed (often.stressed). This dataset shall be regarded as negative symptoms the students experience during their studies.

The preprocessing of the data is a fundamental process of this research. This consists of separate steps for each dataset. For the ‘Health.csv’ the values of the ‘current_smoking’ column are represented as labels and numbers. Therefore, these have to be separated. Moreover, the outliers need to be defined in the ‘current_height’ column. On the other hand, for the ‘FluSymptoms.csv’ the duplicate rows need to be removed, the two instances of specific users at a specific time slot occurred twice with conflicting values are to be removed and each observation has to have its own row so the symptoms can be transformed into rows to reduce zero values. Finally, the datasets are joined and transposed based on the primary key of ‘user.id’ (unit) in order to make a two-mode matrix of the several symptoms and health factors.

The relation between the datasets can be determined as an attribute similarity (Mohr, 1998, p. 354). As stated by John Mohr (1998, p. 354): “Items within a cultural system can also be compared on the basis of the sets of selected attributes that they share”. The symptoms and health issues that are experienced by the students have a causal effect. Since a deteriorating health state is determined by the negative symptoms one is experiencing. In this case, students are compared to one another on the basis of the similarity of their claims that they have experienced through symptoms, which leads to observation of health issues.

For the structure of the findings, a multidimensional scaling and cluster are chosen, since multidimensional scaling enables to reduce the dimension to two dimensions to be able to compare. In this case the two dimensions that are compared are the symptoms and the health factors. The multidimensional space thus represents a series of objects (students) coinciding with symptoms and health factors that if the two items are similar to one another in the matrix, then they are located near each other (Mohr, 1998).

In the case that they are dissimilar, the space between the nodes is increased. Since the data is analyzed in two dimensions, the items convey a relational structure that is embedded. Additionally, the clustering method enables to connect items to clusters that are similar in a pairwise sense, to combine the symptoms and health issues to a student. This method is implemented to designate the subregions within the space and contributes to the multi dimensional scaling findings.

Bibliography

Jang, H.-Y. (1997). Cultural differences in an interorganizational network: Shared public relations firms among Japanese and American companies. *Public Relations Review*, *23*(4), 327–341. https://doi.org/10.1016/S0363-8111(97)90049-7

Madan, A., M. Cebrian, S. Moturu, K. Farrahi, A. Pentland. (2008). "Sensing the 'Health State' of a Community." Pervasive Comupting, Vol. 11, No. 4, pp. 36-45. Retrieved January 31, 2023, from http://realitycommons.media.mit.edu/socialevolution.html

Mohr, J. W. (1998). Measuring Meaning Structures. *Annual Review of Sociology*, *24*(1), 345–370. https://doi.org/10.1146/annurev.soc.24.1.345

Appendix - R Script

```
---
title: "Assignment2"
output: Helge Moes
date: "2023-01-30"
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
```

Jang & Barnett (1994) examined the impact of national culture on organizational culture by analyzing messages directed to external audiences. For a total of 35 Japanese and American businesses, annual reports for 1992 were analyzed by determining the most frequently used words in all 35 reports. Then, the frequency of each word for each company was determined.

There are different packages in R that can do MDS. But they use different algorithms. We use the SmacofSym package. It can both do metric (type=interval) and non-metric MDS (type = ordinal) and gives the stress values. Here we choose metric multidimensional scaling.

To illustrate the use of multidimensional scaling and cluster analysis we first work with a distance matrix that measures the distances between.

```{r}
#Installing packages smacof for multidimensional scaling plot
#install.packages("smacof")
library(smacof)

#read in matrix csv-file
Jang_Barnett <- read.csv2("Data/Jang_Barnett.csv", header=TRUE,row.names = 1)
class(Jang_Barnett)
Jang_Barnett

#Pre-processing
Jang_Barnett.C1 <- as.matrix(Jang_Barnett) #Convert to matrix
Jang_Barnett.C2 <- t(Jang_Barnett.C1)%*%Jang_Barnett.C1 #Transpose matrix
distance_matrix_JB <- dist(Jang_Barnett.C2, method = "euclidean") #Create a distance matrix

class(Jang_Barnett.C2)
```

```
#A one-mode matrix of the American and Japanese companies
Jang_Barnett.C2

#Calculate dissimilarity matrix
#dissimilarity <- dist(Jang_Barnett.C2, method = "euclidean")

#Perform cluster analysis
fit.cluster <- hclust(dissimilarity, method = "ward.D")

#Plot results of Cluster Analysis
plot(fit.cluster, main = "Cluster Analysis Results") + theme( axis.text.x = element_text(angle =
    45, hjust = 1))

#Perform multidimensional scaling
fit.multi <- cmdscale(dissimilarity, k = 2)

#Plot results of Multidimensional Scaling
plot(fit.multi[, 1], fit.multi[, 2], main = "Multidimensional Scaling Results", xlab = "dimension
    1", ylab = "dimension 2")

#Plot a Multidimensional Scaling with an ordinal type through smacof in order to retrieve a
    cluster between the different companies
MDS_Companies <- smacofSym(Jang_Barnett.C2, type = "ordinal", ndim=2)
MDS_Companies$stress
plot(MDS_Companies)
```