**Technische Universiteit Delft**
**Faculteit Elektrotechniek, Wiskunde en Informatica**
**Delft Institute of Applied Mathematics**

# Het onderzoeken van voorkeuren van mensen voor bepaalde voedingsmiddelen gebruikmakend van Thurstone's Pairwise Comparison Model

## (Engelse titel: Thurstone's Pairwise Comparison Model to explore preferences of humans for different food items)

Verslag ten behoeve van het
Delft Institute of Applied Mathematics
als onderdeel ter verkrijging

van de graad van

**BACHELOR OF SCIENCE**
**in**
**TECHNISCHE WISKUNDE**

**door**

**LAURA MIDDELDORP**

**Delft, Nederland**
**Juli 2018**

**BSc verslag TECHNISCHE WISKUNDE**

# "Het onderzoeken van voorkeuren van mensen voor bepaalde voedingsmiddelen gebruikmakend van Thurstone's Pairwise Comparison Model"

**(Engelse titel: "Thurstone's Pairwise Comparison Model to explore preferences of humans for different food items")**


LAURA MIDDELDORP

**Technische Universiteit Delft**


**Begeleider**

Dr. D. Kurowicka

**Overige commissieleden**

Dr.ir. G.F. Nane

Dr. J.G. Spandaw

Juli, 2018

Delft

**Abstract**

The method of pairwise comparisons is a method that is commonly used in psychology to model human preferences for a set of objects. In a pairwise comparison experiment, the preferences of a sample are obtained by comparing the objects in pairs. Each individual in the sample judges every possible pair of objects and expresses which of the two object he prefers. Thurstone's Pairwise Comparison Model can then be used on the gathered data in order to obtain rankings of the objects based on the preferences of the sample. This thesis investigates several properties of Thurstone's Pairwise Comparison Model and how it can be used to model human food preferences. In recent years, people have become more aware of the importance of eating healthy food. It may, however, be questioned whether healthy food is generally preferred by the population. This can be examined by means of Thurstone's Pairwise Comparison Model. In order to examine how the model performs, a data study has been carried out using two sets of food products. One set consisted of general food products, while the other set was comprised of snack products. Both sets of objects consisted of both healthy and unhealthy food products. The influence of an introduction text, emphasizing the importance of healthy food, on the preferences of the individuals has been examined as well. For this, each set of food products was judged by two groups, each group receiving a different introduction text before starting the pairwise comparison experiment. The results of the data study yielded that Thurstone's Pairwise Comparison Model is more suitable for modelling preferences of humans for general food products. Furthermore, the introduction text turned out to be not of influence on the preferences of participants.

## Acknowledgements

Foremost, I would like to thank my loving parents who were always there for me whenever I faced difficulties during my research. Their encouragement and interest were of great help throughout the course of this research.

I would like to express my special gratitude to my dear friend Jenneke, who helped me through this research by introducing me into the world of combinatorics. Furthermore, she was always willing to lend an ear whenever difficulties arose.

To my supervisor, Dr. D. Kurowicka, who guided me through this research and made sure that I could always come by her office whenever I had questions.

And finally, to Dr.ir. G.F. Nane, who thought along with me in times when I encountered peculiar outcomes in my research.

# Contents

# 1 Introduction

For almost a century, researchers in the field of psychology are trying to model human preferences in order to obtain information about how individuals perceive a set of different objects. These objects can be physical, for example, different brands of orange juice or cars, but may also be intangible like statements or loudness of noises. Different methods to model human preferences have been developed over the years, one of the oldest being the method of pairwise comparisons.

There exist a lot of variations on the method of pairwise comparisons, the most common one being the multiple judgment paired comparisons [7]. In this procedure, a sample of individuals from the population is being collected and presented with all possible pairs of objects. Each individual then expresses his preference for every possible pair on the basis of a choice criterion (for example more beautiful, more handy e.g.). The individual prefers the object in the pair that best satisfies the choice criterion. As every possible pair is compared, a preference pattern for every individual can be obtained. This is done for every individual, resulting in a data set of the preference of the sample. When the data is obtained, several models can be applied on the gathered data from the sample to obtain a ranking of the objects that expresses the preference of the sample, each model having his own assumptions [6].
The method of pairwise comparisons is frequently used in areas like consumer psychology and marketing, where, because of the fierce competition, a representation of society's preference is needed, so that the company can fulfill society's needs and take out the competition. Next to this, the method of pairwise comparisons has other applications as well, an example being to rank sport teams based on data from past encounters between teams [19].

This report explores Thurstone's Pairwise Comparison Model, which is one of the models that can be used to obtain information about a set of objects from multiple judgment paired comparisons data. One of the main objectives of this research is to investigate several properties of Thurstone's Pairwise Comparison Model. There is not much known about the behavior of the model itself, as it is generally used only as an application [9, 18, 25]. The properties of Thurstone's Pairwise Comparison Model will be investigated with a simulation study.
Another objective of this report is to examine how the Thurstone's Pairwise Comparison Model can be used to model human food preferences. The importance of eating healthy has been promoted in the media for several years now. More and more people are becoming aware of the consequence of eating unhealthy and therefore try to eat as healthy as possible. Pressurized by nutrition centres and the growing awareness of healthy food within the population, supermarkets, kiosks and even educational institutions have changed their inventory and marketing policies promoting and ordering more healthy foods. For example, the TU Delft recently changed the inventory of the vending machines in most of the faculties: replacing all the unhealthy chocolate bars and bags of crisps by sunflower seeds, muesli bars and rice cookies. However, it can be questioned whether this decision was a good move, since some students stated that they missed the old inventory of the vending machines as they sometimes just crave for an unhealthy snack. For these instances, performing a pairwise comparison experiment and applying Thurstone's Pairwise Comparison Model on the resulting data may be of great use, since they are known to be capable of determining which products are generally liked and disliked and can thus give guidelines to the organization about which products to include and exclude in the inventory and the corresponding quantities that need to be ordered for the included products. However, whether a pairwise comparison experiment really is a suitable method will be examined with a data study. For this, two sets of food objects have been taken into consideration, one set of objects consisted of general food products, while the other contained snacks. This decision has been made such that the suitability of a pairwise comparison experiment can be investigated for both supermarkets, selling all kinds of food products, and kiosks or the vending machines of the TU Delft, that primarily sell snack products. Next to that, the influence of the introduction text at the beginning of the experiment on the outcome of the pairwise comparison experiment will be examined as well: does an introduction text, stating the importance of eating healthy, bias the preferences of the participants in that they prefer the more healthy food products? This would be of importance for the initiators of the experiment as well, since they then might also need to take care of how they formulate the research.

The outline of this report is as follows: Section 2 will give an introduction to Thurstone's Pairwise Comparison model. The assumptions of the model will be formulated and the different cases that encompass the model will be explained. The next section, Section 3, presents several methods and techniques that will be used in the data study. The different cases of Thurstone's Pairwise Comparison model will be the main interest for Section 4, where several properties of the model will be investigated with a simulation study. For example, is it possible to determine beforehand which case is most suitable for the data? Some results of the simulation study are needed for Section 5, where the actual data study is carried out. In Section 5, the set-up of the data study will be explained and the results presented. Section 6 will present the final conclusions of this report and Section 7 will give recommendations for future research.

# 2 Thurstone's Pairwise Comparison Model

Thurstone's Pairwise Comparison model, also called Thurstonian model, was invented in 1927 by Louis Leon Thurstone, an American psychologist known for being a pioneer in the field of psychometrics, the science concerned with the measurement of mental capacities and processes. Thurstone applied his model for the first time in an experiment to measure the attitudes of individuals with respect to different opinions [25]. Since then, the model has been used for all kinds of applications and several extensions of the model have been made, including generalized Thurstonian Models [30].

## 2.1 Assumptions of the model

In a pairwise comparison experiment, one wants to compare $n$ objects, say $O_1, O_2, \ldots, O_n$, by obtaining a ranking of the objects. In order to do so, a sample is being collected from a population and each individual of the sample is being presented with all possible pairs of objects. Each pair is presented to the individual once, leading to a total of $\frac{n(n-1)}{2}$ pairwise comparisons when $n$ objects are involved in the experiment. It is, however, not necessary that all $n$ objects are pairwise compared. There also exist cases of incomplete pairwise comparison experiments, where subsets of the sets of objects are compared [21].

Thurstone's model assumes that when an individual is presented with an object $O_i$, denoted by Thurstone [24] as a stimuli, an impression of the object is created within the mind of the individual, which will be denoted by $X_i$. Thurstone calls these impressions sensations. As every individual in the sample will experience an object differently, $X_i$ can be regarded as a stochastic variable. In Thurstone's model it is assumed that the impressions $X_1, X_2, \ldots X_n$ corresponding to the objects $O_1, O_2, \ldots O_n$ follow a jointly normal distribution with mean $S_i$ and variance $\sigma_i^2$:

$$X_i \sim \mathcal{N}(S_i, \sigma_i^2), \quad i = 1, \ldots n.$$

As the objects are presented in pairs to the individual, each object will give rise to an impression within the individual, that follows a normal distribution. The individual will state which of the two objects in the pair he prefers based on the two impressions. For example, if two objects $O_i$ and $O_j$ are compared, the individual will prefer object $O_i$ over object $O_j$ if and only if $X_i > X_j$. In a pairwise comparison experiment, an individual must always prefer one of the objects over the other, hence no ties are allowed.

As the impressions $X_1, X_2, \ldots, X_n$ are jointly normal distributed, it is possible for them to correlated. The correlation between two impressions of objects is denoted by:

$$\mathrm{Cor}(X_i, X_j) = \rho_{ij} \quad i, j = 1 \ldots n.$$

If the correlation between the objects is equal to zero, it is assumed that the impressions of the objects are independent. Next to independence, it is also possible for the impressions of the objects to have equal variances.

The means of the impressions are also called scale values. According to Thurstone [24] a ranking of the objects can be obtained by looking at the scale values of the different objects. The greater the scale value, the higher the rank of the object in the ranking. The preferences of the individuals, however, do not necessarily need to comply with the ranking of the objects obtained from the preferences of the sample. It may occur that an individual prefers an object with a smaller expectation over an object with a larger expectation. This phenomenon is called an inconsistent choice [25]. The probability that an inconsistent choice occurs depends on several factors. In general, when the distance between the means increases, the probability that an inconsistent choice occurs decreases [29]. If the variance of the object with a smaller mean is greater than the object with a larger mean, or the other way around, the probability of an inconsistent choice will increase as well. An example of this will be given by figure 1.

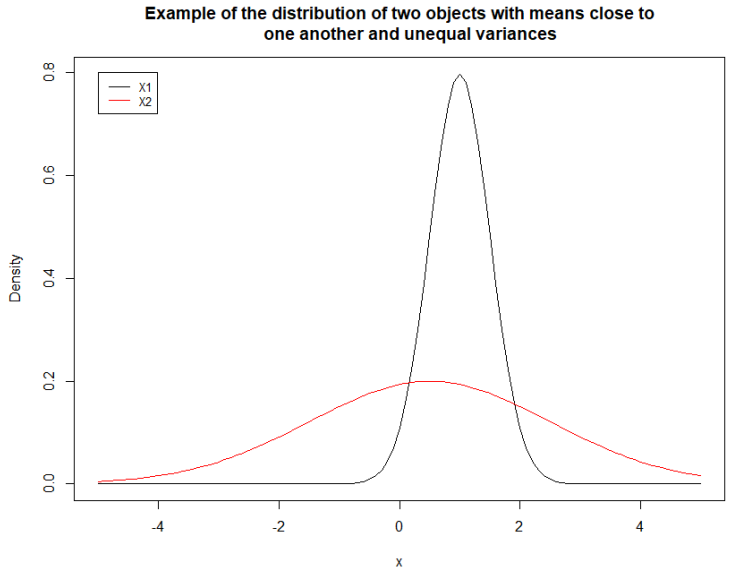



Figure 1: Two distributions with scale values close to one another but great variety in variance: $X_1 \sim \mathcal{N}(0.5, 2)$, $X_2 \sim \mathcal{N}(1, 0.5)$.

Figure 1 shows two objects whose means lie quite close to one another, but the variance of the object with the smaller mean is four times larger than the variance of the higher ranked object. If the objects are uncorrelated, the probability that the first object with a lower mean is regarded greater than the second, is equal to 0.376, which is quite close to 0.5. The same holds when the variance of the object with the larger mean is large compared to the variance of the lower-ranked object. The probability of an inconsistent choice, however, decreases to 0.309, in the case when both objects have variance equal to 0.5. Next to that, the correlation is of influence as well. For instance, when the correlation between the two objects in figure 1 equals 0.5, the probability that the first object is regarded greater than the second object decreases to 0.341 and decreases even further to 0.275 when the correlation equals 0.9. However, when the correlation between the objects is equal to -0.9, the probability of an inconsistent choice increases to 0.405. This has to do with the fact that in the case of a large correlation the relationship between the two objects becomes more linear resulting in a more distinct preference of the individual for the object with a larger mean in the case of a positive correlation, while the objects are more equally preferred when the correlation is negative. In conclusion, it can be stated that the probability of an inconsistent choice increases when the distance between the two objects is small, the variance of one of the objects is large compared to the variance of the other object and the correlation between the objects is large and negative.

An illustrative example for the distribution of four objects, where all objects have equal variance $\sigma^2 = \frac{1}{2}$ and are uncorrelated, can be found in figure 2. From this figure it can be seen that the mean of the first object, $S_1$, is the smallest and the mean of the fourth object, $S_4$, the largest. Following Thurstone's (1927) [24] definition of a ranking, the objects have the following ranking: $(O_1, O_2, O_3, O_4)$.

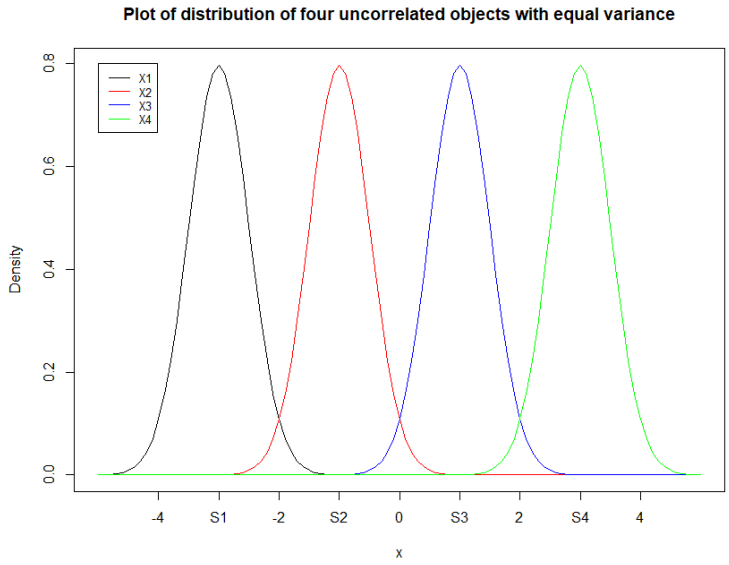



Figure 2: Distribution of four objects, assuming equal variances and zero correlation.

The data obtained from a pairwise comparison experiment consists solely of the proportion of the times that object $i$ has been regarded greater than object $j$ by the sample. In order to get a clear overview of the data, it is convenient to place the proportion data in a matrix or tabular form, also called the proportion matrix. An example of a proportion matrix where three objects $A, B$ and $C$ are compared is given in table 1.

| | A | B | C |
|---|---|---|---|
| A | - | 0.3 | 0.2 |
| B | 0.7 | - | 0.4 |
| C | 0.8 | 0.6 | - |

Table 1: An example of a proportion matrix with gathered data from a pairwise comparison experiment that involves three objects

In general, the number in the $i$-th row and $j$-th column of the proportion matrix is the proportion that the object named at the left of the $i$-th row is judged greater than the object named at the top of the $j$-th column. For example, object A was judged greater than object B in 30 percent of the cases, resulting in a proportion of 0.3. The proportion matrix is sometimes called anti-symmetric as the elements $p_{ij}$ and $p_{ji}$ sum up to 1. Note that the proportion matrix can also be defined the other way around giving the proportion that the columns are preferred over the rows. Furthermore, the diagonal elements of the proportion matrix are left blank, as no comparisons between pairs of the same objects are made.

A Thurstonian model can be regarded as a latent variable model. Latent variables are variables which are not observable, they are deduced through a model from variables that can be observed. A latent variable model links a set of latent variables to a set of observable variables by assuming that the outcomes of the observable variables are the consequence of the individual positions on the latent variables. In the case of a Thurstonian Model, the latent variables are the impressions of the objects. The values of these impressions are drawn from a normal distribution, where the choice of the parameters of the normal distribution depend on the pairwise comparison data of the sample, the observed variables. The dependence of the parameters belonging to the normal distribution of the impressions on the observed pairwise comparison data can be written down in a formula, which is called The Law of Comparative Judgment.

## 2.2 The Law of Comparative Judgment

The Law of Comparative Judgment, invented by Thurstone, expresses the dependence of the unknown parameters of the normal distributions of the objects on the observed pairwise comparison data. Assume that a sample of the population has participated in a pairwise comparison experiment consisting of $n$ objects. Every individual expresses a preference for each pair of objects $(O_i, O_j)$. The individual either chooses $O_i$ or $O_j$ depending on which impression of the objects is greater: choosing object $O_i$ if $X_i > X_j$ and $O_j$ if $X_i < X_j$.

Now, denote by $p_{ij}$ the proportion of times that the sample preferred object $O_i$ over $O_j$, or in other words the proportion of times that $X_i$ exceeded $X_j$. The proportion $p_{ij}$ can therefore be regarded as the probability of $X_i$ to exceed $X_j$ when a pair of objects $(O_i, O_j)$ is presented to a random individual from the sample:

$$p_{ij} = \mathbb{P}(X_i > X_j) = \mathbb{P}(X_i - X_j > 0).$$

As the impressions of the objects $X_i$ and $X_j$ follow a normal distribution, with means $S_i$ and $S_j$, variances $\sigma_i^2$ and $\sigma_j^2$ and correlation $\rho_{ij}$ respectively, their difference also follows a normal distribution:

$$X_i - X_j \sim \mathcal{N}(S_i - S_j, \sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j),$$

where the variance of $X_i - X_j$ has been reformulated the following way:

$$\begin{aligned} \mathrm{Var}(X_i - X_j) &= \mathrm{Var}(X_i) + \mathrm{Var}(X_j) - 2\mathrm{Cov}(X_i, X_j) \\ &= \sigma_i^2 + \sigma_j^2 - 2\mathrm{Cor}(X_i, X_j)\sqrt{\mathrm{Var}(X_i)}\sqrt{\mathrm{Var}(X_j)} \\ &= \sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j. \end{aligned}$$

Since the distribution of $X_i - X_j$ is normal, this distribution can be standardized to the standard normal distribution by subtracting the mean of $X_i - X_j$ from $X_i - X_j$ and dividing by its standard deviation on both sides in the probability.

$$\begin{aligned} p_{ij} &= \mathbb{P}(X_i - X_j > 0) \\ &= \mathbb{P}\left(\frac{X_i - X_j - (S_i - S_j)}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}} > \frac{0 - (S_i - S_j)}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}}\right) \\ &= \mathbb{P}\left(Z_{ij} > \frac{-(S_i - S_j)}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}}\right) \\ &= 1 - \mathbb{P}\left(Z_{ij} \leq \frac{-(S_i - S_j)}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}}\right) \\ &= 1 - \Phi\left(\frac{-(S_i - S_j)}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}}\right) \\ &\overset{1-\Phi(x)=\Phi(-x)}{=} \Phi\left(\frac{S_i - S_j}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}}\right). \end{aligned}$$

Applying the inverse of the cumulative standard normal distribution function on both sides of the equation yields:

$$z_{ij} = \frac{S_i - S_j}{\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}},$$

where $z_{ij} = \Phi^{-1}(p_{ij})$. The formula above is an alternative formulation of the Law of Comparative Judgment, expressing the standard normal deviate corresponding to the proportion in terms of the

parameters of the normal distributions of the objects. Multiplying both sides by $\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}$ gives the most common formulation of the Law of Comparative Judgment as described by Thurstone [24]:

$$S_i - S_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}.$$

When a set of $n$ objects is assumed to be uncorrelated and having unequal variances, there are a total of $n$ means, $n$ variances and $\frac{n(n-1)}{2}$ correlations to be estimated, while only $\frac{n(n-1)}{2}$ pairwise comparisons are given. The solution in this case can not be obtained as the number of parameters exceeds the number of known equations, making the system of equations insolvable [20]. In order to make the system of equations solvable, simplifications are needed [19].
Introducing the Law of Comparative Judgment, Thurstone [24] described five cases of the Law of Comparative Judgment each having their own assumptions that simplify the model. The following subsections will introduce the different cases.

### 2.2.1 Case V

The most simple case of the Law of Comparative Judgment, which is denoted by Thurstone (1927) [24] as Case V, assumes that the objects are uncorrelated and have equal variances. As the objects are independent from one another and all have equal variances, the Law of Comparative Judgment in this case can be written as follows:

$$S_i - S_j = z_{ij}\sqrt{\sigma^2 + \sigma^2} = z_{ij}\sqrt{2\sigma^2}.$$

Mosteller (1951) [19] argues that, without loss of generality, it can be assumed that $\sigma^2 = \frac{1}{2}$, yielding:

$$S_i - S_j = z_{ij}.$$

Assuming that $\sigma^2 = \frac{1}{2}$ has been done for a matter of convenience, making it easier to derive the analytic formula for the estimations of the scale values $S_i$ [27]. It is, however, also possible to assume a different common variance [2].
Note that by assuming that the objects are uncorrelated and having equal variance, the number of parameters to be estimated is reduced to $n$, namely the $n$ scale values of the $n$ objects. This gives that a solution can be obtained if the number of objects in the pairwise comparison experiment is greater than or equal to three, as in that case the number of equations equals the number of parameters to be estimated.

### 2.2.2 Case III and Case IV

Case III of Thurstone's Law of Comparative Judgment assumes that the objects are still uncorrelated, but the variances of the objects are unequal to one another. According to Thurstone (1927) [24], the correlation between the objects are small or almost equal to zero if the objects involved in the comparison experiment are homogeneous. This means that the objects are very much alike, for example when one compares the taste of different chocolate bars. All bars taste like chocolate, but differ in sweetness and consistency. Setting the correlation $\rho_{ij}$ to zero for all pairs $(O_i, O_j)$ gives the Law of Comparative Judgment for Case III:

$$S_i - S_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2}.$$

Assuming zero correlation between the objects decreases the number of parameters to be estimated to $2n$, namely the $n$ scale values and $n$ variances of the objects. When $n = 5$, the number of equations: $\frac{n(n-1)}{2} = 10$ equals the number of parameters to be estimated. Therefore, the system of equations for this case becomes solvable when the number of objects to be compared is at least five.

Next to Case III, Thurstone has also defined Case IV as a simplified case. Case IV assumes that the objects are independent and have unequal variance, however the variances are quite close to one another. One can write, in this case, for the standard deviations of two objects $O_i$ and $O_j$:

$$\sigma_j = \sigma_i + d$$

where $d$ has been chosen small, at least smaller than $\sigma_i$ [24]. Thurstone (1927) [24] has showed that for Case IV the Law of Comparative Judgment can be written in a linear form in the following way:

$$S_i - S_j = \frac{1}{2}\sqrt{2}z_{ij}(\sigma_1 + \sigma_2)$$

The minimum number of objects needed in the pairwise comparison experiment in order for the system of equations to be solvable is the same as in Case III.

### 2.2.3 Case I and Case II

In Case I and Case II of the Law of Comparative Judgment, it is assumed that the objects are correlated and have unequal variances. In these cases, the full version of the Law of Comparative Judgment holds:

$$S_i - S_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}$$

As mentioned before, the estimation for Case I and Case II brings about the problem that the parameters, if formulated in this setting, can not be estimated as the number of equations gathered from the data, which equals $\frac{n(n-1)}{2}$, is less than the number of parameters to be estimated: $2n + \frac{n(n-1)}{2}$. However, Mosteller [20] has shown that this problem can be overcome by relaxing the assumption of different correlations assuming equal correlations between all the objects. The Law of Comparative Judgment, in this case, becomes:

$$S_i - S_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho\sigma_i\sigma_j}$$

Note that assuming constant correlation between the objects reduces the number of parameters to be estimated from $2n + \frac{n(n-1)}{2}$ to $2n + 1$. Therefore, the system of equations for this case becomes solvable if the number of objects participating in the experiment is six or more. Next to that, it must be noted that in order for the covariance matrix to be positive definite, the correlation coefficient $\rho$ must be in the interval $(\frac{1}{n-1}, 1)$ if $n$ objects are compared [16].
One might wonder why a distinction is made between Case II and Case I, as both cases are mathematically identical to one another. The difference between the two cases is the sample size that is involved in the experiment. Thurstone (1927) [24] states that Case I involves only one individual that compares the $\frac{n(n-1)}{2}$ pairs of objects numerous times, while in Case II a group of individuals compares the $\frac{n(n-1)}{2}$ pairs of objects only once.

It is not completely obvious why Thurstone has made this distinction only for the most complex case. Thurstone (1927) [24] talks about a general sample, that can either be an individual or a group of individuals for the more simple cases. Henceforth, as the interest of this report is to investigate the application of Thurstone's Pairwise Comparison model to model the difference in preference of groups, the sample is regarded as a group of individuals for the simplified cases. Next to that, only cases II, III and V will be taken into consideration in the follow-up of this report, since Case IV is too restrictive on account of the variances. It would not make sense to assume that the variances are quite close to one another as the preference of the sample is not known beforehand and thus uncertain. Because of this uncertainty, it is necessary to allow for all possible variances.

## 2.3 Methods to estimate the parameters of the distributions

In this section, several solving methods for cases II, III and V are presented as these cases will be taken into consideration in the simulation study of Section 4.

Before the solving methods are explained, a restriction on the estimated scale values and a unit of measurement for the variances are introduced first.
A restriction on the scale values is needed so that the scale values can be obtained and compared. Remember that the proportion $p_{ij}$ can be expressed as a function of the parameters of the distributions which, in all cases, included the scale separations $S_i - S_j$. For example, in Case V:

$$p_{ij} = \Phi\left(\frac{S_i - S_j}{\sqrt{2\sigma^2}}\right).$$

When all the scale separations are known, there is no unique zero point for the scale values [26]. One can compare this to the situation where one only knows the difference between prices of certain goods. Knowing the difference in price does not say anything about the price of the goods itself. According to Thurstone & Jones (1957) [26] numerical values can be assigned to the objects only by setting an arbitrary origin.
Thus, in order to obtain scale values, one must set an origin beforehand. Several options for the origin exist, for example one may choose to set the scale value of the first object equal to zero [19, 26]. The choice of origin that will be used in this case is defined by setting the sum of the scale values equal to zero:

$$\sum_{i=1}^{n} S_i = 0.$$

In general, defining an origin for the scale values brings about another advantage with respect to detecting the degree of variability of the scale values. In this case, if the means are close to one another, all mean values would be close to zero.

Next to a restriction on the scale values, a unit of measurement for the variances is needed as well. Several units of measurement for the variance have been defined [2]. The unit of measurement chosen in this case is the one invented by Burros (1951) [5], who has defined the unit of measurement of the variance to be equal to the number of objects:

$$\sum_{i=1}^{n} \sigma_i^2 = n.$$

The necessity of the choice of a unit of measurement for the variances can be explained by looking at the formula of Case III which expresses the proportion $p_{ij}$ as a function of the means and variances:

$$p_{ij} = \Phi\left(\frac{S_i - S_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right)$$

If the variances are equal to one another, one would have, by the unit of measurement, that $\sigma_i^2 = \sigma_j^2 = 1$. This implies that the proportion $p_{ij}$ would only depend on the difference between the scale values. If, however, the variances differ, the proportions $p_{ij}$ would depend on both the variances and means. The purpose of the unit of measurement is to capture the difference in variance for the different cases so that they can be mutually compared to one another. For instance, if the estimated scale values after fitting both Case III and V on the proportion matrix yield almost the same outcome and the estimates of the variances in Case III are all close to one, it can be concluded that Case V is sufficient enough for this case.

### 2.3.1 Parameter estimation for Case V

The only parameters to be estimated in Case V are the scale values. Torgerson (1958) [27] has stated that reliable estimates of the scale values can easily be obtained by averaging the rows of the standard normal deviate matrix of the proportion matrix and multiplying it by two times the square root of the common variance. In other words:

$$\widehat{S_i} = \frac{\sqrt{2\sigma^2}}{n} \sum_{j=1}^{n} z_{ij}.$$

Since

$$S_i - S_j = z_{ij}\sqrt{2\sigma^2},$$

the estimator for the scale values can easily be derived by minimizing the sum of squares of the estimates:

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} (z_{ij}\sqrt{2\sigma^2} - (\widehat{S_i} - \widehat{S_j}))^2,$$

where $\sigma^2$ is the common variance of the objects. Differentiating the sum of squares above to $\widehat{S}_i$ and setting the derivative equal to zero yields:

$$\begin{aligned}
\frac{\partial \sum}{\partial \widehat{S}_i} \sum_{\substack{j=1 \\ j \neq i}}^{n} (z_{ij}\sqrt{2\sigma^2} - (\widehat{S}_i - \widehat{S_j}))^2 &= 0 \\
2 \sum_{\substack{j=1 \\ j \neq i}}^{n} (-z_{ij}\sqrt{2\sigma^2} + \widehat{S}_i - \widehat{S_j})) &= 0 \\
\sum_{j=1}^{i-1} -z_{ij}\sqrt{2\sigma^2} + \widehat{S}_i - \widehat{S_j} + \sum_{j=i+1}^{n} -z_{ij}\sqrt{2\sigma^2} + \widehat{S}_i - \widehat{S_j} &= 0 \\
-\sum_{j=1}^{n} z_{ij}\sqrt{2\sigma^2} + (n-1)\widehat{S}_i - S_1 \ldots - S_{i-1} - S_{i+1} \ldots - S_n &= 0 \\
-\sum_{j=1}^{n} z_{ij}\sqrt{2\sigma^2} + n\widehat{S}_i - \sum_{j=1}^{n} \widehat{S_j} &= 0 \\
n\widehat{S}_i &= \sqrt{2\sigma^2} \sum_{j=1}^{n} z_{ij} \\
\widehat{S}_i &= \frac{\sqrt{2\sigma^2}}{n} \sum_{j=1}^{n} z_{ij}
\end{aligned}$$

Note that in this case it is assumed that $z_{ii} = 0$, such that the two sums of the $z_{ij}$ can be merged together. For this, the diagonal elements of the proportion matrix, which were previously left blank, are now assumed to be 0.5 such that

$$z_{ii} = \Phi^{-1}(p_{ii}) = \Phi^{-1}(0.5) = 0$$

for all $i$. Presuming that the diagonal elements of the proportion matrix are equal to 0.5 is needed in order to complete the derivation of the analytic formula for the estimation of the scale values [19]. The restriction placed on the scale values, $\sum_{i=1}^{n} S_i = 0$, is needed for the derivation of the analytic formula for the scale values as well.

As the variance restriction chosen in this report dictates that the variances must sum up to $n$, the common variance $\sigma^2$ of the objects must be chosen equal to 1, which yields:

$$\widehat{S}_i = \frac{\sqrt{2}}{n} \sum_{j=1}^{n} z_{ij}.$$

Furthermore, when the proportion $p_{ij}$ equals zero or one, a solution can not be obtained, since in that case the inverse of the standard normal deviate equals (minus) infinity:

$$\lim_{x \to 0} \Phi^{-1}(x) = -\infty \text{ and } \lim_{x \to 1} \Phi^{-1}(x) = \infty.$$

This is being resolved by slightly changing the proportions $p_{ij}$ using the sample size. Let $N$ be the sample size, then the proportions $p_{ij}$ are transformed in the following way:

$$\widetilde{p_{ij}} = \begin{cases} \frac{1}{N} & \text{if } p_{ij} = 0 \\ 1 - \frac{1}{N} & \text{if } p_{ij} = 1 \end{cases}$$

Note that transforming the $p_{ij}$ in this way applies in all three cases.

### 2.3.2 Parameter estimation for Case II and Case III

The estimates for the parameters in Case II and III of Thurstone's model can not, as opposed to Case V, be derived analytically. Therefore, the estimations have to be computed numerically. Several methods exist to compute the estimations, such as maximum likelihood, weighted non linear least squares, also known as the Minimum Pearson $\chi^2$, and unweighted non linear (ordinary) least squares [10]. All methods have one thing in common: in order to obtain the estimates of the parameters, the objective function resulting from the chosen method has to be minimized by making use of non-linear optimization. In this case, the method of unweighted least squares has been chosen for the estimation of the parameters in cases II and III. This choice has been made based upon the fact that unweighted least squares has been employed in Case V for the parameter estimation as well, therefore making it easier to compare the performance of the models.

The objective function $f$ to be minimized in the case of unweighted non linear least squares is defined as follows:

$$f = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (\widehat{z_{ij}} - z_{ij})^2,$$

where $\widehat{z_{ij}}$ equals the Law of Comparative Judgment for the estimated parameters and $z_{ij}$ is the Law of Comparative Judgment for the true values of the parameters. For Case III, this gives the following:

$$\widehat{z_{ij}} = \frac{\widehat{S_i} - \widehat{S_j}}{\sqrt{\widehat{\sigma_i^2} + \widehat{\sigma_j^2}}}$$

and

$$z_{ij} = \frac{S_i - S_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}$$

For Case II, the same formulas can be employed, adding the estimated and true correlation term in the denominators of the fractions. The estimates for the parameters are found by minimizing the objective function $f$.
In general, the parameters that minimize the objective function are exactly these values that set the gradient of the objective function equal to the zero.
Note that, since there are two constraints placed on the parameters as well, one has to make use of non-linear constrained optimization in order to obtain the estimates. Several methods exist for this purpose of which two which will be explained.

The first and also well-known method is the method of Lagrange multipliers. This method tries to find local optima (either minima or maxima) of an objective function on its feasible region. Suppose that a function $f$ needs to be minimized having the following constraints:

$$\begin{aligned} &\min f(\mathbf{x}) \\ &\text{subject to } \mathbf{g}(\mathbf{x}) = \mathbf{a} \end{aligned}$$

where $\mathbf{x}$ is the vector of parameters, $\mathbf{g}$ is the vector with constraints and $\mathbf{a}$ is the vector of values which the constraints should satisfy. In this case, there are only two constraints, one for the scale values and one for the variances, and the value vector $\mathbf{a} = (0, n)$. The method of Lagrange tries to find the local optimum by ensuring that the gradient of the objective function is parallel to the gradients of the constraints. As the minimum arises when the level curves of the function $f$ are tangent to the curves of the constraints and the gradient of a function is perpendicular to the level curves, the gradient of $f$ is perpendicular to the curves of the constraints. This would imply that the gradients of both $f$ and the constraints are parallel to one another. In other words, the following equation holds:

$$\nabla f = \sum_{i=1}^{M} \lambda_i \nabla g_i,$$

where $\lambda_i$ are the Lagrange multipliers.
In order to solve for the parameters, the following auxiliary function is introduced:

$$\mathcal{L}(\mathbf{x}, \lambda_1, \ldots, \lambda_M) = f(\mathbf{x}) - \sum_{i=1}^{M} \lambda_i (g_i(\mathbf{x}) - a_i).$$

If the gradient of $\mathcal{L}$ is set to zero, it can be seen that:

$$\begin{aligned} &\nabla \mathcal{L}(\mathbf{x}, \lambda_1, \ldots, \lambda_M) = 0 \\ &\Rightarrow \nabla f(\mathbf{x}) - \sum_{i=1}^{M} \lambda_i \nabla g_i(\mathbf{x}) = 0 \\ &\Rightarrow \mathbf{g}(\mathbf{x}) = \mathbf{a} \end{aligned}$$

So by solving the equation $\nabla \mathcal{L} = 0$, the parameters that both satisfy the constraints and minimize the objective function are obtained.

Next to the method of Lagrange multipliers, which uses derivatives in order to obtain the minimum, one can also make use of derivative free optimization methods. These methods are especially useful in the case when derivatives are unknown or are computationally difficult. One of the derivative free optimization methods is the Constrained optimization by linear approximation (COBYLA). This method is capable of numerically finding the minimum of the objective function in the feasible region without using the gradient of the function.
The COBYLA method is an iterative method that approximates an $n$-dimensional constrained optimization problem by making use of a simplex of $(n+1)$ vertices. During each iteration, a linear approximation of the objective function is constructed and then solved by making use of the Simplex algorithm. The solution obtained from the Simplex algorithm is then evaluated in the original constrained problem and the simplex for the next iteration is set up by replacing one of the vertices in the simplex by the new solution. The decision which vertex to replace is controlled by avoiding the degenerate situation where the volume of the simplex decreases to zero. The algorithm ceases when the solution of the Simplex algorithm can not be further improved [22].

Only one of the two methods described above is implemented in the statistical software program R. There does not exist a function for the method of Lagrange multipliers in R. However, there exists a function in R which minimizes the objective function using the augmented Lagrange method, which solves the constrained problem by replacing the constrained problem by a series of unconstrained problems while adding a penalty term to the objective. The function for this method in R is called solnp and by default minimizes the objective function unless stated otherwise. The function for the COBYLA method in R is called cobyla. Nevertheless, the solnp function seemed to have problems finding the optimal solution as it stated that the obtained solution was not reliable because of problems inverting the Hessian when some examples were tried. Therefore only the cobyla function will be used in this report.

It should be noted that the cobyla function requires starting parameters in order to obtain a solution. It was discovered during the implementation that the cobyla function is very sensitive giving different solutions for different initial conditions. The question that arises is: how should these starting parameters be chosen? If the distribution is known beforehand, one could use the parameters of the distributions as the initial condition since these values are already quite close to the optimal solution. Nonetheless, this would not seem very realistic in the case of real data as the distribution is simply not known.
A solution for this would be to make use of a two-step optimization process. In this process, starting parameters for the cobyla function are determined by making use of a method which minimizes the unconstrained problem. The unconstrained optimization method chosen in this case is the Quasi-Newton method. This method is a more computationally efficient variant of Newton's method, where the Hessian matrix is approximated in each iteration using values of the gradient of the objective function obtained from the previous iterations. The resulting parameters from this optimization method are then used as the starting parameters for the cobyla. The performance of this two-step optimization process is examined in the simulation study in Section 4.

## 2.4 A method to determine the best model for the data

After performing a pairwise comparison experiment it is not known in advance which of the cases of Thurstone's model is most applicable for the gathered data. A few guidelines on this matter have been given in literature [20, 24], however they are quite vague and prone to one's own interpretation. For example, Thurstone (1927) [24] states that it is a safe assumption to assume that the correlation between the objects is zero when they are similar, therefore excluding Case II. Yet, this raises the question: when are objects similar to one another? The definition of similarity is subjective, differing per individual: some stating a set of objects as similar while others do not.

There exist several methods that can be used to determine the best fitting model for the data set. In this case, the choice has been made to use the Akaike Information Criterion (AIC). This criterion is defined as follows:

$$\text{AIC} = 2k - 2\ln(\widehat{L}),$$

where $k$ equals the number of parameters of the model and $\widehat{L}$ corresponds to the maximum value of the likelihood function of the model. The value of the likelihood function is obtained at the values of the estimated parameters:

$$\widehat{L} = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \widehat{p_{ij}}^{f_{ij}} (1 - \widehat{p_{ij}})^{N - f_{ij}},$$

where $N$ is the sample size that participated in the experiment, $f_{ij}$ is the frequency: $f_{ij} = Np_{ij}$, that object $i$ was regarded greater than object $j$ in the experiment and $\widehat{p_{ij}}$ equals the estimated proportion resulting from the estimated parameters. For example, in Case III, the estimated proportion equals

$$\widehat{p_{ij}} = \Phi\left(\frac{\widehat{S_i} - \widehat{S_j}}{\sqrt{\widehat{\sigma_i^2} + \widehat{\sigma_j^2}}}\right).$$

The best fitting model for a data set is determined as follows: one fits all three cases on the data set in question and computes the AIC values of the estimated parameters. The model that has the smallest AIC value is the best fitting model for the data set. In general, adding too many parameters to the model may lead to the model over-fitting the data, meaning that the model perfectly fits the existing data but might fail to fit additional data. The Akaike Information Criterion is a penalized information criterion that takes care of this by adding two times the number of parameters to the term of two times the logarithm of $\widehat{L}$.

The AIC appears to be a suitable method that can be used to determine the best model for a given data set. However, not much is known about how the AIC performs in the case of Thurstone's Pairwise Comparison Model as no tests regarding the performance of the AIC have been conducted so far. As the main interest of Thurstone's Pairwise Comparison Model is to obtain the rankings of the objects, it may be questioned whether the ranking of the best model according to the AIC always coincides with the true ranking. Section 4 will investigate whether the choice of the model based on the AIC recovers the true ranking and how this is influenced by other factors, such as the sample size and the number of objects.

# 3 Methods to determine the number of intransitive preferences and agreement within and between groups

One of the main objectives of this report is to investigate how food preferences of individuals can be modelled by making use of Thurstone's Pairwise Comparison Model and how the preferences of individuals might be influenced by an introduction text given at the beginning of the experiment. These questions, as mentioned earlier, will be answered by means of a data study. This section will introduce several methods needed later on in the analysis of the data.

## 3.1 Intransitive preferences of individuals

As mentioned earlier, an individual may sometimes make an inconsistent choice in a pairwise comparison experiment by preferring a lower ranked object over a higher preferred object. These inconsistent choices may lead to the individual having intransitive preferences for the objects [29]. For example, when presented with three objects $(O_1, O_2, O_3)$ an individual might judge $O_1 > O_2$ and $O_2 > O_3$, but instead of judging $O_1 > O_3$, he judges $O_3 > O_1$. This phenomenon is also called an intransitive response or circular triad as $O_1 > O_2 > O_3 > O_1$ [3]. Next to circular triads of length three, there also exist circular triads of length four or higher. Kendall and Babington-Smith (1940) [13], however, showed that a circular $n$-ad contains at least $n - 2$ circular triads for $n \geq 4$. The occurrence of circular triads in the preferences of individuals is a general characteristic of pairwise comparison experiments [29]. They may occur because the individual makes a mistake in his judgments or the individual might not have a preference for the alternatives, meaning that he is indifferent for some pairs of objects [12]. Whether an individual makes a mistake in his judgments or has a lack of preference for the objects involved in the experiment when his preferences contain circular triads is not easily determined [11]. Therefore, if one wants to retrieve why an individual made an intransitive choice, the only option would be to personally ask this individual for the reason.
Moreover, intransitive preferences of individuals can not be detected from the proportion matrix, as this matrix gives the combined judgment of the sample. So, how can one observe whether an individual has circular triads in his preferences?

For this purpose, Kendall and Babington-Smith (1940) [13] have derived a formula that determines the total number of circular triads made by an individual by looking at the preferences of the individual. The formula in question is defined as follows:

$$CT = \frac{n(n^2-1)}{24} - \frac{T}{2}$$

where $CT$ is the number of circular triads, $n$ the number of objects, $a_i$ the number of times object $i$ is preferred over the other objects by the individual and

$$T = \sum_{i=1}^{n} \left( a_i - \frac{n-1}{2} \right)^2$$

The proof of this formula is based on graph theory, where the preferences of the individual are transformed into a directed graph.

*Proof.* Suppose that an individual has made all $\frac{n(n-1)}{2}$ paired comparisons of the $n$ objects, where ties were not permitted. Then his preferences can be re-expressed in the language of graph theory in the following way: the vertices $A_1, \ldots, A_n$ in the graph represent the $n$ objects involved in the experiment, where each node is connected with every other node. The direction of the arcs between the vertices indicate the preference of the individual, where the vertex from which the arc issues is being preferred. An example of a preference matrix of an individual and the corresponding graph representation when $n = 5$ can be seen in table 2 and figure 3.

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $A_1$ | - | 0 | 1 | 1 | 1 |
| $A_2$ | 1 | - | 1 | 0 | 0 |
| $A_3$ | 0 | 0 | - | 0 | 0 |
| $A_4$ | 0 | 1 | 1 | - | 1 |
| $A_5$ | 0 | 1 | 1 | 0 | - |

Table 2: Example of a preference matrix of an individual for 5 objects.

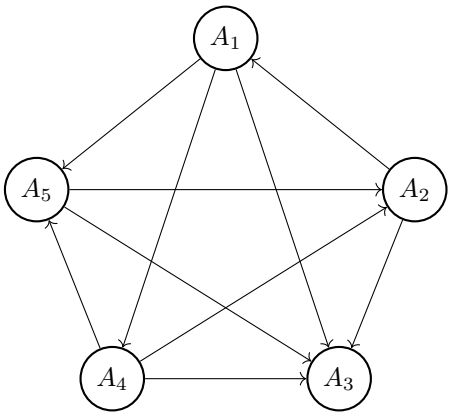



Figure 3: Directed graph representation of the preference matrix.

In the example in figure 3, it can be seen that the individual prefers object 1 over objects 3, 4 and 5 as the lines from $A_1$ are directed to the vertices $A_3, A_4$ and $A_5$. Now, if we call $a_i$ the number of arcs that are leaving vertex $A_i$, then we have in this case that $a_1 = 3$.
Note that:

$$\sum_{i=1}^{n} a_i = \frac{1}{2}n(n-1)$$

since $\frac{1}{2}n(n-1)$ comparisons are made and the individual always prefers one of the two objects of the pair.

The number of circular triads in the preferences of the individual can be determined by looking at the difference between the total number of triads that can occur and the total number of non-circular triads in the individual's preferences.

Note that a triad will not be circular if and only if two of its arcs issue from one of the triad's vertices. For example, looking at the graph in figure 3 it can be seen that the triad $A_1 - A_4 - A_5$ is non-circular as two arrows descend from $A_1$ making a circular tour between the vertices impossible.
The total number of non-circular triads in the preferences of an individual is equal to the sum of all the combinations of two that can be made from the number of outgoing lines from the vertices:

$$\begin{aligned}\sum_{i=1}^{n}\binom{a_i}{2} &= \sum_{i=1}^{n}\frac{a_i!}{(a_i-2)!2!} \\ &= \frac{1}{2}\sum_{i=1}^{n} a_i(a_i-1) \\ &= \frac{1}{2}\sum_{i=1}^{n} a_i^2 - \sum_{i=1}^{n} a_i.\end{aligned}$$

Now, if we define $\overline{a} = \frac{1}{2}(n-1)$ as the mean of the $a_i$'s, then it follows that $T$ can be rewritten as follows:

$$\begin{aligned}T &= \sum_{i=1}^{n}(a_i - \overline{a})^2 \\ &= \sum_{i=1}^{n} a_i^2 - 2a_i\overline{a} + \overline{a}^2\end{aligned}$$

such that

$$\sum_{i=1}^{n} a_i^2 = T + 2\overline{a}\sum_{i=1}^{n} a_i - n\overline{a}^2$$

Furthermore, we have that: $\sum_{i=1}^{n} a_i = n\overline{a}$. Putting everything together yields that the total number of non circular triads equals:

$$\begin{aligned}
\sum_{i=1}^{n} \binom{a_i}{2} &= \sum_{i=1}^{n} \frac{a_i!}{(a_i-2)!2!} \\
&= \frac{1}{2}\sum_{i=1}^{n} a_i^2 - \sum_{i=1}^{n} a_i \\
&= \frac{1}{2}\left(T + 2\overline{a}\sum_{i=1}^{n} a_i - n\overline{a}^2 - n\overline{a}\right) \\
&= \frac{1}{2}T + \left(\overline{a} - \frac{1}{2}\right) n\overline{a} - \frac{1}{2}n\overline{a}^2 \\
&= \frac{1}{2}T + \frac{1}{2}\left(\overline{a}-1\right) n\overline{a} \\
&= \frac{1}{2}T + \frac{1}{2}\left(\frac{1}{2}(n-1)-1\right) n\frac{1}{2}(n-1) \\
&= \frac{1}{2}T + \frac{1}{4}n(n-1)\left(\frac{1}{2}n - \frac{3}{2}\right) \\
&= \frac{1}{2}T + \frac{1}{8}n(n-1)(n-3).
\end{aligned}$$

Now, as the total number of triads equals all possible combinations of three of the $n$ objects, we have that the number of circular triads is equal to:

$$\begin{aligned}
CT &= \binom{n}{3} - \frac{1}{2}T + \frac{1}{8}n(n-1)(n-3) \\
&= \frac{1}{6}n(n-1)(n-2) - \frac{1}{2}T + \frac{1}{8}n(n-1)(n-3) \\
&= \frac{1}{6}n^3 - \frac{1}{2}n^2 + \frac{1}{3}n - \frac{1}{8}n^3 + \frac{1}{2}n^2 - \frac{3}{8}n - \frac{1}{2}T \\
&= \frac{1}{24}n^3 - \frac{1}{24}n - \frac{1}{2}T \\
&= \frac{n(n^2-1)}{24} - \frac{T}{2},
\end{aligned}$$

which proves the formula. □

Next to the number of circular triads, Kendall and Babington-Smith (1940) also determined the maximum number of circular triads that may occur in the preferences of an individual dependent on $n$. They discovered that the maximum number of circular triads is equal to:

$$CT_{\max} = \begin{cases} \frac{n(n^2-1)}{24} & \text{if } n \text{ is odd} \\ \frac{n(n^2-4)}{24} & \text{if } n \text{ is even} \end{cases}$$

*Proof.* The proof of this statement will make use of the function $T$ defined in the proof of the number of circular triads. First, the function $T$ is rewritten in the following way:

$$\begin{aligned}
T &= \sum_{i=1}^{n}\left(a_i - \frac{n-1}{2}\right)^2 \\
&= \sum_{i=1}^{n} a_i^2 - \frac{1}{4}n(n-1)^2
\end{aligned}$$

Now, we will first prove that if the direction of a preference is altered and the effect is to increase the number of circular triads by say $p$, then the function $T$ will be reduced by $2p$ and the other way around.

For this, consider the preference $A \to B$, where "$\to$" means "is preferred over". The only triads that are affected by reversing the preference to $B \to A$ are those triads containing the line $AB$. Suppose there are $\alpha$ preferences where $A \to X$ (including $B$) and $\beta$ preferences where $B \to$ X. Then there are four possible types of triads:

$$\begin{aligned}
&A \to X \leftarrow B, && \text{say } x \text{ in number}\\
&A \leftarrow X \to B, &&\\
&A \to X \to B, && \text{which must number } \alpha - x - 1\\
&A \leftarrow X \leftarrow B, && \text{which must number } \beta - x
\end{aligned}$$

Looking at the four possibilities it can be seen that the first three triads are non-circular and the last one is circular. Now, if the preference $A \to B$ is reversed, the first two triads remain non circular, while the third one becomes circular and the last one becomes non-circular. The increase in the number of circular triads is therefore equal to:

$$\alpha - x - 1 - \beta - x = \alpha - \beta - 1 = p$$

By changing the preference $A \to B$, the number of preferences $A \to X$ decreases by 1, while the number of preferences $B \to X$ increases by 1. Therefore, the reduction in $T$ by changing the preference $A \to B$ equals:

$$\begin{aligned}
\Delta T &= \alpha^2 - (\alpha - 1)^2 + \beta^2 - (\beta + 1)^2\\
&= 2(\alpha - \beta - 1)\\
&= 2p
\end{aligned}$$

Furthermore, it is clear from the definition of $T$ that its maximum value is obtained when the preferences are ranked, which means that no circular triads occur in his preferences. This gives that the maximum value of $T$ equals:

$$\begin{aligned}
T_{\max} &= \sum_{i=1}^{n} a_i - \frac{1}{4}n(n-1)^2\\
&= \sum_{i=0}^{n-1} i - \frac{1}{4}n(n-1)^2\\
&= \frac{n(n-1)(2n-1)}{6} - \frac{1}{4}n(n-1)^2\\
&= \frac{n^3 - n}{12}
\end{aligned}$$

In order to determine the minimum value of $T$, one can consider a polygon like in figure 3 with nodes $A_1, \ldots, A_n$ and construct a full preference scheme in the following manner: first set up the preferences $A_1 \to A_2 \to \cdots \to A_n \to A_1$. Then set up the preferences $A_1 \to A_3 \to A_5 \to$. If the current scheme does not yield a closed tour of all points in the graph, continue to the next unvisited vertex $A_k$ and set up the preference scheme $A_k \to A_{k+2} \to$ etc. After that, set up the preferences $A_1 \to A_4 \to A_7 \to$, and continue until the whole preference scheme is complete.

Now, if $n$ is odd, the preference scheme will consist of circular tours of the polygon. Every vertex $A_i$ will have $\frac{1}{2}(n-1)$ arrows departing from the vertex, giving that $a_i = \frac{1}{2}(n-1)$ for all $i$. This gives that the minimum value of T is equal to zero when $n$ is odd.
If $n$ is even, the preference $A_1 \to A_{\frac{1}{2}n+1}$ will not be a part of a circular tour but a single line joining the two vertices. This gives that there will be $\frac{n}{2}$ vertices with $a_i = \frac{n}{2}$ and $\frac{n}{2}$ vertices with $a_i = \frac{n}{2} - 1$.
The minimum value of $T$ in this case equals: $T = \sum_{i=1}^{n} \frac{1}{4} = \frac{n}{4}$.

This gives that $T$ can range from 0 or $\frac{n}{4}$ to $\frac{n^3-n}{12}$ depending on whether $n$ is odd or even. Since an increase of two in $T$ brings about a decrease of unity in the number of circular triads, it can be derived that the maximum number of circular triads equals the half of the difference between the minimum and

maximum value of $T$ yielding:

$$CT_{\max} = \begin{cases} \frac{1}{2}\left(\frac{n(n^2-1)}{12}\right) = \frac{n(n^2-1)}{24} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(\frac{n^3-n}{12} - \frac{n}{4}\right) = \frac{n^3-4n}{24} & \text{if } n \text{ is even} \end{cases}$$

which is exactly what needs to be proven. □

Furthermore, Kendall and Babington-Smith (1940) [13] have defined a measure of consistency for a directed graph dependent on the number of circular triads $CT$. This measure is called the coefficient of consistence and is denoted by $\zeta$:

$$\zeta = \begin{cases} 1 - \frac{24CT}{n(n^2-1)} & \text{if } n \text{ is odd} \\ 1 - \frac{24CT}{n(n^2-4)} & \text{if } n \text{ is even} \end{cases}$$

The coefficient of consistence displays the number of circular triads compared to the maximum number of circular triads. If the individual has no circular triads in his preferences, the value of $CT$ equals zero, which means that the coefficient of consistence $\zeta$ equals 1. As the number of circular triads increases, the value of $\zeta$ decreases. If the maximum number of circular triads is obtained, which is $\frac{n^3-n}{24}$ for $n$ odd and $\frac{n^3-4n}{24}$ for $n$ even, the coefficient of consistence $\zeta$ will be equal to zero.

However, one might wonder for which values of $\zeta$ the preferences of an individual can be regarded as too random. Since the coefficient of consistence $\zeta$ is a function of the number of circular triads, the significance of $\zeta$ can be determined by looking at the number of circular triads. For this, Kendall and Babington-Smith determined the number of circular triads that would occur in the preference of an individual when an individual would randomly judge a set of $n$ objects in a pairwise comparison experiment, where $2 \leq n \leq 7$. Then, for each number of objects $n$, the frequency $f$ for every number of circular triads has been written down in a table together with the probability that the number of circular triads will be obtained or exceeded. This table can be seen in figure 4.

*Frequency (f) of values of d and probability (P) that values will be attained or exceeded*

| Value of $d$ | $n=2$ | | $n=3$ | | $n=4$ | | $n=5$ | | $n=6$ | | $n=7$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $f$ | $P$ | $f$ | $P$ | $f$ | $P$ | $f$ | $P$ | $f$ | $P$ | $f$ | $P$ |
| 0 | 2 | 1·000 | 6 | 1·000 | 24 | 1·000 | 120 | 1·000 | 720 | 1·000 | 5,040 | 1·000 |
| 1 | | | 2 | 0·250 | 16 | ·625 | 120 | ·883 | 960 | ·978 | 8,400 | ·998 |
| 2 | | | | | 24 | ·375 | 240 | ·766 | 2,240 | ·949 | 21,840 | ·994 |
| 3 | | | | | | | 240 | ·531 | 2,880 | ·880 | 33,600 | ·983 |
| 4 | | | | | | | 280 | ·297 | 6,240 | ·792 | 75,600 | ·967 |
| 5 | | | | | | | 24 | ·023 | 3,648 | ·602 | 90,384 | ·931 |
| 6 | | | | | | | | | 8,640 | ·491 | 179,760 | ·888 |
| 7 | | | | | | | | | 4,800 | ·227 | 188,160 | ·802 |
| 8 | | | | | | | | | 2,640 | ·081 | 277,200 | ·713 |
| 9 | | | | | | | | | | | 280,560 | ·580 |
| 10 | | | | | | | | | | | 384,048 | ·447 |
| 11 | | | | | | | | | | | 244,160 | ·263 |
| 12 | | | | | | | | | | | 233,520 | ·147 |
| 13 | | | | | | | | | | | 72,240 | ·036 |
| 14 | | | | | | | | | | | 2,640 | ·001 |
| Total | 2 | — | 8 | — | 64 | — | 1,024 | — | 32,768 | — | 2,097,152 | — |

Figure 4: Table of frequencies of the number of circular triads for $n$ objects together with the probability that the number of circular triads will be attained or exceeded. Source: Kendall and Babington-Smith (1940) [13].

The numbers in this table were determined in the following way: when $n$ objects are compared, there are $2^{\binom{n}{2}}$ possible configurations of preferences. For example, when three objects, $A, B$ and $C$, are com-

pared, the total number of possible configurations equals $2^{\binom{3}{2}} = 8$. Of these eight possible configurations, six are rankings as there are $3! = 6$ ways to rank six objects. The other two configurations are circular triads: $A > B > C > A$ and $A > C > B > A$. Therefore, in the case of three objects, the probability of the preferences having zero circular triads equals $\frac{6}{8} = 0.75$ and the probability of one circular triad equals $\frac{2}{8} = 0.25$. This can also be seen in table 4. Now, the distribution of the number of circular triads when $n = 4$ is determined by going through the different possible configurations of preferences that arise when an extra vertex $D$ is added to the vertex set and determining in which cases extra circular triads would occur based on the distribution for $n = 3$. Therefore, the method used to determine the distribution of the number of circular triads consists of going from the distribution of $n$ objects to the distribution of $(n + 1)$ objects.

However, when exactly can the number of circular triads be regarded as being too large? For this, one can apply a hypothesis test. The null hypothesis $H_0$ of this test states that the individual does not demonstrate a clear preference, making his judgments at random.
Now, if one defines a significance level $\alpha$, the maximum allowable number of circular triads $CT_{\max}$ an individual can make such that he is regarded as being competent of making judgments is the largest number of circular triads $CT$ such that:

$$\mathbb{P}(CT \leq CT_{\max}) < \alpha.$$

For example, when $\alpha = 0.05$ and the number of objects equals six, the maximum number of circular triads allowed equals 1, as the probability that the number of circular triads is smaller than or equal to 2 equals $1 - 0.949 = 0.0501 > 0.05$. Still, Kendall and Babington-Smith [13] were only able to complete the table for a number of objects up to seven, since determining frequencies of the number of circular triads for a number of objects larger than seven by making use of combinatorics is computationally expensive.

Kendall (1955) [12] has shown that if the number of objects is larger than seven, a transformation can be applied on the number of circular triads $CT$ that follows an approximate $\chi^2$ distribution with $d$ degrees of freedom:

$$\chi^2 = \frac{8}{n-4}\left(\frac{1}{4}\binom{n}{3} - CT + \frac{1}{2}\right) + d$$

where

$$d = \frac{n(n-1)(n-2)}{(n-4)^2}.$$

Hence, in the case of more than seven objects a hypothesis test can be applied to the transformed number of circular triads $CT$ to determine whether the expert's preference contains too many circular triads, the null hypothesis being that the expert specified his preferences randomly. The proof of this statement relies on showing that the first four moments of the distribution of the triads coincide with the first four moments of the $\chi^2$ distribution after a transformation has been applied on the number of circular triads. By showing that the first four moments coincide, the transformed number of circular triads can be regarded as approximately $\chi^2$ distributed.

## 3.2 Tests for agreement within and between groups

Next to the inconsistencies in the preferences of an individual, it is also of interest to explore the similarities of preferences in a group of individuals. Does a group of individuals have a clear preference for the objects? Suppose that a sample of $m$ individuals has participated in a pairwise comparison experiment involving $n$ objects, making comparisons between all possible $\frac{n(n-1)}{2}$ pairs of objects. Let $f_{ij}$ be the elements of the resulting frequency matrix, which denotes the number of times that object $i$ was judged greater than object $j$. Note that $f_{ij}$ may consist of any number from 0 to $m$ and by symmetry: $f_{ji} = m - f_{ij}$.
Define

$$\Sigma = \sum_{i \neq j} \binom{f_{ij}}{2},$$

the summation taking place over all the $n(n-1)$ terms of the frequency matrix. $\Sigma$ is then equal to the sum of the number of agreements between pairs of individuals as it sums over all the numbers of possible pairs that can be made from the $f_{ij}$ individuals preferring object $i$ over object $j$.
In order to capture the degree of agreement within a sample, Kendall and Babington-Smith (1940) [13] have defined a coefficient of agreement, $u$, which depends on $\Sigma, n$ and $m$. The coefficient of agreement $u$ is defined follows:

$$u = \frac{2\Sigma}{\binom{m}{2}\binom{n}{2}} - 1$$

If the individuals are in complete agreement with one another, all the $m$ individuals would prefer the same object in each pair, which would result in the frequency matrix consisting of $\frac{n(n-1)}{2}$ elements equal to $m$ and the other $\frac{n(n-1)}{2}$ element equal to 0. In this case the value of $\Sigma$ would be equal to $\binom{n}{2}\binom{m}{2}$ and thus the coefficient of agreement:

$$u = 2\frac{\binom{m}{2}\binom{n}{2}}{\binom{m}{2}\binom{n}{2}} - 1 = 2 - 1 = 1$$

So when there's complete agreement between the individuals within the sample, the coefficient of agreement $u$ equals 1.

However, the more the elements $f_{ij}$ of the frequency matrix tend to deviate from the numbers $m$ and 0, the smaller $u$ becomes. The minimum number of agreements within the sample occurs when all the elements $f_{ij}$ equal $\frac{m}{2}$ for $m$ even or $\frac{m\pm1}{2}$ for $m$ odd, where in the case of $m$ odd $\frac{n(n-1)}{2}$ elements of the frequency matrix are equal to $\frac{m-1}{2}$ and the other $\frac{n(n-1)}{2}$ elements equal to $\frac{m+1}{2}$. In that case, the preferences of the sample are not in agreement with one another, therefore resulting in a coefficient of agreement $u = \frac{-1}{n-1}$ for $m$ even and $\frac{-1}{n}$ if $m$ is odd.

It is possible to test whether the coefficient of agreement $u$ is significant. Note that, as the coefficient of agreement is dependent on $\Sigma$, one could look at the probability that in the case of $m$ individuals and $n$ objects the value of $\Sigma$ would arise by chance if the preferences of all the individuals were assigned at random. For this purpose, Kendall and Babington-Smith [13] have determined the distributions of $\Sigma$ for various number of objects ($n \leq 8$) and small sample sizes ($m \leq 6$), when the preferences of the individuals in the sample would be assigned at random. This was done by making use of combinatorics. The significance of the value of $\Sigma$ can then be tested in the same way as in the case of the number of circular triads.
However, when the number of objects and sample size increases the probability that a certain value of $\Sigma$ occurs by chance is harder to identify by combinatorics [13]. In that case, Kendall (1955) [12] has shown that for larger values of $m$ and $n$ the statistic

$$\frac{4}{n-2}\left[\Sigma - \frac{1}{2}\binom{n}{2}\binom{m}{2}\frac{m-3}{m-2}\right]$$

is approximately $\chi^2$ distributed with

$$\binom{n}{2}\frac{m(m-1)}{(m-2)^2}$$

degrees of freedom. The proof of this statement is based on showing that the first four moments of $\Sigma$ around the origin comply with the first four moments of a transformed $\chi^2$ distribution, which, after some algebraic computation, yields the desired result.

Besides the agreement within a group, one can also test whether two groups are in agreement with one another when both groups have judged the same set of object in the experiment. In order to test for concordance between groups, one could make use of a non-parametric statistic. The only downside, however, is that no non-parametric tests for between-group concordance have been formulated for pairwise comparison experiments. There do exist several tests for measuring the significance of concordance between groups in ranking experiments [23]. One of these tests is Kraemer's test for between-group concordance. This test, invented by Kraemer (1981) [15], can be used to measure concordance between

$g$ groups, where $g \geq 2$, and is based on Kendall's coefficient of within-group concordance for a ranking experiment.

Kendall and Babington-Smith (1939) [14] have defined a measure of within-group concordance in ranking experiments in the following way: Suppose there are $m$ judges each ranking $n$ objects, where no ties are allowed. First define:

$$R_i = \sum_{j=1}^{m} r_{ij}$$

as the total sum of the ranks of object $i$ assigned by the $m$ judges, where $r_{ij}$ is the rank of object $i$ corresponding to judge $j$ and

$$\overline{R} = \frac{1}{n} \sum_{i=1}^{n} R_i$$

as the mean of the total ranks of the objects. Now, let:

$$A = \sum_{i=1}^{n} (R_i - \overline{R})^2$$

be the sum of squares of the deviations of the $R_i$ from the mean of the total ranks. Then, Kendall's coefficient of concordance is defined as:

$$W = \frac{12A}{m^2(n^3 - n)}$$

In the case of complete concordance within a group the coefficient of concordance $W$ will be equal to 1 and decreases when the degree of agreement diminishes [6].
Note that Kendall's coefficient of concordance is similar to the coefficient of agreement, as both measure the degree of agreement within a group of individuals. The main difference between the two is that the coefficient of concordance was invented especially for ranking problems, while the coefficient of agreement was originally formulated for pairwise comparisons. The coefficient of concordance can also be tested for significance. Siegel (1956) [28] has determined critical values for $S$ when the number of objects is between 3 and 7 and the sample size between 3 and 20. For larger $m$ and $n$, the statistic

$$W' = \frac{S}{\frac{1}{12}nm(m+1)}$$

is approximately $\chi^2$ distributed with $n - 1$ degrees of freedom [28].

Now, suppose that two samples of size $m_1$ and $m_2$ respectively rank the same set of $n$ objects, each sample having their own coefficient of concordance $W_i$. Kramer's measure for between-group concordance $W_B$ is defined as the ratio of the coefficient of concordance for the total sample, consisting of the two groups together, to the weighted averages of the within-group coefficients of concordance. In other words:

$$W_B = \frac{W_T}{W_W}$$

where

$$W_W = \frac{m_1 W_1 + m_2 W_2}{m_1 + m_2}$$

and $W_T$ is the coefficient of concordance for the total sample.

The weighted average of the coefficients of concordance $W_W$ acts as an upper bound for the total concordance. This means that if there's no within-group concordance, there can be no between-group concordance as well. The latter makes sense, as testing whether the groups are in agreement to one another is reasonable only when there is agreement within the groups [15].

The coefficient of between-group concordance $W_B$ is thus easily computed. However, which values of

$W_B$ can be regarded as significant? For this, Kraemer (1981) [15] has stated that under the null hypothesis of the groups not being in concordance,

$$\eta = \frac{W_B}{1 - W_B}$$

is approximately $\mathrm{F}_{(n-1),(n-1)}$ distributed.

Note that in the case of pairwise comparisons the objects are inherently ranked in the preferences of an individual by adding the rows of the preference matrix of the individual. For example, in the preference matrix in table 2 in Section 3.1, object $A_3$ is never preferred over any other object and therefore ends in the fifth place in the ranking. Objects $A_2$ and $A_5$ are both two times preferred over one of the other objects therefore managing a tie for third place in the ranking. The objects $A_1$ and $A_4$ are tied for first place as both objects are preferred three times over any of the other objects. In this case the ranking of the individual contains ties because of the fact that intransitive preferences are present. The tests for concordance between groups in ranking experiments, however, do not allow ties in the rankings. This would imply that the concordance between groups in pairwise comparison experiments can be tested, but only if the preferences of all the individuals in the sample are without circular triads, as in that case the preferences of individuals can be expressed as a ranking without ties. This implies that if one wants to test for concordance between two groups in a pairwise comparison experiment, all the individuals having circular triads in their preferences will be dropped from the sample. If, however, the vast majority of the individuals in the sample have intransitive preferences, the remaining transitive sample may be of a too small size to be representative for the population. For example, when two samples of size 50 are participating in a pairwise comparison experiment and in both samples 90 percent of the individuals has at least one circular triad in their preferences, the remaining sample size will be five for each group. This sample size is too small to be a reliable representation of the preferences of the population.

Therefore, alternatives are needed in the case of circular triads. It is difficult to determine whether two groups are in concordance with one another by merely looking at the rankings of the frequency matrices or the proportion matrices of the two groups. In the most extreme case, when the proportion matrices are exactly the same or when the rankings of the frequency matrices coincide, it can be concluded that the groups are in concordance with one another. However, when this is not the case, between-group concordance is harder to detect.

A method that might serve as an alternative in the case of circular triads would be the following: Given two proportion matrices, $P_1$ and $P_2$, of two groups with sample size $m_1$ and $m_2$ respectively, one first determines the best case of Thurstone's model for proportion matrix $P_1$ by making use of the AIC. Next, 500 matrices are simulated from the distribution given by the estimated parameters for the first sample. For each simulated matrix, compute the Mean Square Error (MSE) of the matrix with both $P_1$ and $P_2$. This will yield 500 MSE's for each matrix. The concordance between the groups can be determined by testing whether the MSE's of both matrices are from the same distribution by making use of a two-sample Kolmogorov Smirnov test. The Kolmogorov-Smirnov test uses the supremum of the absolute difference between the empirical distribution functions as the test statistic, yielding significant results if the difference becomes too large. If the MSE's turn out to be from the same distribution, it can be stated that the matrices of the two groups are approximately the same as well, hence the groups are in concordance with one another. This method, however, has never been tested before and it is thus unsure whether it will perform as desired. This will be tested in the simulation study in Section 4.

# 4 Simulation study

The different cases of Thurstone's Law of Comparative Judgment, that encompass Thurstone's Pairwise Comparison Model, will primarily be the guidelines in the simulation study, which will consist of four parts. The results of this simulation study are of importance for the data study. In the first two parts of the simulation study, the influence of the spacing of the scale values and variances on the accuracy of the estimation of the parameters in the different cases will be investigated. Furthermore, in the second part of the simulation study, the two-step optimization process explained in Section 2.3.2 will be tested for cases II and III. Does the two-step optimization process always yield an accurate estimation for the parameters? The third part of the simulation study explores the performance of the AIC. It is known that the AIC is capable of giving back the best fitting model for a data set, but it has not been tested whether the AIC can be used for determining the best model in the case of Thurstone's Pairwise Comparison Model and how this is related to the sample size and number of objects. It is of interest to see whether the best model is capable of retrieving the true rankings. Next to that, the relationship between the coefficient of agreement and the choice of the AIC will be investigated as well. In the fourth and final part of the simulation study, the proposed test for between-group concordance based on the MSE's of the matrices will be examined.

## 4.1 Case V

Case V is the most simple formulation of the Law of Comparative Judgment, assuming equal variances and zero correlations. In this simulation study the influence of the spacing of the scale values on the quality of the fit has been examined. In order to investigate this, both equidistant spacing and random spacing have been considered. In this case, the common variance $\sigma^2$ has been set equal to 1, so that the variance restriction is satisfied.
Note, however, as other variances restrictions can be applied as well, the common variance does not necessarily have to be equal to 1. Another simulation study, where the influence of the size of the common variance and the effect of the placement of the scale values on the accuracy of the estimations has been examined, has been executed as well. The results of this study can be found in appendix A.1, since the outcome didn't seem to bring about new information.

In each of the cases, the number of objects is equal to six. This number has been chosen based upon that six is the smallest number of objects such that every case of Thurstone can be fitted on the data. Moreover, the means of the objects, denoted by $S_i$, are written in vector notation and sum up to zero to satisfy the scale restriction. Next to that, the sample size is chosen equal to 1000. During the simulation, the process of the simulation of the proportion matrix is repeated twenty-five times so that an average can be obtained and the responses of the experts can not be dedicated to coincidence, such as the risk of having a biased sample.

In the case of equidistant spacing, five different configurations have been deployed which can be seen in table 3, the distance between the scale values ranging from 0.2 to 4.

| **S** | Mean vector | Distance |
|---|---|---|
| $\mathbf{S_1}$ | (-0.5, -0.3, -0.1, 0.1, 0.3, 0.5) | 0.2 |
| $\mathbf{S_2}$ | (-1.25, -0.75, -0.25, 0.25, 0.75, 1.25) | 0.5 |
| $\mathbf{S_3}$ | (-2.5, -1.5, -0.5, 0.5, 1.5, 2.5) | 1 |
| $\mathbf{S_4}$ | (-5, -3, -1, 1, 3, 5) | 2 |
| $\mathbf{S_5}$ | (-10, -6, -2, 2, 6, 10) | 4 |

Table 3: The five equidistantly chosen mean vectors used in the simulation.

What happens with the estimation of the means when the distance between the means increases? Because the means are all differently spaced, it is impossible to compare them by just plotting all the estimates and the real means in one figure. In order to really compare the accuracy of the estimation, the estimated

scale values all have been transformed onto a $[0, 1]$-scale by using the following transformation:

$$\widehat{\mathbf{T}_{\mathbf{i}j}} = \frac{\widehat{\mathbf{S}_{\mathbf{i}j}} - \max_j(\mathbf{S}_{\mathbf{i}j})}{\max_j(\mathbf{S}_{\mathbf{i}j}) - \min_j(\mathbf{S}_{\mathbf{i}j})}$$

where $\widehat{\mathbf{T}_{\mathbf{i}j}}$ is the transformed $j$-th index of the estimate of the mean vector, $\widehat{\mathbf{S}_{\mathbf{i}}}$, $\widehat{\mathbf{S}_{\mathbf{i}j}}$ equals the $j$-th index of the estimate of the mean vector $\widehat{\mathbf{S}_{\mathbf{i}}}$ and $\max_j(\mathbf{S}_{\mathbf{i}j})$ and $\min_j(\mathbf{S}_{\mathbf{i}j})$ are respectively the minimum and maximum values of the true mean vector $\mathbf{S}_{\mathbf{i}}$. A perfect estimation of the scale values yields, according to this transformation, values that lie perfectly on the line $y = x$. In the case of six objects, this means that the estimated scale values are equal to: $(0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1)$.
However, as the scale values may also be over- or underestimated, the transformed estimates of the scale values are not necessarily completely contained in the $[0, 1]$ interval. Nevertheless, this is not a big problem as the transformation still reflects the accuracy of the estimates such that the influence of the distances between the scale values on the estimates can be compared simultaneously. The plot of the transformed estimates for the different mean vectors given in table 3 together with the line $y = x$ displaying the perfect estimation can be found in figure 5.

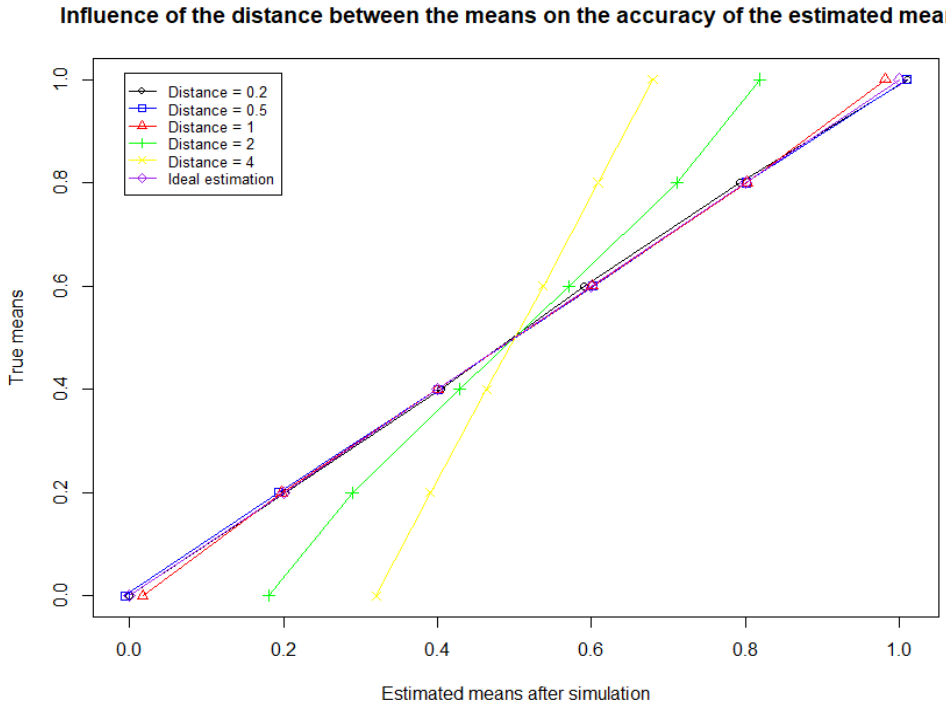



Figure 5: Influence of the distance between the means on the accuracy of the estimated means, where the estimated means are transformed to a $[0, 1]$ interval.

From figure 5, it can be seen that the accuracy of the estimation decreases when the distance between the means increases. This is confirmed by figure 6 as well, where the sum of squared errors of the estimates increase when the distance between the means increases.

This result seems counter-intuitive at first sight. As the common variance of the objects is equal to 1 in all cases, the overlap of the distributions of the objects becomes smaller when the distance between the scale values increases. Less overlap between the distributions means that the probability of making an inconsistent choice becomes smaller and according to Siraj et al. (2015) [29] consistent judgments result in a better estimate for the parameters.

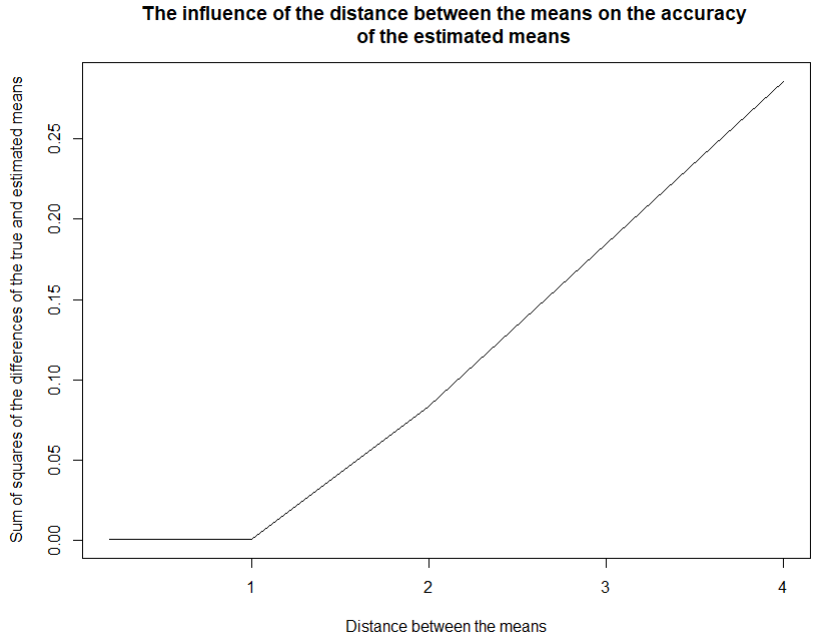



Figure 6: The distance between the means plotted against the sum of squared errors of the estimates of the scale values.

The fact that the estimates decrease in accuracy has to do with the choice of the sample size in the simulation. In this case, the sample size has been chosen equal to 1000. If the means of the objects are widely spaced such that the distributions have no overlap at all, every judge or individual automatically chooses that object in the pair that has a higher scale value. In that case, the proportion matrix would consist of zeros above the diagonal and ones below the diagonal. However, as explained in Section 2.3, zeros and ones are not tolerated in the proportion matrix. Therefore the proportion matrix would consist, in the case of a sample size of 1000, of elements being equal to 0.001 above the diagonal and 0.999 below the diagonal. In the case of six objects, this gives the following proportion matrix $P$:

$$P = \begin{pmatrix} 0.5 & 0.001 & 0.001 & 0.001 & 0.001 & 0.001 \\ 0.999 & 0.5 & 0.001 & 0.001 & 0.001 & 0.001 \\ 0.999 & 0.999 & 0.5 & 0.001 & 0.001 & 0.001 \\ 0.999 & 0.999 & 0.999 & 0.5 & 0.001 & 0.001 \\ 0.999 & 0.999 & 0.999 & 0.999 & 0.5 & 0.001 \\ 0.999 & 0.999 & 0.999 & 0.999 & 0.999 & 0.5 \end{pmatrix}$$

Now, applying the inverse of the standard normal cumulative distribution on the elements in the proportion matrix and applying the formula for the estimations of the scale values gives that, in the case of six objects, the maximum distance between the scale values of the least and most preferred object is equal to:

$$\frac{5\sqrt{2}\Phi^{-1}(0.999)}{6} - \frac{5\sqrt{2}\Phi^{-1}(0.001)}{6} = \sqrt{2}\left(\frac{15.45116}{6} - \frac{-15.45116}{6}\right) \approx 7.284.$$

So in order to obtain an adequate estimate of the scale values, when the sample size is chosen to be 1000, the distance between the scale values of the least and most preferred object may be 7.284 at most. The fact that the maximum distance can not be larger than 7.284 explains why the transformed estimations of the scale values deviated from the real scale values when the distance between the scale values is 2 or 4.

If the distance between the scale values of the most and least preferred objects is greater than 7.284, the estimates of the scale values are not reliable as they simply can not be attained. This has to do with the fact that the inverse of the standard normal deviate has a limiting effect. When $x$ goes to zero, the

inverse converges to infinity:

$$\lim_{x \to 0} \Phi^{-1}(x) = -\infty.$$

However, it converges slowly to infinity as the standard normal distribution converges exponentially to zero when $y$ converges to minus infinity. This means that, for example, in order to get a reliable estimate for the scale values when the distance of the scale values of the most and least preferred objects is 20 and six objects are investigated, the sample size of the experiment must at least be equal to $10^{20}$ experts, which is about the size of the population of the entire unknown galaxy.
The limiting effect of the standard normal distribution in relation to the sample size when six objects are investigated can be seen in table 4.

| Sample size | 10 | $10^2$ | $10^3$ | $10^4$ | $10^5$ | $10^6$ | $10^{10}$ | $10^{15}$ | $10^{20}$ | $10^{25}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Max. distance | 3.021 | 5.843 | 7.284 | 8.766 | 10.052 | 11.204 | 14.994 | 18.718 | 21.832 | 24.561 |

Table 4: The limiting effect of the sample size on the maximum distance between the scale values of the most and least preferred objects, when six objects are involved in the experiment and the common variance $\sigma^2 = 1$.

Note that the limiting effect also occurs in the case when a general number of objects is used in a pairwise comparison experiment. The maximum interval distance that can be obtained in the estimation therefore not only depends on the sample size, but also on the number of objects that is considered. Furthermore, it should be noted that when the sample size decreases, the accuracy of the scale values decreases as well, since the proportion matrix less delineates the true distribution in the case of a smaller sample size [2].

When the means of the objects are randomly spaced, the results do not differ significantly from the case where the scale values were equidistantly spaced indicating that the distance between the means doesn't seem to influence the accuracy of the estimates. The model seems to perform quite well, giving reasonable estimates as long as the interval distance of the scale values do not exceed 7.284. A simulation has been executed with three different randomly spaced vectors, where the distance between scale values of the outermost and innermost objects increases. The mean vectors used for the simulation, the length of the interval where the means are defined and the sum of squared errors of the estimates are reported in table 5.

| $S$ | Mean vector | Distance | Sum of squares of the errors |
|---|---|---|---|
| $\mathbf{S_6}$ | (-0.5, -0.3, -0.2, 0.1, 0.2, 0.7) | 1.2 | $2.571 \cdot 10^{-4}$ |
| $\mathbf{S_7}$ | (-2.3, -1.5, -0.6, -0.1, 1.8, 2.7) | 5 | 0.023 |
| $\mathbf{S_8}$ | (-5.3, -3.5, -0.8, 0.9, 3.8, 4.9) | 10.2 | 11.076 |

Table 5: The randomly spaced mean vectors employed in the simulation, the distance between the smallest and largest mean and the sum of squares of the errors of the estimation.

Looking at the sum of squares, it is obvious that the sum of squares increases as the distance of the interval where the means are defined increases for the same reason as in the case of equidistantly spaced means. Therefore, randomly spaced means do not seem to affect the performance of the model in a peculiar way, thus having a similar effect on the accuracy of the estimation.

In conclusion, it can be stated that one has to be careful that the scale values of the most and least preferred objects are close enough to each other, the maximum distance determined by the number of objects and the sample size, so that an accurate estimation can be made. However, making sure that the distance is small enough can not be taken care of, as the data gathered from the pairwise comparison experiment can not be influenced beforehand. Nevertheless, the discovery in this part of the study still brings about a useful aspect. When the proportion matrix has elements close to zero and one, it can be stated that the means are spaced far-away from one another and the resulting estimation might be inaccurate.

## 4.2 Case III and Case II

The primary objective of this part of the simulation study is to investigate the performance of the two-step optimization process proposed for cases III and II of Thurstone's model. This is done by comparing

the accuracy of the estimates of parameters obtained from the two-step optimization with the estimates of the parameters obtained from using the true parameters as initial condition for the cobyla. Next to that, it is also of interest to examine how the variances and scale values together influence the accuracy of the scale values.

For Case III of Thurstone's Law of Comparative Judgment, a total number of $2n$ parameters have to be estimated: $n$ means and $n$ variances. As the variances in this case differ from one another one could wonder how the variability in variance together with the mean distance affects the accuracy of the estimates of both the means and variances. In order to investigate this, two different variance vectors have been considered, each differing in variability. These vectors can be seen in table 6.

| $\boldsymbol{\sigma^2}$ | Variance vector | Degree of variability in variance |
|---|---|---|
| $\boldsymbol{\sigma_1^2}$ | (0.7, 0.8, 0.9, 1.1, 1.2, 1.3) | Low |
| $\boldsymbol{\sigma_2^2}$ | (0.01, 0.07, 0.1, 1.5, 1.4, 2.92) | High |

Table 6: The two variance vectors used for simulation in Case III.

Next to that, the equidistantly mean vectors $\mathbf{S_1}$ and $\mathbf{S_3}$ from table 3 have been used as mean vectors in the simulation to examine how the distance between the means interacts with the variance on the precision of the estimates. These two vectors have been chosen on the basis of the results of the simulation of Case V, where it was discovered that the mean distance can not be too large, the distance depending on the sample size and number of objects, if accurate estimations are desired. Furthermore, the simulation of Case V has pointed out that the placing of the means are irrelevant on the precision of the estimates. Therefore, $\mathbf{S_1}$ and $\mathbf{S_3}$ have been chosen as the mean vectors for the simulation.

The sample size for the simulation is again chosen equal to 1000. A simulation involving a smaller sample size of 100 has been executed as well but will not be discussed here, since this simulation yielded similar results, the only difference being that the estimations of the parameters were less accurate. The latter occurs because of the earlier mentioned fact that a smaller sample size decreases the quality of the approximation of the estimates. The Quasi-Newton method used in the first step of the two-step optimization process requires, just like the cobyla, starting parameters. In this case it has been chosen to set the starting parameters of the Quasi-Newton method equal to zero for the means and one for the variances such that they comply to the parameter restrictions. The simulation is replicated twenty-five times resulting in twenty-five estimations for every parameter. This decision was made on computational grounds, as the simulation would otherwise take very long to complete. Since the number of samples used in the simulation was small, it was decided not to make use of uncertainty bounds or boxplots, but to investigate the accuracy of the estimations based on the means and the minimum and maximum estimates of the twenty-five estimations. In order to compare the accuracy of both optimization routines, the minimum and maximum estimates together with the mean of the twenty-five estimations for the two routines were plotted side by side in one figure. At the same time, the plots obtained also capture the influence of the variability in variance and mean distance on the accuracy of the estimates.

First, the influence of the variability in variance on the approximation of the estimates when the mean distance is small has been explored. The plots obtained from this simulation yielded the same outcome for the accuracy of the estimations for both low and high degrees of variability in variance. Therefore, only the results in the case of a high variability in variance will be discussed. The plots of a low variability in variance can be found in appendix A.4. Figures 7 and 8 show the minimum, maximum and mean estimations of the means and variances for both optimization approaches in the case of a small mean distance and high variability in variance.

From figures 7 and 8 it can be deduced that the estimations of the parameters are quite accurate for both optimization approaches. The estimations of the parameters are more accurate when the true parameters are used as the initial condition for the cobyla function. This primarily holds for the estimations of the variances, as the minimum and maximum estimation deviate a lot more from the true parameters in the case of the two-step optimization process. Nevertheless, the estimations obtained from the two-step optimization process are still sufficiently adequate, seeing as the means of the estimations are close to the

true parameters. Next to that, figures 7 and 8 show that the higher the variance of an object, the more the minimum and maximum estimates of the parameters deviate from the mean estimate. This can be explained by the fact that in the case of a larger variance, the preferences of the individuals vary more, which results in more variation in the row of the proportion matrix corresponding to the object with a larger variance compared to the rows belonging to the objects having smaller variances. As a result, the estimations of the parameters of the object with a larger variance also show more variation.

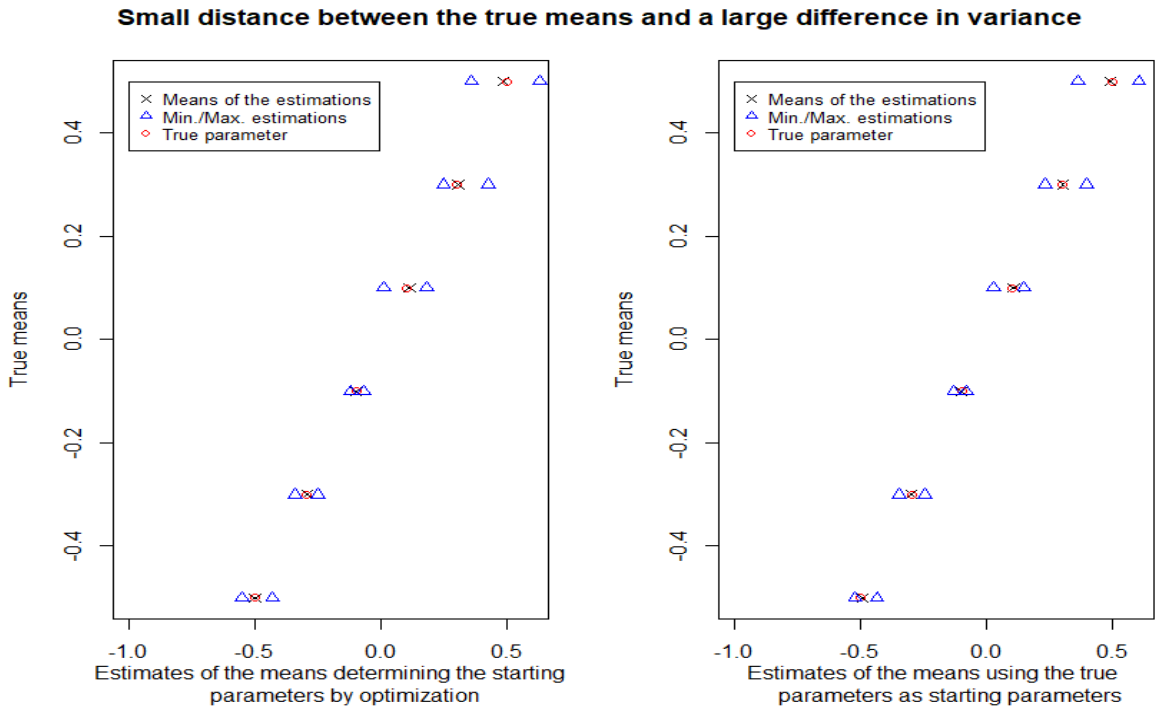



Figure 7: The accuracy of the estimations of the means for both optimization routines, when the distance between the means is small and the variability in variance is high.

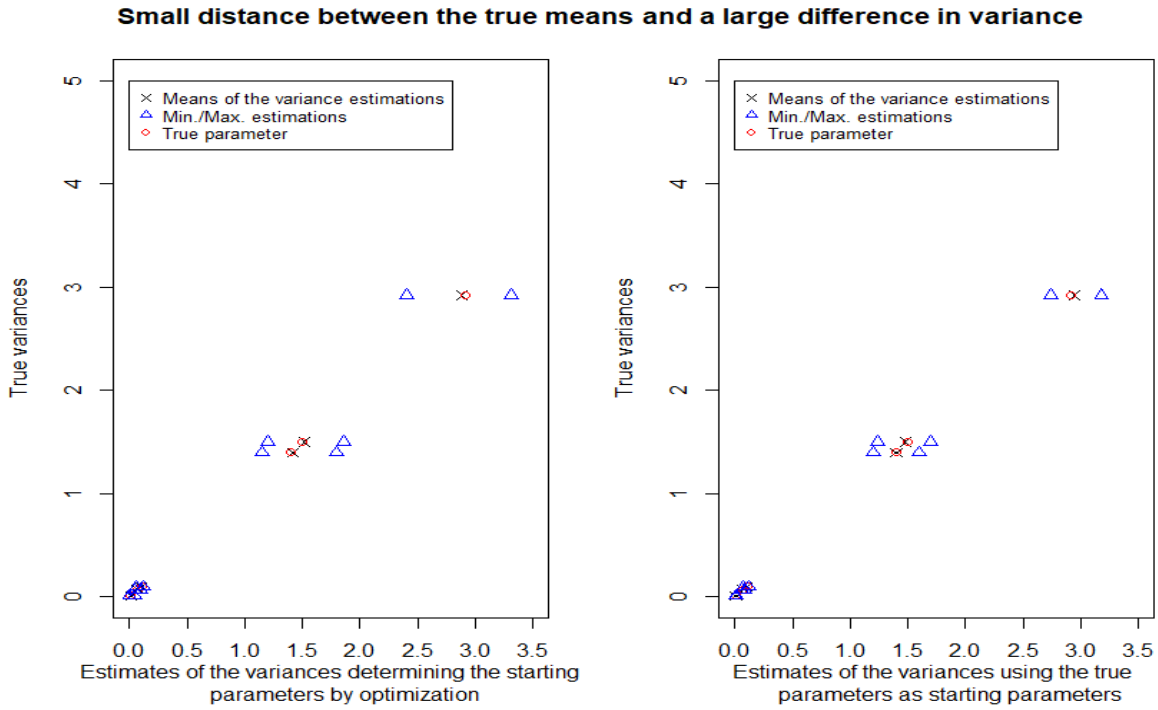



Figure 8: The accuracy of the estimations of the variances for both optimization routines, when the distance between the means is small and the variability in variance is high.

As the estimations of a low variability in variance yielded the same results for the accuracy of the estimations, it can be concluded that in the case of a small mean distance the variability in variance of the objects is not of influence on the estimations of the parameters.

In the case of a larger mean distance, the estimations of the parameters decreased in accuracy when compared to a small mean distance. This primarily concerned the variances, since the accuracy of the estimated means were still adequate. As the same outcome was obtained for both high and low variability in variance, only the plots obtained of a large mean distance and low variability in variance will be displayed, the other plots can be found in appendix A.4. Figures 9 and 10 show the estimates of the means and variances in the case of a large mean distance and low variability in variance.

Looking at figures 9 and 10, the estimations of the means appear to be quite precise for both the two-step optimization as the optimization using the true parameters as initial condition. The estimations of the means are, just as in the case of a small mean distance, slightly more accurate when the true parameters are used as the initial condition. The estimations of the variances, however, are not adequately estimated in both cases. When the true parameters are used as the initial condition, the estimations of the variances are not inline with the true parameters and the order of the variances are switched. This effect worsens in the case of the two-step optimization process, as the mean of the estimates for the smallest variance is the largest mean of all the means of the estimates for the variances. The latter is also detected in the estimations of the mean of the object with the smallest variance. In figure 9 it can be seen that, for both optimization routines, the minimum and maximum estimate of the object with the smallest variance deviate the most from the mean of the estimations. The reason why the variances are not adequately estimated might be due to the fact that, as the means are widely spaced from one another, the variances are not that much of an influence on the judgments of individuals, since there's almost no overlap between the distributions. As the variances are barely of influence on the preferences, they are less apparent in the proportion matrix which results in a poorer estimation for the variances.

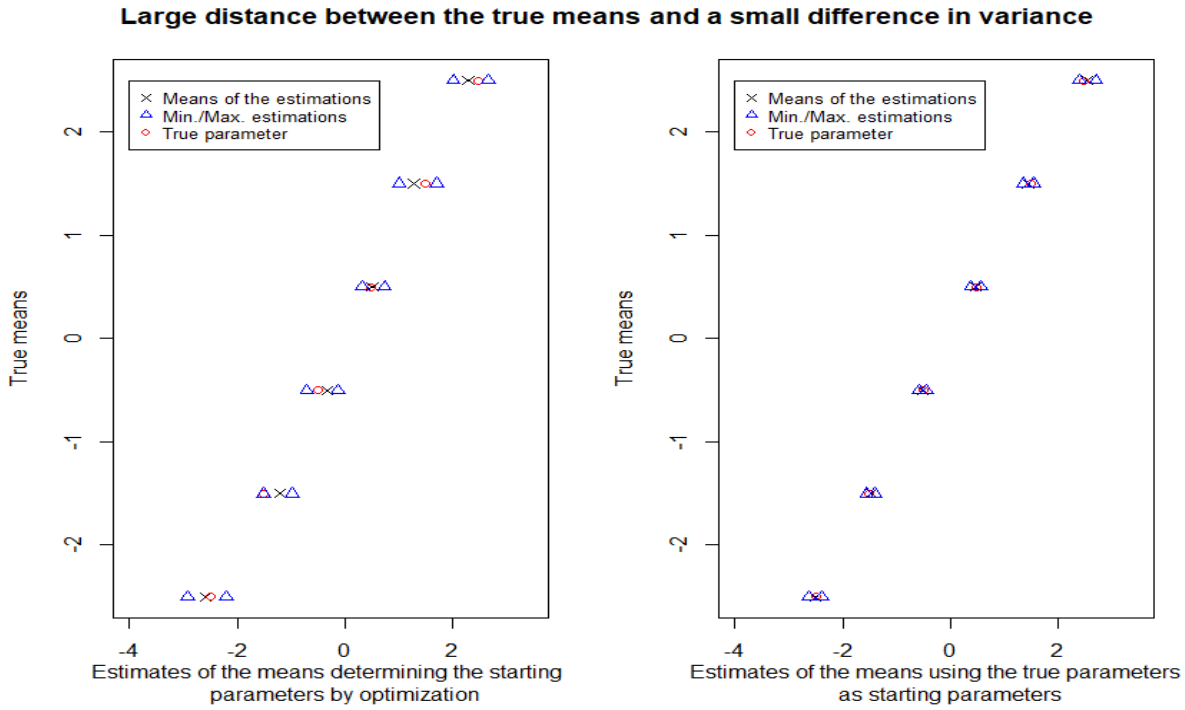



Figure 9: The accuracy of the estimations of the means for both optimization routines, when the distance between the means is large and the variability in variance is low.

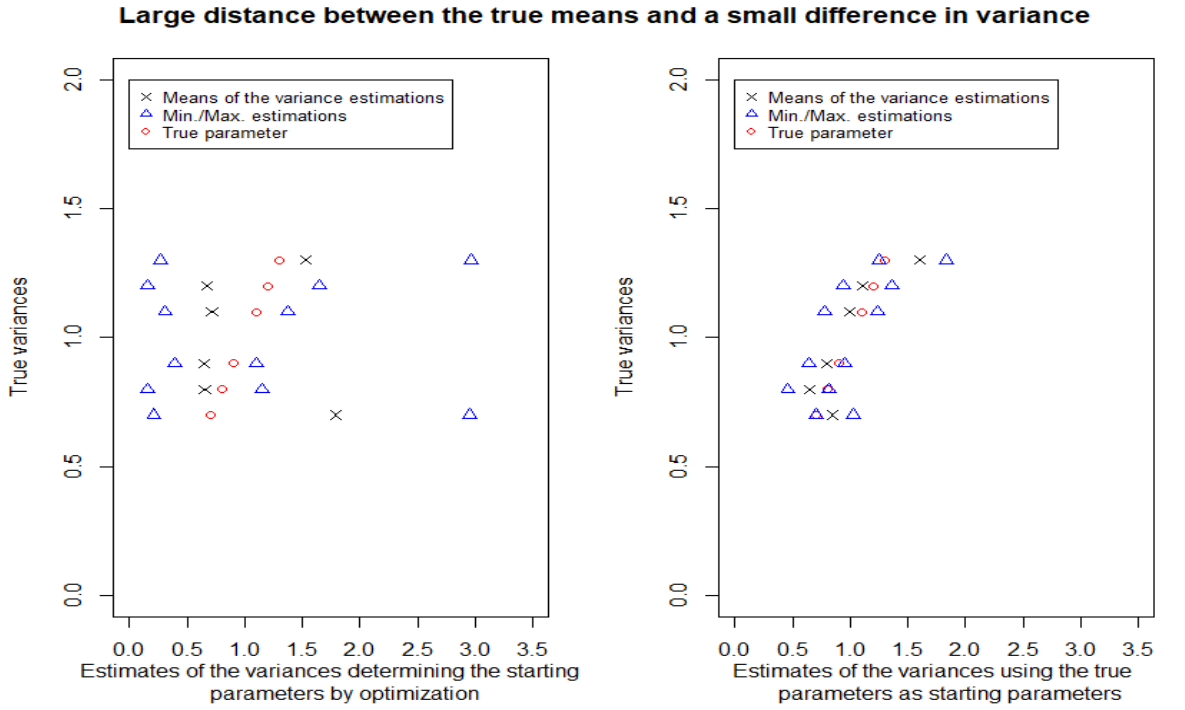



Figure 10: The influence of the variability in variance on the accuracy of the estimations of the variance, when the distance between the means is large and the variability in variance is low.

In conclusion, it can be stated that the two-step optimization process yields satisfactory results for Case III, giving accurate estimations for the parameters when the distance between the means is small. When the distance between the means increases, the accuracy of the estimations decreases: especially the estimations of the variances become unwieldy. The latter not only holds for the two-step optimization process, but also for the optimization using the true parameters as initial condition. Thus when the distance between the means is large, the variances are in general not adequately captured. Therefore, it might be better not to allow for different variances in the model when the distance between the means is large, hence preferring Case V.

Next to Case III, the performance of the two-step optimization process for Case II has been tested as well in the simulation study. For this, the same setting as in Case III has been employed, the only difference being that the correlation $\rho$ was added to the equation and varied between $-0.1$ and 0.9. The results obtained from this simulation study were, unfortunately, inadequate. When the true parameters were used as the initial condition for the cobyla, the optimization process is capable of accurately estimating the means, variances and correlation coefficient. The two step optimization process, however, seems to have problems minimizing the objective function. The estimations obtained from this simulation study were poor. For instance, when the correlation coefficient was set equal to 0.9 and 25 simulations were performed, the estimates for the parameter of the correlation coefficient varied anywhere between $-0.16$ and 0.95. Because of the fact that the correlation coefficient was not effectively captured, the estimations of the means and variances were poor as well. In order to solve this, several other unconstrained optimization routines for the first step in the optimization process were attempted as well, such as Particle Swarm Optimization, Nelder-Mead and Hooke-Jeeves. However, none of them were successful, all yielding the same results as in the case of Quasi-Newton. Hence, it may be questioned whether a two-step optimization is a suitable method for solving the parameters for Case II.

## 4.3 The performance of the AIC and the relationship of the AIC with the coefficient of agreement

The main interest of this section is to test whether the AIC is a suitable method to determine the best case of Thurstone's Pairwise Comparison Model for a given data set. For this, the performance of the AIC was investigated for both the two-step optimization process and the situation where the true distribution is known. As mentioned in Section 2.4, it is of interest to examine whether the best model is capable of retrieving the true ranking of the objects and how this is influenced by the sample size and number of objects. In order to investigate this, 200 proportion matrices were simulated from several multivariate normal distributions. Then, the best model was determined by the AIC for each proportion matrix, after which the Mean Square Errors (MSE's) of the parameter rankings from the true rankings were computed. Next to that, the rankings of the matrix, which are obtained by adding the rows from the frequency matrix, were computed as well, together with the number of circular triads present in the preferences and the coefficient of agreement. The matrix rankings were compared to the parameter rankings for each proportion matrix in order to investigate the relationship between the two. Furthermore, the relationship between the coefficient of agreement and the outcome of the AIC was examined as well. Before continuing it should be noted that it was discovered that the number of circular triads was always equal to zero in the simulation study. This had to do with the fact that the values corresponding to the objects were drawn from the multivariate distribution at once, making the preferences consistently transitive.

The estimation part of the simulation study has shown that the parameters are quite accurately estimated in all three cases of Thurstone's model when the distribution of the parameters is known, while the two-step optimization is capable of only adequately capturing the estimations of the parameters for cases V and III. Therefore, only Case III and V were included in the AIC procedure for the two-step optimization process, as it would not make sense to consider a model that is not capable of estimating the parameters as a possible candidate for the best model.

The study design was approximately the same for both studies: the number of objects investigated were equal to six, seven, ten and fifteen and the sample size varied between 10, 50, 100 and 200. Fur-

thermore, the same mean and variance vectors were used in the two studies, the only difference being that correlation was not included in the two-step optimization process. The results of the performance of the AIC in the two-step optimization process will be discussed, as these results are more relevant for the data study since the true parameters are not known in that case. Nevertheless, some relevant results of the other AIC simulation study concerning the true parameters as initial condition in the optimization will sometimes be briefly mentioned as well. The latter yielded approximately the same results, the main difference being that the correlation coefficient was of influence as well.

First, the performance of the AIC in the case of six objects was investigated. The two different mean and variance vectors used in the simulation are tabulated in table 7.

| Mean vector | Variance vector |
|---|---|
| $\mathbf{S_{6A}}$ = (-1.1, -0.6, -0.1, 0.3, 0.6, 0.9) | $\boldsymbol{\sigma^2_{6eq}}$ = (1.0,1.0, 1.0, 1.0, 1.0, 1.0) |
| $\mathbf{S_{6B}}$ = (-0.9, -0.5, -0.1, 0.2, 0.5, 0.8) | $\boldsymbol{\sigma^2_{6uneq}}$(0.5, 1.5, 0.8, 1.2, 1.1, 0.9) |

Table 7: The two mean and variance vectors used in the simulation to investigate the performance of the AIC in the case of six objects.

Note that the mean vectors are chosen such that the resulting parameters can be accurately estimated for all sample sizes. Furthermore, in the simulation study involving Case II, the correlation was varied between $-0.1$ and 0.9.

As the results of the performance of the AIC for the two mean vectors were similar to one another, only the results obtained from the first mean vector will be discussed.

Table 8 displays the outcome of the AIC for the first mean vector and unequal variances for different sample sizes in the case of the two-step optimization process.

| Sample size | 10 | 50 | 100 | 200 |
|---|---|---|---|---|
| V | 200 | 142 | 0 | 0 |
| III | 0 | 58 | 200 | 200 |

Table 8: The outcome of the AIC for the first mean vector, unequal variances and different sample sizes.

The question that now arises is: How many of the rankings of the best model according to the AIC are correct? If the sample size is equal to 10, the result is disastrous. Only 22 percent of the 200 rankings of the parameters coincide with the true rankings. Figure 11 shows the MSE's of the rankings of the parameters with the true rankings, together with the MSE's of the rankings of the matrices.

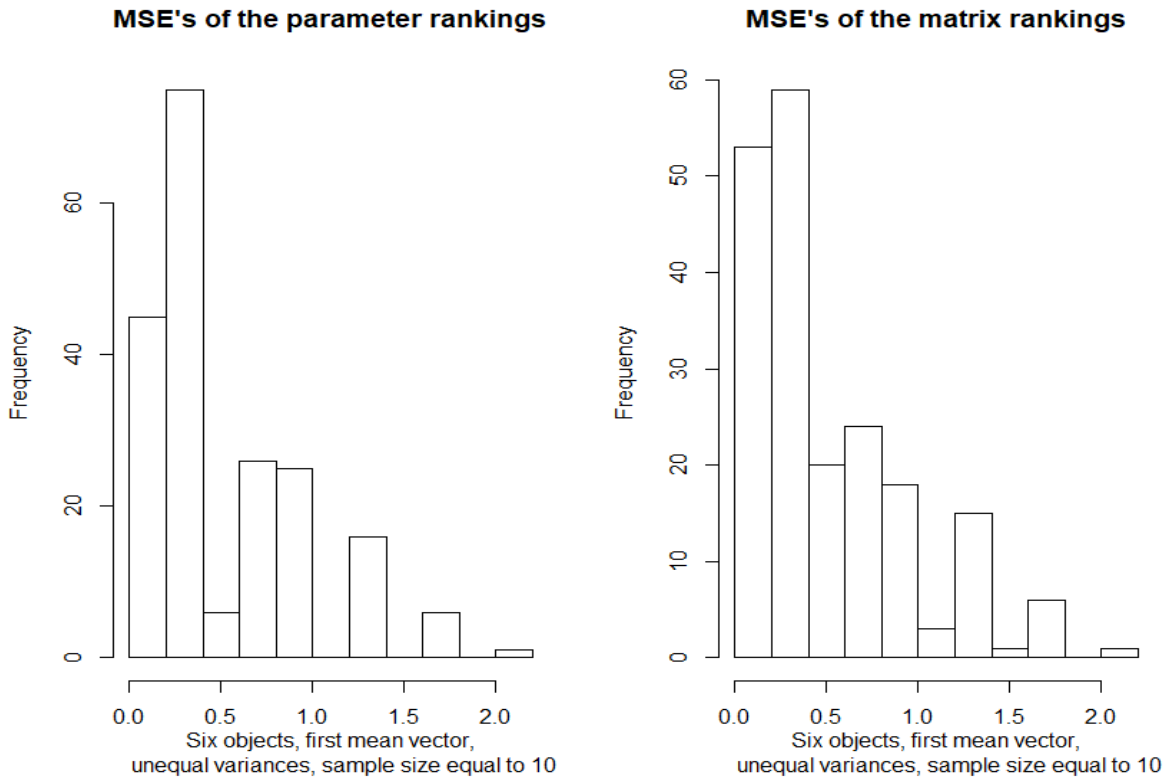



Figure 11: MSE's of the parameter rankings and matrix rankings in the case of unequal variances and a sample size equal to 10.

Looking at figure 11, it can be deduced that the performance of the AIC is far from optimal in the case of a small sample size since the MSE's of the parameter rankings are quite large. This, however has nothing to do with the AIC not being capable of determining the best model. The MSE's of the parameter rankings of the model not recommended by the AIC are large as well. The cause of this is simply that the sample size is too small to accurately estimate the parameters, which results in the rankings of the parameters getting mixed up. This is further confirmed by the MSE's of the matrix rankings. These are large as well indicating that a small sample size does not properly reflect the underlying distribution. In the case of correlation being included, the same conclusions were drawn. However, it should be mentioned that the MSE's of the rankings decreased as the correlation increased. This has to do with the fact that the preference of the sample is more in compliance with the ranking of the distribution when the correlation is large, since, as explained in Section 2.1, the probability that an object with a smaller mean is preferred over an object with a larger mean decreases when the correlation increases. The latter does not only hold for a small sample size, but applies for general sample sizes as well.

When the sample size increases to 50, the AIC is performing better. Out of the 142 times the AIC recommended Case V, 113 parameter rankings coincided with the true ranking, while for Case III 40 parameter rankings of the total 58 were identical to the true ranking. This is a significant improvement. The majority of the MSE's that were not equal to zero, equaled $\frac{1}{3}$, which implied that, as six objects were considered, only 2 objects adjacent to one another in the ranking were shifted. This means that, even though the ranking of the parameters did not exactly coincide with the true ranking, the ranking was still close to the latter. Furthermore, for both cases, the MSE's of the rankings of the parameters of the model that was not recommended by the AIC were slightly larger. This implies that the best model according to the AIC actually is the best model, since the parameter rankings of the AIC coincide more often with the true rankings. Figure 12 displays the MSE's of the recommended cases III and V together with the MSE's of the models that were not recommended.

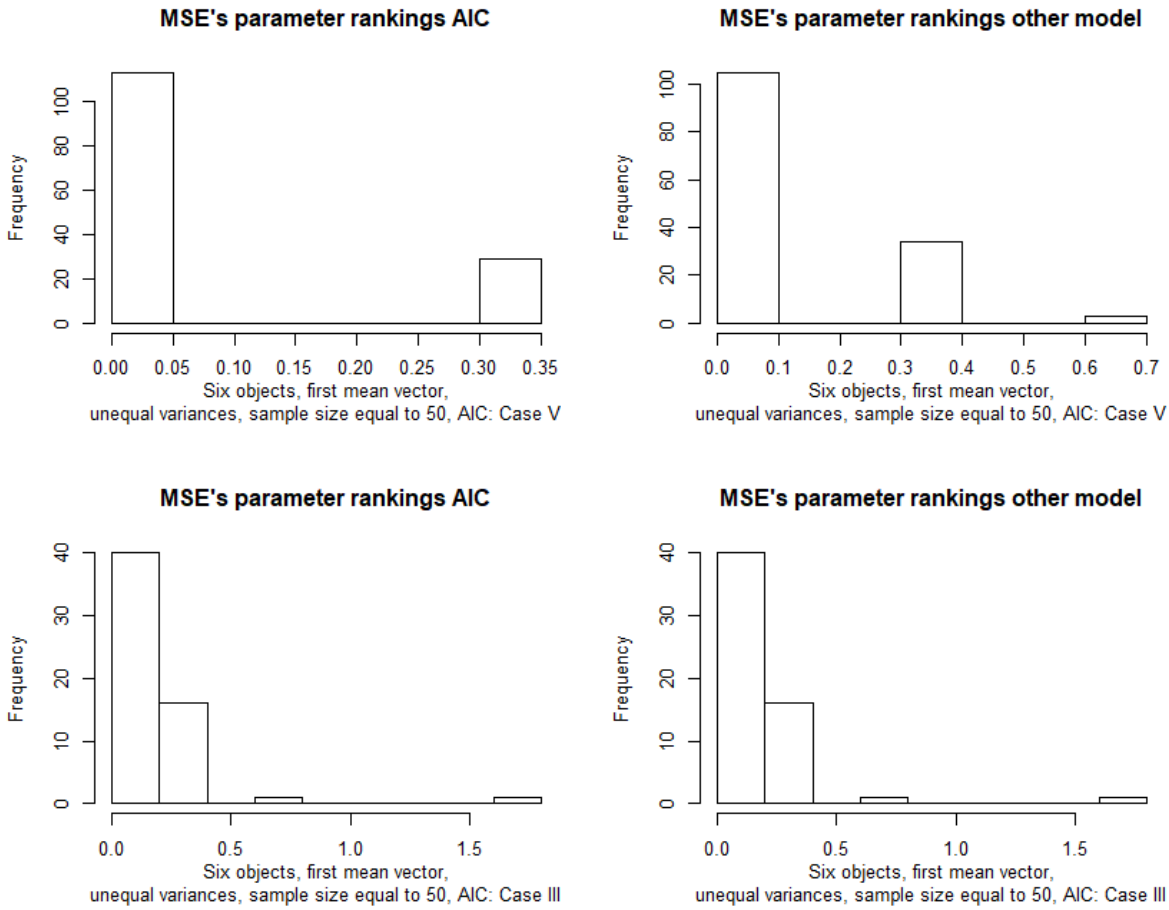



Figure 12: MSE's of the parameter rankings of the best model according to the AIC and the model not recommended by the AIC for both cases in the case of unequal variances and sample size equal to 50.

For a sample size equal to 100 or 200, the rankings of the parameters of the best model almost always coincide with the true ranking. For example, when the sample size equals 200, only 1 of the 200 parameter rankings is not identical to the true ranking. It makes sense that the rankings of the parameters are more in agreement with the true rankings in the case of a larger sample size, since, as aforementioned, the preference of the sample more clearly reflects the underlying distribution of the objects. Figure 13 shows the MSE's of the parameter rankings for both sample sizes.

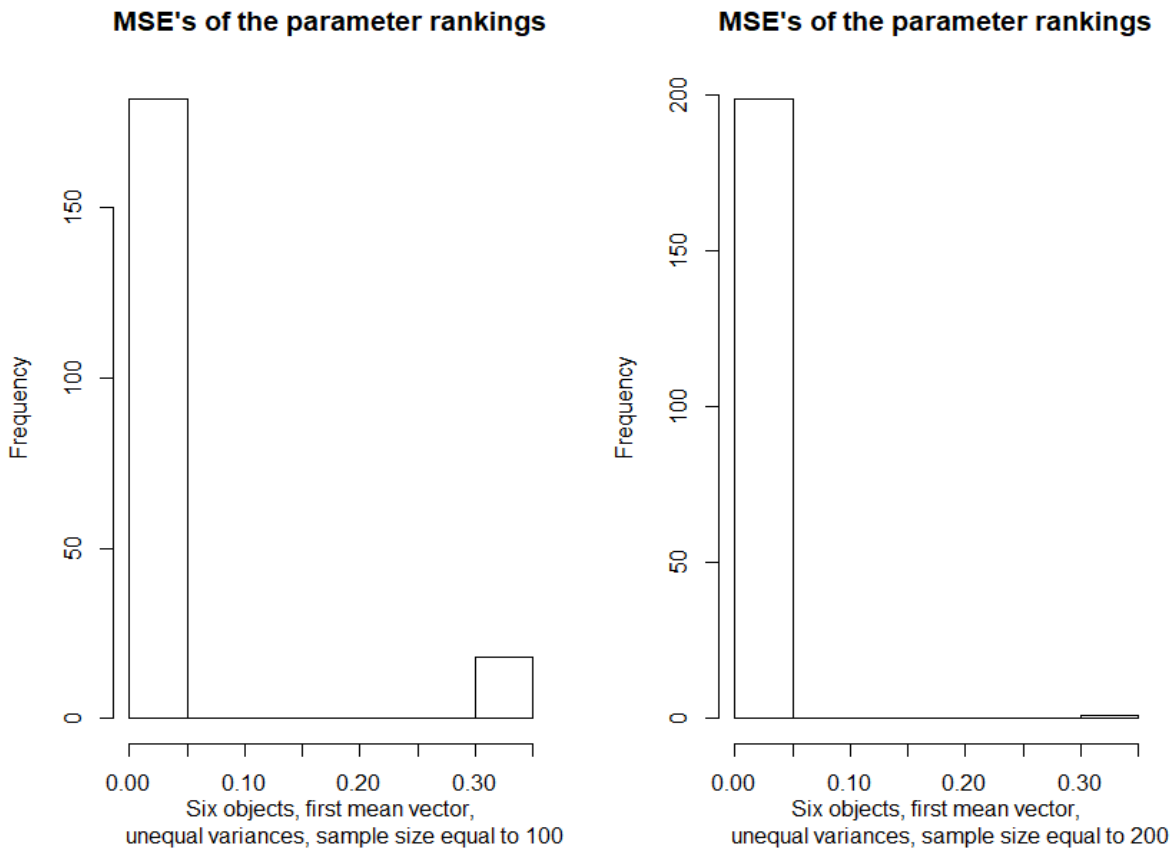



Figure 13: MSE's of the parameter rankings of the best model in the case of unequal variances and sample sizes equal to 100 and 200.

Just as in the case of the sample size being equal to 50, the MSE's of the rankings of the parameters of the model that is not recommended by the AIC were of the same size or larger as the MSE's of the best model. This again confirms that the AIC is capable of determining the best model for a given data set.

Looking at the histograms in figure 11, it can be seen that the distributions of the MSE's of the matrix and parameter rankings are very much alike. Indeed, applying a Kolmogorov-Smirnov test yielded a $p$-value of 0.9972 which implies that the MSE's are coming from the same distribution. Moreover, the $p$-value of the Kolmogorov-Smirnov test became closer and closer to 1 when the sample size increased. The explanation of this phenomenon can be seen in figure 14.

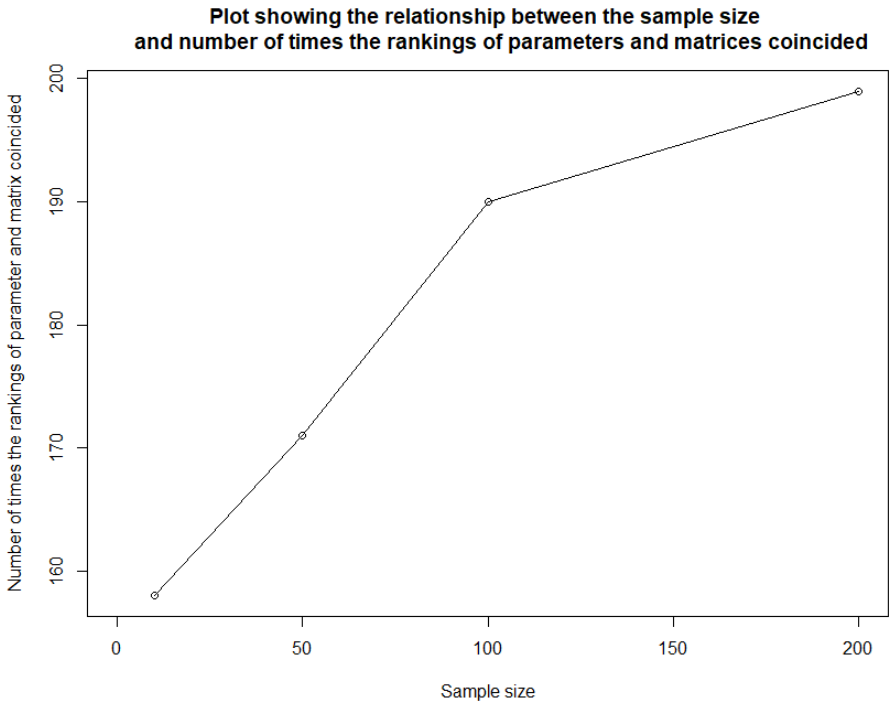



Figure 14: Plot showing the number of times the parameter and matrix rankings coincided dependent on the sample size.

Figure 14 shows that the number of times the rankings of the parameters of the best model and matrices coincide increases when the sample size increases, therefore resulting in a distribution that is more similar which causes the $p$-values to increase.

The outcome of the AIC in the two-step optimization process for the first mean vector and equal variances are tabulated in table 9.

| Sample size | 10 | 50 | 100 | 200 |
|---|---|---|---|---|
| V | 200 | 152 | 3 | 0 |
| III | 0 | 48 | 197 | 200 |

Table 9: The outcome of the AIC for the first mean vector, equal variances and different sample sizes

When the sample size is equal to 10, the MSE's of the parameter rankings are, just as in the case of unequal variances, quite big because of the same reason of the sample size being too small for the models to adequately capture the rankings. However, if the sample size becomes larger than 10, the AIC begins to behave peculiarly. As the sample size increases, Case III is recommended more and more often as the best model, even though the objects have equal variances. The cause of this may lie in the fact that the AIC inherently has a preference for choosing a more complicated model compared to other information criteria such as the Bayesian Information Criterion (BIC) [1]. The BIC not only penalizes for the number of parameters of the model but also takes into account the sample size that was used in the experiment and may therefore, unlike the AIC, recommend Case V as the best model in the case of equal variances. However, even though Case III is recommended by the AIC, the rankings of the parameters are still quite accurate, as can be seen in figure 15.

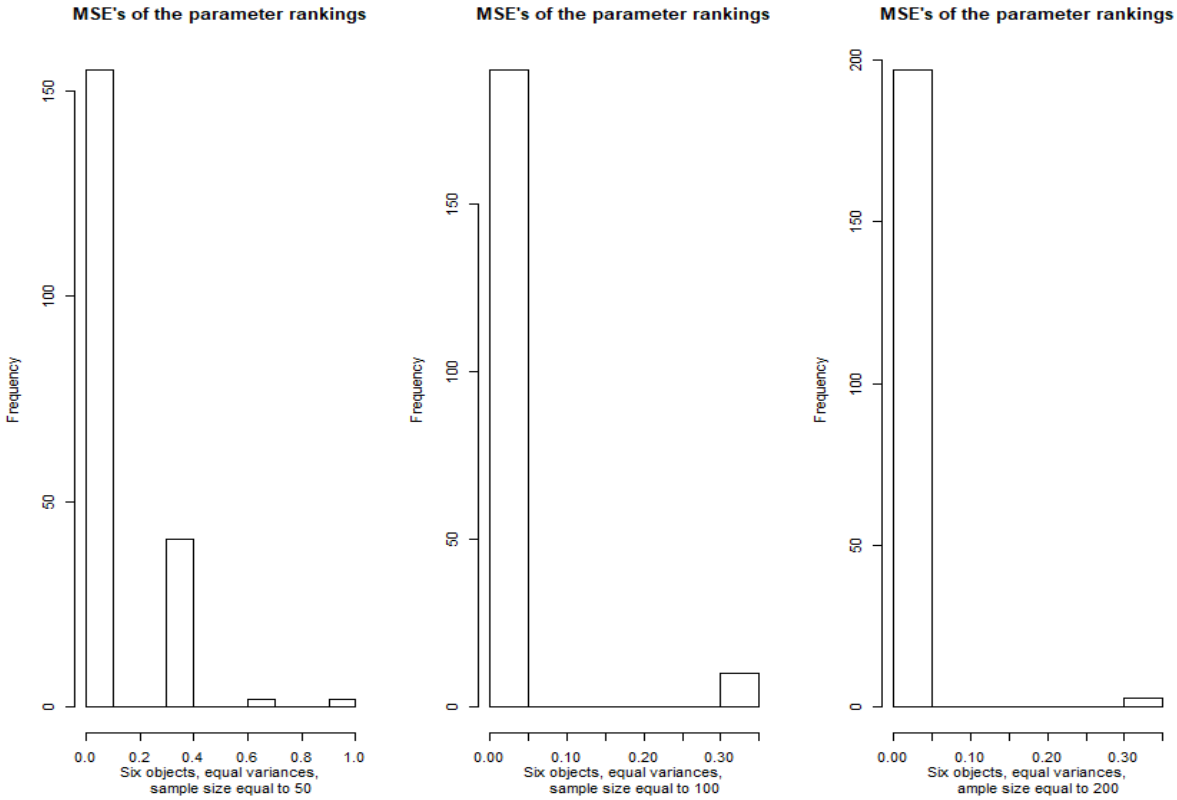



Figure 15: MSE's of the parameter rankings of the best model according to the AIC in the case of equal variances.

For all three sample sizes, the MSE's of the rankings of the parameters of the model not recommended by the AIC were the same or larger implying that the AIC also in this case recommends the best model based on the accuracy of the ranking. Examining the estimations of the variances for Case III, it was found that the estimations were quite close to 1 and thus reasonably accurate. From this it can be concluded that the fact that the AIC chooses a more complicated model is not a big problem. The same phenomenon occurred in the simulation study where the true parameters were used as the initial condition for the optimization, yielding Case III as the best model when the variances were equal. However, the AIC did not have problems distinguishing Case III and Case II as the AIC never recommended Case II when no correlation was present.

In conclusion, it can be stated that in the case of six objects the AIC is a suitable criterion to use for prescribing the best model for a given data set when the sample size is greater than or equal to 50. The latter holds for both the two-step optimization process as the situation where the parameters of the distribution are known. Furthermore, it was discovered that the MSE's of the parameter rankings of the best model are more likely to coincide with both the true rankings and matrix rankings when the sample size increases. Hence, the larger the sample size, the better the performance of the AIC.

Next to six objects, the performance of the AIC has been investigated for seven, ten and fifteen objects as well for both optimization routines. The results of the simulation study of the AIC involving seven objects yielded the same results as the simulation regarding six objects. Therefore, the results for seven objects will not be discussed here.

In the simulation study involving ten objects, one mean vector and two different variance vectors, one having equal and the other having unequal variances, were considered. The mean vector considered was

equal to:

$$\mathbf{S_{10}} = (-1.0, -0.9, -0.75, -0.45, -0.1, 0.3, 0.45, 0.6, 0.85, 1.0)$$

and the vector of unequal variances:

$$\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{10uneq}} = (0.4, 1.6, 0.35, 0.55, 1.0, 1.65, 1.45, 1.0, 0.1, 1.9)$$

Note that the mean vector was again chosen such that the parameters could be estimated accurately for all the sample sizes and all the models.

Table 10 displays the outcome of the AIC for ten objects having unequal variances for different sample sizes for the two-step optimization process and figure 16 the MSE's of the corresponding parameter rankings for the different sample sizes.

| Sample size | 10 | 50 | 100 | 200 |
|---|---|---|---|---|
| V | 192 | 12 | 3 | 0 |
| III | 8 | 188 | 197 | 200 |

Table 10: The outcome of the AIC for ten objects, unequal variances and different sample sizes

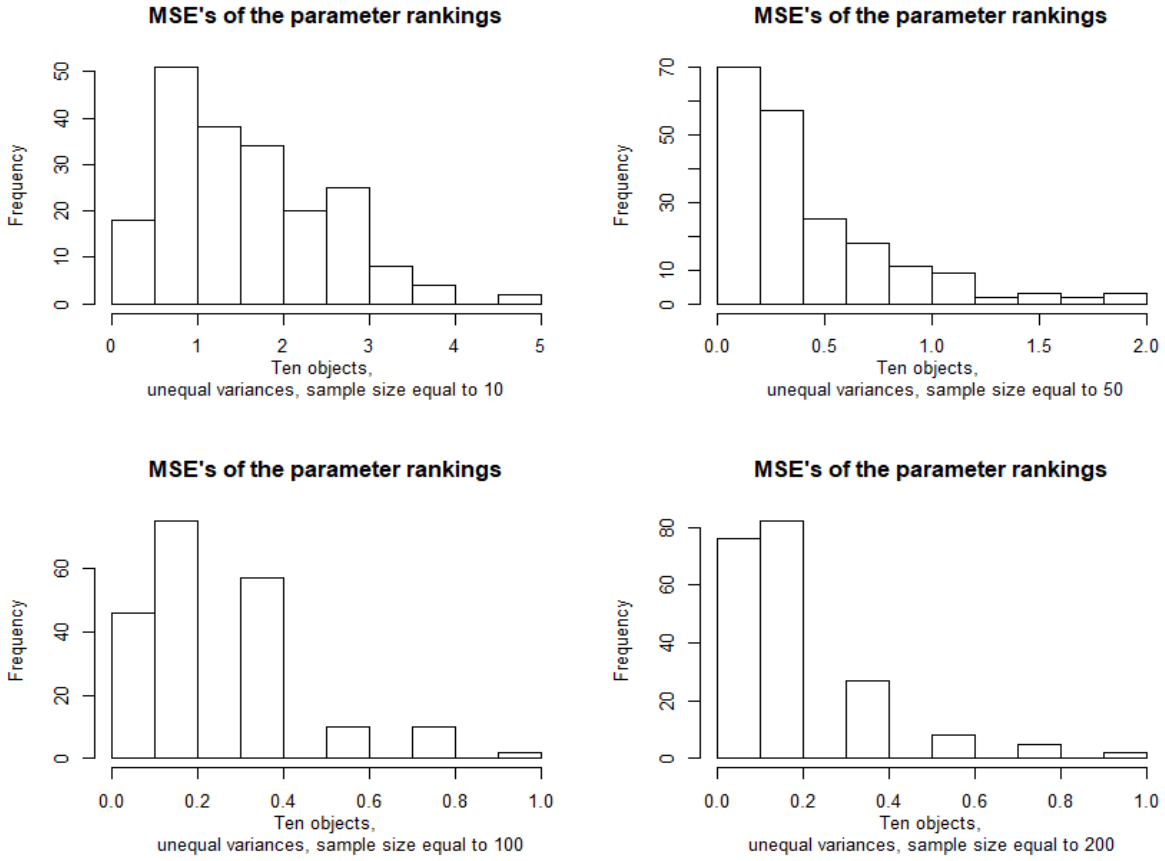



Figure 16: MSE's of the parameter rankings of the best model according to the AIC in the case of ten objects and unequal variances.

From figure 16, it can be seen that the MSE's of the rankings of the parameters are quite large. Even in the case of the sample size being equal to 200, only 38 percent of the parameter rankings of the best model are identical to the true rankings. This, however, does not imply that the AIC is incapable of choosing the best model, since the MSE's of the parameter rankings of the model not recommended by

the AIC are large as well. The reason why the parameter rankings of the best model are poor has to do with the sample size being too small to accurately estimate the parameters. Note that the distance between the means is smaller when compared to, for instance, six objects. This gives that people are more likely to make an inconsistent choice, which results in a less clear representation of the true rankings in the proportion matrix and parameters. In order to solve this, one could for example employ a different mean vector. However this is not realistic as the means of the objects can not be chosen beforehand. The AIC must perform accurately for every possible mean distance. A solution for the problem would be to use a larger sample size. Figure 17 shows that a larger sample size indeed improves the performance of the AIC, as the MSE's of the parameter rankings of the best model in the case of ten objects and a sample size equal to 500 and 1000 are significantly smaller.

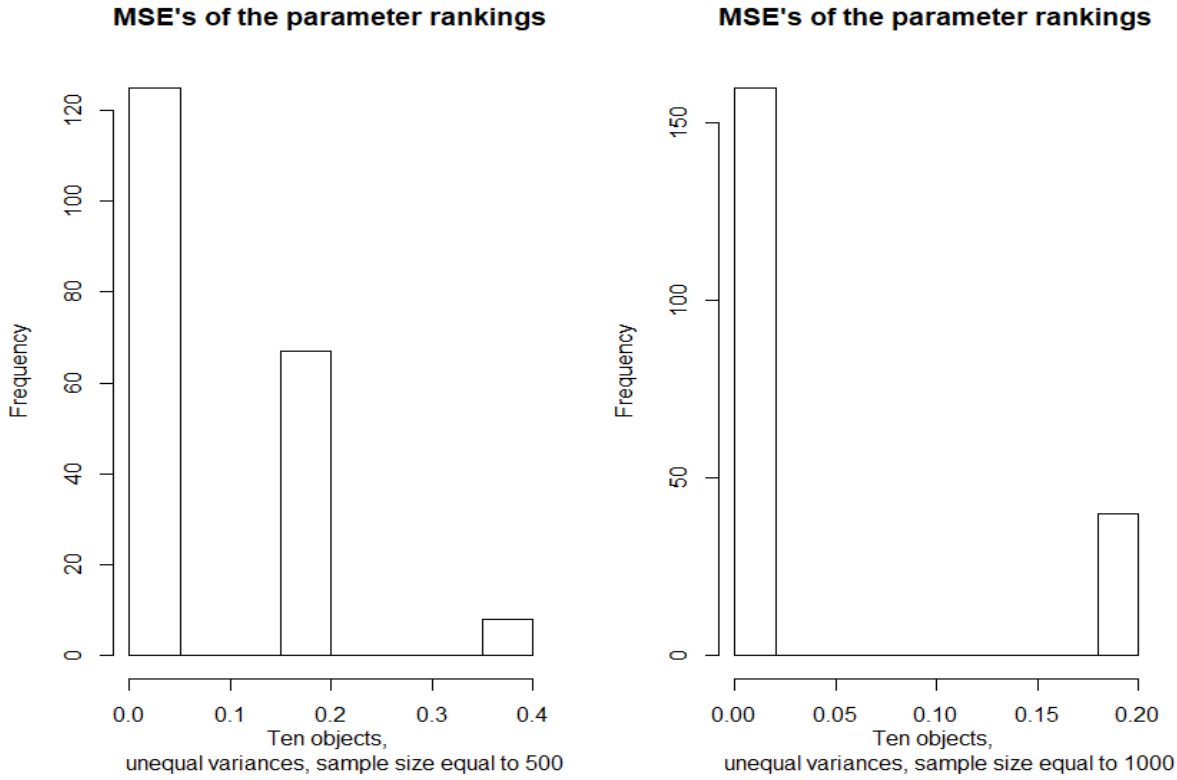



Figure 17: MSE's of the parameter rankings of the best model according to the AIC in the case of ten objects and unequal variances for larger sample sizes.

Hence, the presence of inconsistent choices seems to weigh less when the sample size is large, which results in better estimates for the parameters and as a consequence more accurate rankings of the objects. Therefore, it can be concluded that the AIC performs well for ten objects only if the sample size is sufficiently large. The same results were obtained for ten objects having equal variances and will thus not be discussed here. In the case of fifteen objects, the MSE's of the parameter rankings were even larger compared to ten objects and an even greater sample size was needed for the parameter rankings of the best model to coincide with the true rankings for the same reason. Also, it should be mentioned that for a larger number of objects having equal variances, the AIC recommended Case III as the best model as well. Nevertheless, the estimations of the variances were still quite accurate provided that the sample size was large.

In the estimation part of Case III of the simulation study it was discovered that the estimations of the variances were not accurately captured if the mean distance is large, therefore implying that Case V might be more appropriate in this situation. Whether this is really the case has been investigated by

using the AIC. For this, 200 proportion matrices were simulated from a distribution with mean vector $\mathbf{S_3}$ from table 3 and unequal variance vector $\boldsymbol{\sigma^2_{6uneq}}$ from table 7. Surprisingly, the AIC did not always yield Case V as the best model, tending to recommend Case III more often as the sample size increased. This implies that the unequal variances, even though not perfectly estimated, are still of importance and thus need to be incorporated in the estimation. The reason for this might be that the distance between the means, although spaced quite far apart, are still close enough to one another such that the variances are of significance. In order to examine this, the simulation has been repeated using a mean vector with a larger mean distance, $\mathbf{S_5}$ from table 3, together with the same variance vector. In that case, the AIC yielded Case V as the best model for all instances, thus implying that the variances are not of influence only when the distance between the means is really large.

Next to the performance of the AIC, the coefficient of agreement and the relationship between the coefficient of agreement and the outcome of the AIC have been examined as well. From this, it was discovered that the average coefficient of agreement was significantly lower when the sample size was equal to 10 compared to the other sample sizes. Figure 18 displays the boxplots of the coefficients of agreement obtained from the simulation of the two-step optimization process involving six objects, the first mean vector from table 7 and unequal variances.

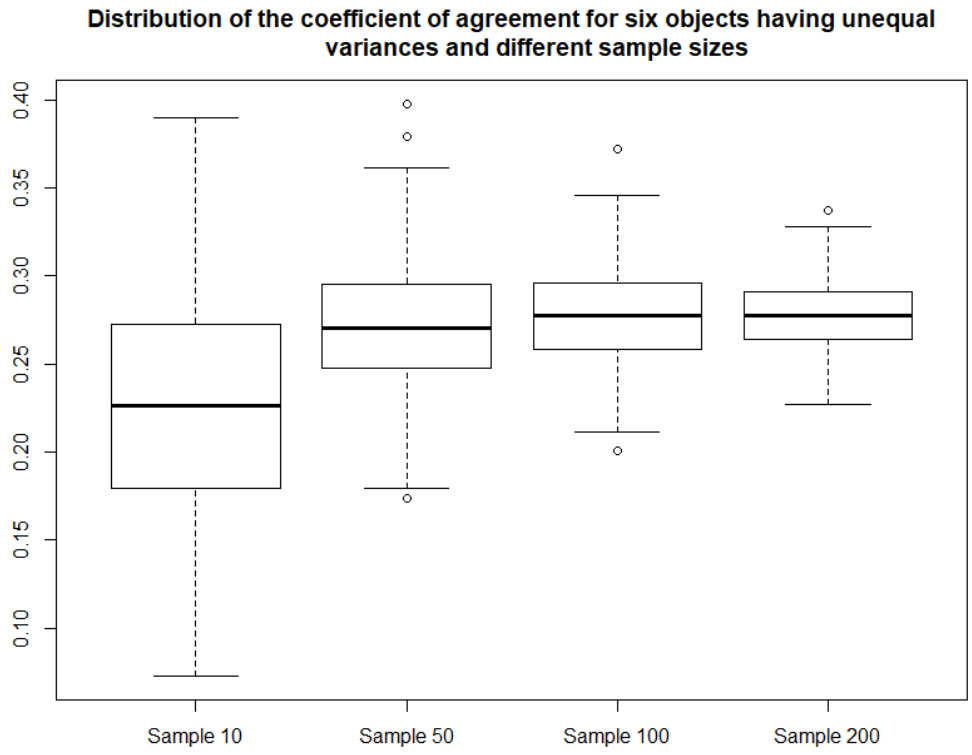



Figure 18: The distribution of the coefficient of agreement for six objects having unequal variances and different sample sizes.

The fact that the coefficient of agreement is significantly lower in the case of a smaller sample size has to do with how the coefficient of agreement is defined. When the sample size is small, the total number of agreements between pairs of individuals is generally lower as well, resulting in a smaller value of the coefficient of agreement. However, even though the median and mean of the coefficient of agreement for a sample size equal to 10 are significantly lower, figure 18 also shows that the values of the coefficient of agreement show more variation when the sample size is small. As the sample size increases, the values of the coefficient of agreement become more stable. The reason why this happens can be explained by the

fact that in the case of a smaller sample size the judgments of the individuals vary widely per sample size, which results in more extreme values for the coefficient of agreement. When the sample size increases, the judgments of the individuals are becoming more similar, resulting in less variation in the coefficient of agreement.

The number of objects was also found to be of influence on the value of the coefficient of agreement. The latter namely decreased when the number of objects increased. This, however, was not only caused by the number of objects itself but was primarily due to the mean distance and the sample size. Note that the means used in the simulation have a small mean distance. As mentioned earlier, if the means are closely spaced, an individual is more likely to make inconsistent choices. This results in less agreement between the individuals and thus a smaller coefficient of agreement. However, the coefficient of agreement again increases for a larger number of objects if the mean distance or sample size is increased. For example, the coefficient of agreement was equal to 0.190 for the case of a close mean distance and sample size equal to 50, but increased to 0.256 when the sample size was equal to 500. This can be explained by the results of the performance of the AIC, where it was discovered that inconsistent choices become of less influence when the sample size increases. The latter might have an impact on the coefficient of agreement as well: since the inconsistencies weigh less heavily in the case of a larger sample size the individuals are, on average, more in agreement with one another resulting in a larger coefficient of agreement.

Besides the sample size, the number of objects and the mean distance, the correlation was also of influence on the coefficient of agreement. In the simulation study of the AIC where the true distribution of the objects is known, it was discovered that the coefficient of agreement increases when the correlation increases. This makes sense, since, as mentioned in Section 2.1, a larger correlation entails that an individual is less likely to prefer an object with a smaller mean over an object with a larger mean. As this holds for all individuals in the sample, the preferences of the individuals will be more in line with one another resulting in a larger coefficient of agreement.

The relationship between the outcome of the AIC and the coefficient of agreement was examined by comparing the mean values of the coefficient of agreement for the different outcomes of the AIC. In both instances, it was discovered that the larger the coefficient of agreement, the more likely the AIC yields a more complicated model as the best model. For example, in the case of six objects having mean vector $\mathbf{S_{6A}}$, unequal variances and a sample size equal to 50, the average coefficient of agreement was equal to 0.289 for the 58 times Case III was recommended and 0.268 for Case V. This is also displayed by figure 19, showing the boxplots of the coefficients of agreement in both cases.

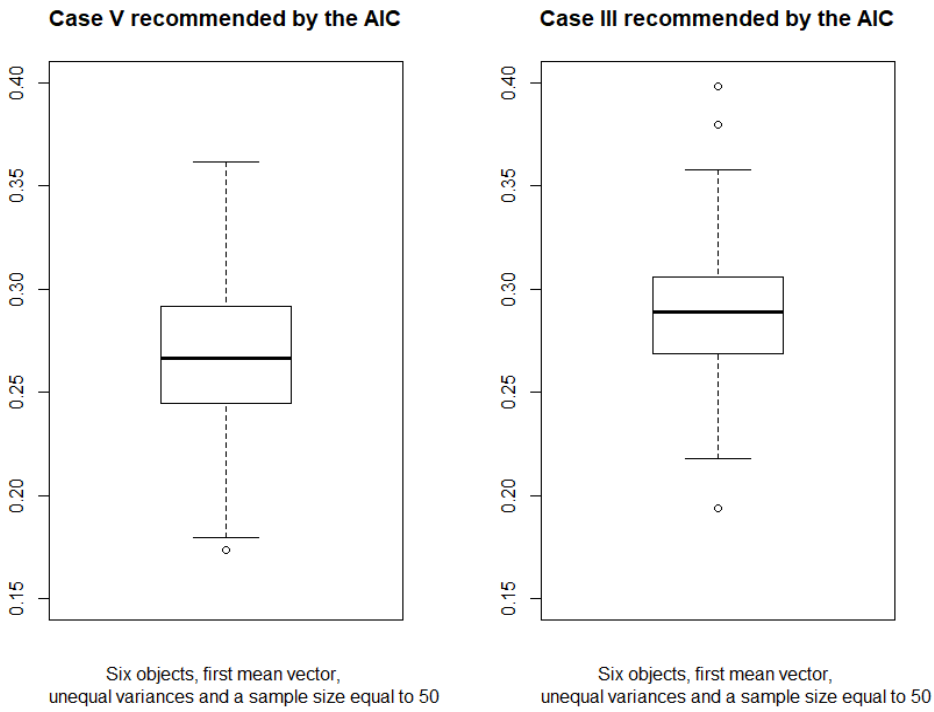



Figure 19: The distribution of the coefficient of agreement for Case V and Case III for six objects and a sample size equal to 50.

Altogether, the AIC seems like a suitable approach that can be used to determine the best case of Thurstone's Pairwise Comparison Model for a given data set. In the case of six and seven objects, the AIC performed rather well for a sample size greater than or equal to 50. When the number of objects increases, however, a larger sample size is needed in order for the AIC to perform properly. The cause of this has not do with the fact that the AIC is incapable of determining the best model, as the parameter rankings of the other cases not recommended by the AIC were just as poor, but with the fact that the sample size needs to be large enough in order for the parameters of the models to be accurately estimated. Furthermore, the value of the coefficient of agreement turned out to depend on the sample size, the number of objects, the mean distance and the eventual correlation included in the distributions. In addition, it was discovered that the AIC is more likely to yield a more complicated model as the best model when the coefficient of agreement is larger.

## 4.4 An alternative method for testing between-group concordance in the case of intransitive preferences

As mentioned in the methodology section, tests for measuring the between-group concordance have been defined only for ranking experiments and can therefore be solely applied when no circular triads are present in the preferences of the sample. An alternative method for measuring between-group concordance when individuals have transitivity problems, based on the MSE's of the proportion matrices, has been proposed in the same section. The performance of this method has been tested with a simulation study, where several distributions and sample sizes were investigated. The sample sizes were chosen equal to 50, 100 and 250. Furthermore, the number of objects was chosen equal to seven. This decision has been made based on both the results of the simulation study of the AIC, yielding adequate results for the rankings in the case of seven objects and the fact that the data study will consider seven objects as well. Four different distributions were examined in the simulation study consisting of two different mean vectors and variance vectors. One of the mean vectors had a small mean distance, while the means of the other mean vector were spaced further apart. The two different mean and variance vectors that were employed in the simulation can be seen in table 11.

| Mean vector | Variance vector |
|---|---|
| $\mathbf{S_{alt1}}$ = (-1.0, -0.8, -0.3, 0.1, 0.3, 0.7, 1.0) | $\boldsymbol{\sigma}^2_{\mathbf{eq\ alt}}$ = (1.0, 1.0, 1.0, 1.0, 1.0, 1.0) |
| $\mathbf{S_{alt2}}$ = (-2.50, -1.75, -0.75, 0.00, 0.75, 1.75, 2.50) | $\boldsymbol{\sigma}^2_{\mathbf{uneq\ alt}}$ = (0.8, 1.3, 1.9, 0.1, 1.2, 0.7, 1.0) |

Table 11: The two mean and variance vectors used in the simulation of the alternative method.

For each distribution and possible sample size, twenty-five pairs of proportion matrices were drawn from the same distribution and sample size and tested for between-group concordance by making use of the alternative method. The decision only to simulate twenty-five pairs was based on computational reasons, as it took a lot of time for one iteration to complete. Since the preferences in the simulation study are always transitive, the twenty-five pairs of matrices were also tested for between-group concordance with Kraemer's test. The outcome of the alternative method was then compared to the outcome of Kraemer's test to examine how many times the results of both methods coincided. By comparing the outcome of both methods, the power of the alternative method can be examined.

Before the results will be presented, it must be noted that the outcome of Kraemer's test was found to be significant in all twenty-five cases for all distributions and sample sizes investigated.

When the distance between the means was small, the same result for the alternative method was obtained for all sample sizes and variance vectors investigated. The alternative method never yielded that the concordance between the groups was significant, as the $p$-value of the Kolmogorov-Smirnov test turned out to be significant in all the twenty-five cases. From this, it can be concluded that in the case of a small mean distance, the performance of the alternative method is extremely poor, as the outcomes of both methods never coincided. Nevertheless, when the sample size was increased, the alternative method began to perform slightly better when the objects had equal variances. For example, in the case of equal variances and a sample size equal to 500, the alternative method gave a significant between-group concordance for one of the twenty-five pairs. When the sample size was even further increased to 1000, the result of Kraemer's test and the alternative method coincided in three of the twenty-five cases. The increase in sample size, however, was not of influence on the performance of the alternative method when the objects had unequal variances.

The reason why the performance of the alternative method was poor has to do with the fact that in the case of a small mean distance the preferences of the individuals differ a lot, which results in two proportion matrices that are quite dissimilar from one another. Especially when the sample size is small, the proportion matrices, although drawn from the same distribution, may differ a lot. Furthermore, in the case of a small sample size, the estimations of the parameters may not be very precise when compared to the true parameters. As a consequence, the simulated matrices from the distribution using the estimated parameters of the first proportion matrix are close to the first proportion matrix, while they deviate a lot more from the second proportion matrix. This results in MSE's that are significantly different from one another, which explains why the Kolmogorov-Smirnov test almost always yielded a significant $p$-value.

If the means of the objects are spaced further apart, the result of the alternative method yields more often that the between-group concordance is significant. For example, when the sample size is equal to 50 and the variances of the objects are unequal to one another, the alternative method returns a significant result in two of the twenty-five cases, while for the case of the objects having equal variances the alternative method yields significant between-group concordance four out of twenty-five times. Similar results were found in the case of the sample size being equal to 100 or 250 and will therefore not be discussed in detail. The number of times, out of the total twenty-five times, that the alternative method yielded a significant result in the case of a larger mean distance varied between one and five for the different sample sizes and was generally higher when the variances were equal to one another. However, unlike the smaller mean distance, a larger sample size did not seem to improve the performance of the alternative method, as the number of times the alternative method and Kraemer's test coincided remained roughly the same. In general, it can be stated that the alternative method performs better when the distance between the means is larger. This can be easily explained by the fact that the preferences of the individuals are more distinct in the case of a larger mean distance, resulting in proportion matrices that are more alike.

However, even though the performance of the alternative method improves when the mean distance increases, the results of the simulation study still show that the alternative method does not work quite optimal in the case of a larger mean distance, since the results were significant in maximum twenty percent of the cases. From this, it can be questioned what the minimum distance between the smallest and largest mean must be in order for the alternative method to return significant between-group concordance for all twenty-five pairs of matrices. This has been investigated as well by looking at the accuracy of the alternative method for different mean distances and sample sizes. In this study it was discovered that in order for the alternative method to yield significant results in all the twenty-five cases the distance between the smallest and largest mean needs to be extremely large. For example, when the sample size equals 50 and the variances are unequal to one another, the distance between the smallest and largest mean must at least be equal to 25. The minimum distance, however, decreased as the sample size increased. Nevertheless, even when the sample size was equal to 1000, the minimum distance needed was still quite large, being equal to 10. The fact that such a large distance is needed in order for the alternative method to be 100 percent accurate, shows that the alternative method is not a suitable method to use for determining between-group concordance since the estimations of the parameters can not be accurately obtained in that case. In Section 4.1 it was discovered that the estimations of the scale values can be accurately estimated if and only if the distance between the largest and smallest mean are spaced not too far away from one another, the maximum distance determined by the number of objects and the sample size. In the case of seven objects and a sample size equal to 50, the maximum distance between the smallest and largest mean such that accurate estimations can be obtained is equal to 4.979, which is much smaller than 25.

In conclusion it can be stated that the alternative method is not an appropriate method to employ when one wants to determine whether the concordance between two groups is significant. The results have shown that in order for the alternative method to be 100 percent accurate, the minimum distance between the smallest and largest mean must be large. However, the minimum distance needed is that large that accurate estimations of the parameters can not be obtained. As the main interest of applying Thurstone's Pairwise Comparison Model is to obtain accurate rankings of the objects, it is better to use Kraemer's test when one wants to determine whether two groups are in concordance with one another.

# 5 Data study

The data study performed serves two objectives. The first of which is to examine how pairwise comparison experiments and Thurstone's model can be employed to model human food preferences. What are the advantages and disadvantages and how does this relate to the similarity in products? Next to that, the influence of an introduction text given at the beginning of the experiment, emphasizing the importance of eating healthy, on the preferences of individuals for certain food products is investigated as well. This section will explain the design of the data study that was employed, followed by a presentation of the results.

## 5.1 The set-up of the data study

The data study performed considered four groups that were independent of one another. Of these four groups, two groups judged seven general food products with a low degree of similarity, while the other two groups judged seven snacks, food products that were very similar to one another. The pairwise comparison experiment was not carried out in person, but digitally by means of a Google form. This meant that the individuals judged the food products by looking at images of them. Concerning the images, it was ensured that the image of every product had a neutral background, so that the individuals only focused on the food products and were not distracted by anything else. Figures 20 and 21 display the products and their corresponding images used in the pairwise comparison experiments for the general and similar products respectively.

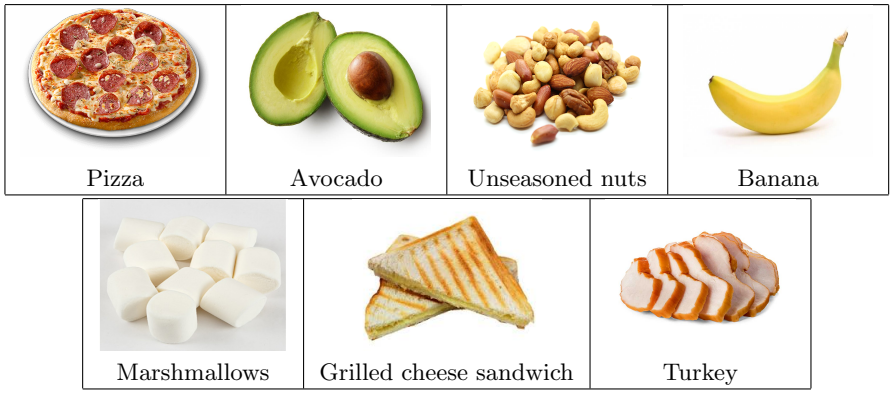



Figure 20: Images of the general food products used in the pairwise comparison experiment.

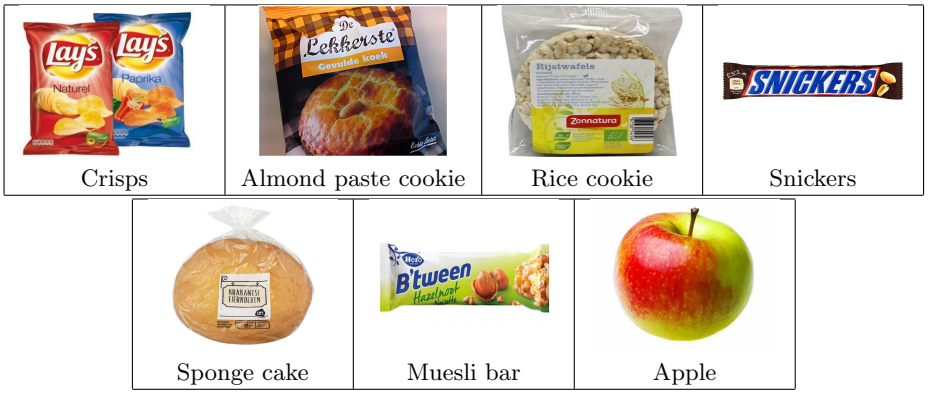



Figure 21: Images of the similar food products (snacks) used in the pairwise comparison experiment.

The language of the forms used in the data study were entirely in Dutch and the minimum threshold for the sample size of the groups was set equal to 50. This decision has been made based on the results

of the simulation study, since the AIC can only be used to determine the best model if the sample size is sufficiently large such that accurate estimations can occur. An attempt was made to compose samples of individuals from different backgrounds by sending the Google forms to relatives, friends and acquaintances with the request to share the form with their own network. The images of the food products were presented side by side. Figure 22 shows an example of how the food products were presented in the form.

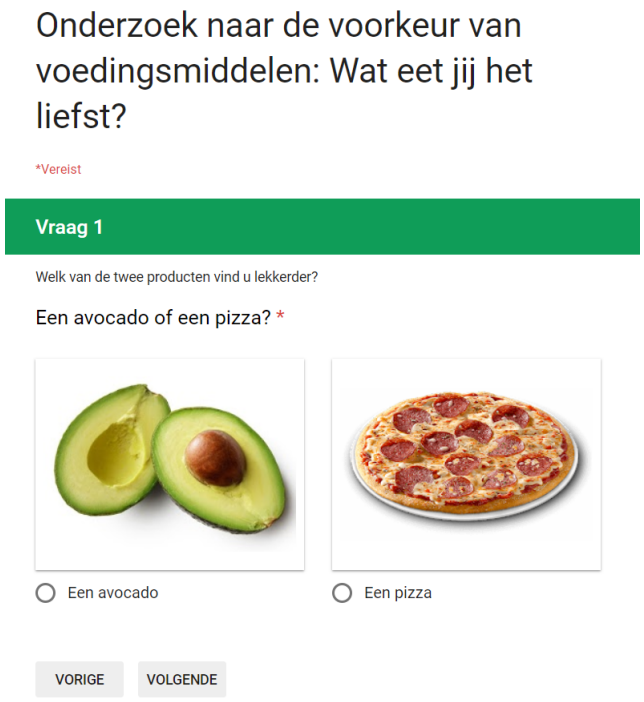



Figure 22: An example of a question used in the form.

Every pair of groups judging the same set of food products were each given a different introduction text at the beginning of the experiment. One of the two groups in the pair received an introduction text that asked the individuals to just give their preferences for the several food products based on what they would like to eat. The other introduction text, however, not only asked for the preference of the individuals for the food products, but also emphasized the importance of eating healthy food. The introduction texts were exactly the same for the general food and snack products. Lastly, after the pairwise comparison experiment was finished, participants were also asked to answer a few questions about their eating habits: which aspects of food products do they find most and least important? The options in the case of general food products were how healthy the food products are, the price of the food products and the ease of preparation. The latter means how easy the food can be eaten, for example a banana is prepared in 2 seconds: you peel the skin off and you can eat it. A pizza, however, take more time to prepare, as it needs to be baked in the oven. For the snack products, the participants could choose between the price of the snacks, how much the snacks are capable of filling the stomach and how healthy the snacks are. In this way, it can also be tested whether the introduction text is of influence on the beliefs of the individuals: do the individuals in the group receiving the introduction text accentuating the importance of healthy food state more often that they find the health aspect of food more important? For reasons of convenience, the group that receives an introduction text at the start of the experiment emphasizing the importance of eating healthy will from now on be referred to as the healthy group. In addition, the other group will be referred to as the tasty group.

By means of setting up the data study as mentioned, the influence of the introduction text on the preferences of individuals can be examined: are the individuals in the healthy group more prone to prefer the healthier food products? At the same time, the relationship of the similarity in products with both

the influence of the introduction text and the (intransitive) preferences of the individuals in general can be investigated as well.

Before going to the results, a few remarks about the process of gathering the data and how the data was analyzed will be made. First of all, the collection of the data went very easily. Within a week, the minimum number of respondents needed for the samples was obtained and even exceeded. The sample sizes for the different groups varied between 50 and 79. Furthermore, some of the respondents even sent a message to me afterwards stating how much they enjoyed filling in the questionnaire.
Next to that, the analysis of the data was performed in a specific order for both general and similar products and will also be presented in this order in the results section. A brief explanation of this process will be given. First, the number of circular triads in the preferences of the individuals were determined for both the tasty and the healthy group and compared to one another in order to see whether there's a difference between the competences of the individuals in the two groups. Subsequently, the product trios that caused the most intransitive preferences in the individuals were determined for the two groups. By doing this, it can be investigated which products are troublesome in a pairwise comparison experiment. After this, the coefficient of agreement was computed for both groups and tested for significance. Then, the best case of Thurstone's model for both data sets was determined by the AIC and the resulting parameters and matrix rankings compared and conclusions about the concordance between the groups were drawn based on the rankings. Furthermore, if the number of individuals being completely transitive was not too small, the decision was made to remove all the intransitive individuals from the sample and test for between group-concordance by making use of Kraemer's test. Before the test was applied, however, the best model for the data sets of the transitive samples was again determined by the AIC. This was done to see how the rankings of the food products are influenced by the intransitive individuals in the sample.

## 5.2 Results of the data study

The results of the data study will be presented in three parts. In the first part of this section, the results for the general food products are discussed. After that, the outcome of the data study for the snack products are elaborated on. In the final part of this section, the results of both experiments are compared and final conclusions will be drawn.

### 5.2.1 Results of the general food products

The number of participants in the pairwise comparison experiment involving general products was equal to 55 for the tasty group and 50 for the healthy group. First, the number of circular triads occurring in the two groups will be discussed. Table 12 displays how many individuals had a certain number of circular triads in their preferences for both groups:

| Number of circular triads | Tasty group | Healthy group |
|---|---|---|
| 0 | 40 | 38 |
| 1 | 11 | 7 |
| 2 | 2 | 3 |
| 3 | 1 | 1 |
| 4 | 0 | 1 |
| 5 | 0 | 0 |
| 6 | 0 | 0 |
| 7 | 1 | 0 |

Table 12: Total number of individuals having a certain number of circular triads in their preferences for the two groups for the pairwise comparison involving general food products.

Looking at table 12 it can be seen that the number of individuals having no circular triads is approximately the same for both groups. However, since the healthy group consisted of a slightly smaller sample size, the total percentage of individuals being transitive was slightly higher in that group: 76 percent against 72,7 percent for the tasty group. Furthermore, the individuals having too many circular triads

in their preferences were determined from Kendall and Babington-Smith's table introduced in Section 3.1. Using a significance level of $\alpha = 0.05$ it can be deduced from figure 4 that the maximum number of circular triads that may occur within an individual is equal to three. This means that in both groups, one individual has too many intransitive preferences. However, the number of circular triads differs a lot for both individuals, the individual in the healthy group having four circular triads, while the preferences of the one in the tasty group consists of a whopping seven circular triads. In general, it can be stated that the healthy group is slightly more consistent in their preferences since the preferences of an individual contain on average 0.4 circular triads in the healthy group, while the average number of circular triads per individual is equal to 0.455 in the tasty group.

The product triads that frequently occurred as circular in the preferences of the individuals in both groups were the following triads: Pizza-Avocado-Grilled cheese sandwich and Pizza-Banana-Grilled cheese sandwich. However, the number of times that these triads occurred as circular was not very high. The first circular triad occurred three times in the preferences of the tasty group and two times in the preferences of the healthy group. The second circular triad occurred an equally number of times in the preferences of the two groups, but the other way around. For both groups, the maximum number of times that a triad occurred as circular in the preferences was equal to three and thus not very high. Generally, if a triad was circular in the preferences of the group, it mostly occurred only one time. Furthermore, of the $\binom{7}{3} = 35$ possible triads, 17 never occurred as circular in the tasty group, while in the healthy group 18 triads were never judged intransitively. Even though the number of times the triads occurred as circular were not extremely high, some products did occur more often as part of a circular triad than the others. Primarily the pizza and grilled cheese sandwich were frequently judged intransitively in both groups. Of the other products, the marshmallows occurred the least number of times in a circular triad. The fact that the pizza and grilled cheese sandwich often occurred as part of a circular triad in the preferences of the individuals might be because they are both well-liked by people but in some instances, the latter prefer other objects as well. One individual of the tasty group, for example, stated after the pairwise comparison experiment: "Normally I would always prefer a pizza over anything, because who doesn't like pizza. But when I'm just done with my exercise and all sweaty, I need to reload my carbohydrates and would therefore prefer a banana.". The reason why the marshmallow occurred the least number of times in a circular triad in both groups can be explained by the fact that a lot of individuals dislike marshmallows. One individual of the healthy group even said: "Marshmallows are disgusting, they are just too sweet.". The unpopularity of the marshmallows is also reflected in the frequency matrices of the preferences of both groups, which can be found in appendix B.1. As marshmallows are not very popular, the other two objects in the triad are more likely to be preferred over the marshmallows. This automatically makes the triad non-circular, since the three objects, with the marshmallows in last place, are ranked.

The coefficient of agreement for both groups was found to be significant using a significance level of $\alpha = 0.05$, therefore indicating that the individuals within both groups had a common preference. The coefficient of agreement was slightly higher in the healthy group, being equal to 0.232 while the coefficient of agreement in the tasty group equaled 0.187. At first sight one would think that this makes sense, since the healthy group made less intransitive judgments and thus have preference that is more distinct. However, this statement is generally not true, since personal agreement in the form of transitivity and agreement within groups are independent of one another. For example, every individual in the sample may be completely transitive in his judgments, but the preferences of the individuals in the group may still differ. Therefore, the cause of the higher coefficient of agreement of the healthy group is simply that the preferences of the individuals in that group are more in agreement with one another.

After the analysis of the coefficient of agreement, the AIC was used to determine the best case of Thurstone's model for the data sets. For both groups, the best model turned out to be Case III. Table 13 displays the parameters of Thurstone's model for both groups and the rankings of the products based on the scale values. Moreover, the rankings of the matrix, obtained by summing the rows of the frequency matrix, coincided exactly with the rankings of the scale values. The sums of the rows of the frequency matrix, including the matrix rankings and the plots of the distributions of the food products for both groups can be found in the appendix.

| | Means tasty | Means healthy | Var. tasty | Var. healthy | Rank. tasty | Rank. healthy |
|---|---|---|---|---|---|---|
| Pizza | 0.239 | 0.762 | 0.750 | 1.344 | 3 | 2 |
| Avocado | -0.342 | -0.344 | 2.373 | 3.560 | 6 | 5 |
| Unseasoned nuts | 0.126 | -0.131 | 1.085 | 0.422 | 4 | 4 |
| Banana | 0.712 | 0.141 | 1.176 | 0.540 | 1 | 3 |
| Marshmallows | -1.000 | -0.685 | 0.300 | 0.034 | 7 | 7 |
| Grilled sand. | 0.457 | 0.767 | 0.771 | 0.990 | 2 | 1 |
| Turkey | -0.192 | -0.510 | 0.549 | 0.110 | 5 | 6 |

Table 13: Means, variances and rankings of the general food products for both groups.

From table 13 it can be seen that the means of the tasty group are spaced further apart from one another than the means of the healthy group. For instance, the means of the top two of the healthy group differ only 0.005 from one another, which is not much. Next to that, it can be deduced that the healthy introduction text, given at the beginning of the pairwise comparison experiment is not of influence on the preferences of the individuals. It is not the case that the more healthy food products are preferred in the healthy group, since the pizza and grilled cheese sandwich, which are both unhealthy, are the most preferred food products. Ironically, the banana ends up in first place in the rankings of the tasty group, well ahead of the (unhealthy) number 2 and 3: the pizza and the grilled cheese sandwich. The introduction text is thus not really of influence on the preferences itself. The same was discovered for the eating habits of the individuals. In the tasty group, a total of 63,6 percent stated that they found the health aspect of food the most important one, while only 28 percent of the individuals in the healthy group paid most attention to how healthy the food products are. The majority in the healthy group stated that the price of the food products was most important to them. How easy the food products are prepared was found to be the least important factor for both groups.

Since the coefficient of agreement is significant for both groups and the rankings of both groups are quite close to one another, it can also be determined whether the groups are in concordance. However, as circular triads are present in the preferences of both groups and the alternative method was found to be unsuccessful, between-group concordance is hard to determine. One could attempt to investigate between-group concordance by comparing the rankings of both groups. Examining the latter, it is apparent that the rankings of the food products do not coincide completely, but are rather in concordance. The marshmallows end up in last place in both rankings and additionally also have the smallest variance in both cases. The top three of both groups consist of the same three products: the pizza, banana and grilled cheese sandwich but the three are not ranked in the same order in the two rankings. The same holds for the avocado and turkey, that end up in fifth and sixth place. Furthermore, the avocado has the largest variance in both instances. This can be explained by the fact that the avocado is either fancied or disliked. For example, one individual in the healthy group stated: "I would prefer an avocado over anything, they are just so delicious!", while another individual from the same group said: "I will never eat an avocado, yuck.". Altogether, the preferences of the groups are suspected to be quite in concordance with one another.

Since the number of individuals having transitive preferences are still quite high in both groups, the decision was made to remove all the intransitive individuals from both groups and use Kraemer's test for between-group concordance. Removing the intransitive individuals from the groups did not seem to affect the rankings of the food products, since the same rankings were obtained as in the case of intransitivities being present. This implies that the preferences of the intransitive individuals did not deviate a lot from the preferences of the transitive sample. The frequency matrices of the transitive groups, the parameter and matrix rankings and the plots of the distributions can be found in appendix B.1.

The result of Kraemer's test turned out to be very significant. The value for the between-group concordance in Kraemer's test, $W_B$, was equal to 0.900, which resulted in $\eta$ being equal to 9.03445. The $p$-value corresponding to this statistic was 0.0085, which is significant using significance level $\alpha = 0.05$. Therefore, the null hypothesis of the groups' preferences being totally random is rejected. This confirms the suspicion that the preferences of the groups might be in concordance with one another. It can thus be stated that the groups indeed have the same preferences and furthermore confirms that the introduc-

tion text given at the beginning of the experiment was indeed not of influence on the preferences of the individuals.

### 5.2.2 Results of the snack products

The sizes of the groups in the pairwise comparison experiment for the snack products were equal to 58 for the tasty group and 79 for the healthy group. The number of times that a certain number of circular triads occurred in the preferences of the individuals in the two groups are tabulated in table 14

| Number of circular triads | Tasty group | Healthy group |
|---|---|---|
| 0 | 35 | 56 |
| 1 | 17 | 12 |
| 2 | 3 | 7 |
| 3 | 2 | 1 |
| 4 | 1 | 1 |
| 5 | 0 | 2 |

Table 14: Total number of individuals having a certain number of circular triads in their preferences for the two groups for the pairwise comparison involving snacks.

Table 14 shows that the individuals in the healthy group were slightly more consistent in their judgments. In the healthy group, 70,9 percent of the individuals were completely transitive, while only 60,3 percent of the individuals in the tasty group had no circular triads in their preferences. Furthermore, the average number of circular triads in the preferences of an individual equaled 0.569 for the tasty group and 0.544 for the healthy group. However, even though the healthy group had a lower average number of circular triads, there were more individuals in the healthy group having too many circular triads in their preferences: three individuals in the healthy group have preferences that contain too many intransitive judgments, whereas just one individual in the tasty group has preferences that contain too many circular triads.

There was one product triad that appeared regularly as circular in both groups. This triad consisted of the crisps, the almond paste cookie and the muesli bar and occurred three times in the tasty group and four times in the healthy group. There were no triads that occurred significantly more often as circular than other triads: the highest number of times a triad appeared as circular was equal to four for both groups. Moreover, the number of times a product was part of a circular triad was approximately the same for all products. This implies that the snacks are equally judged inconsistently by the individuals, regardless whether the snacks were healthy or not.
In the tasty group, 14 triads, out of the total 35, never occurred as circular in the preferences of the individuals. This number was much lower for the healthy group, where only 8 triads were never circular. The reason why this number is lower in the healthy group probably has to do with the size of the healthy group being larger and the fact that there were no triads that occurred more often as circular, therefore increasing the probability of more and diverse circular triads.

The coefficient of agreement turned out to be significant for both groups. The value of the coefficient of agreement was equal to 0.091 for the tasty group and 0.114 for the healthy group. Therefore, the individuals in the healthy group are slightly more in agreement with one another than the individuals in the tasty group. The best case of Thurstone's model according to the AIC turned out to be Case V for both data sets. The scale values for the snacks for both groups, together with the ranks are displayed in table 15.

| | Means tasty | Means healthy | Rank. tasty | Rank. healthy |
|---|---|---|---|---|
| Crisps | 0.150 | 0.289 | 3 | 2 |
| Almond paste cookie | 0.046 | -0.374 | 4 | 7 |
| Rice cookie | -0.574 | -0.287 | 7 | 6 |
| Snickers | 0.015 | -0.142 | 5 | 4 |
| Sponge cake | 0.298 | 0.172 | 1 | 3 |
| Muesli bar | -0.230 | -0.182 | 6 | 5 |
| Apple | 0.293 | 0.524 | 2 | 1 |

Table 15: Means and rankings of the snacks for both groups.

The ranks of the matrix were found to be identical to the ranks of the scale values. The latter, together with the sums of the rows of the frequency matrices of both groups and the plot of the distributions of the snacks, can be found in appendix B.2. Table 15 shows that the preferences of the two groups deviate a lot from one another. There are no snacks that have the same position in both of the rankings. The greatest difference in the rankings being the almond paste cookie, which is ranked in last place in ranking of the healthy group and in fourth place in the ranking of the tasty group. Furthermore, the introduction text at the beginning of the experiment seems not to be of influence on the preferences of the individuals. The sponge cake, which is quite healthy, is ranked in first place by the tasty group, while the healthy group prefers an apple. Note, however, that the scale values of the most and second-most preferred snacks by the tasty group are quite close to one another. The scale value of the sponge cake only differs a mere 0.005 from the scale value of the apple. The reason why the preferences of the groups seem to differ a lot from one another, may be explained by the fact that every individual prefers different kinds of snacks. For instance, one individual said: "I'm a sugar lover and therefore not really into crisps or rice cookies.", while another participant said: "I'm trying to snack as healthy as possible and would never eat something like a snickers or crisps.". This entails that the rankings of the snacks for the two groups may differ as well. Still, it should be noted that the rankings of the two groups are not entirely different from one another. For example, the top three of both groups consisted of the same products, but were ranked in a different order. The reason why the almond paste cookie is ranked differently in both groups is probably just coincidence: more individuals in the tasty group prefer almond paste cookies over the other snacks. Even though the introduction text was not of influence on the preferences of the individuals, 43 percent in the healthy group stated that they found the health aspect of snacks the most important, whereas the tasty group declared that they mainly focus on the stomach filling property of the snacks. Only 31 percent of the individuals in the tasty group valued the health aspects of the snacks the most. The price of the snacks was the least important factor for both groups.

As the coefficient of agreement was found to be significant for both groups, the groups were also tested for between-group concordance. This was done by looking at the rankings. As mentioned earlier, the rankings of the two groups were not really in line with one another, therefore the expectation is that the groups are probably not in agreement with one another. Whether the latter was really the case was tested by Kraemer's test for between-group concordance.

Since the number of transitive individuals was sufficiently large for both groups, the intransitive individuals were excluded from the sample. First, the influence of the preferences of the intransitive individuals on the rankings were examined. The best case of Thurstone's model according to the AIC for the transitive samples was again Case V. Furthermore, the matrix rankings of both groups contained ties; the apple and sponge cake were tied in first place in the ranking of the tasty group and the muesli bar and rice cookie were tied in fifth place in the ranking of the healthy group. The matrix rankings, scale values, resulting parameter rankings and plots of the distributions for both groups are displayed in appendix B.2.
Moreover, it turned out that the intransitive individuals in the healthy group did not have any influence on the rankings of the snack products: the rankings of the snacks remained the same. However, the rankings of the snacks did change for the tasty group when the intransitive individuals were excluded. The apple and sponge cake switched positions in the ranking and the spacing between the scale values increased as well from 0.005 tot 0.023. In addition, the almond paste cookie and snickers switched positions as well. Investigating the frequency matrix of the intransitive individuals, it was discovered that the intransitive individuals primarily preferred almond paste cookies and were not very fond of apples. This

explains why the snacks switched positions in the ranking. Furthermore, the rankings of both groups still differed quite a lot from one another, only the snickers were assigned the same ranking. Therefore, looking at the rankings of the products, it is likely that the groups are not in concordance with one another.

Kraemer's test, however, contradicted this, as it yielded a significant $p$-value. The coefficient of between-group concordance $W_B$ was equal to 0.889, which resulted in $\eta$ being equal to 8.045. The $p$-value corresponding to this value was equal to 0.011, which is significant at significance level $\alpha = 0.05$. This result shows that between-group concordance is hard to determine just by comparing the rankings of the two groups. Hence, a hypothesis test is needed in order to determine whether the two groups are in concordance with one another. Furthermore, the outcome of Kraemer's test affirms that the introduction text was not of influence on the preferences of the individuals.

### 5.2.3 Conclusions

Comparing the results of the general food products and snack products it was discovered that more circular triads occurred in the case of snacks. This can be explained by the fact that the preferences of the individuals for the general food products are more distinct. They know exactly which products they like and which ones they dislike, while in the case of snacks the products are in general equally preferred. One individual, which participated in both experiments, stated: "In the case of general food products, I know exactly what I would eat and what not. However, choosing between the snacks was really hard for me as I actually liked all of them.". The fact that the preferences of the individuals are more distinct in the case of the general food products entails that the number of circular triads is significantly lower for the general food products.

The fact that most of the individuals like most of the snacks, in contrast to the general food products, also causes a lower coefficient of agreement for the two groups judging the snacks. The individuals find it harder to make a decision between the snack products, which results in a greater variety of preferences for the snacks and thus a lower coefficient of agreement. Furthermore, the relationship between the coefficient of agreement and the outcome of the AIC was noticeable as well. In the case of a higher coefficient of agreement, it was discovered that the AIC is more likely to yield a more complicated model as the best model. This was also the case for the food products: Case III was recommended for the general food products, while Case V turned out to be the best model for the snack products.

Furthermore, the introduction text given at the beginning of the pairwise comparison experiment did not have any influence on the preferences of the individuals for both the general food products and snacks: the individuals in the healthy groups did not have a more distinct preference for the healthy products. After the experiment was completed, some participants of the healthy group were asked for their opinions on the introduction text. A few individuals stated that they didn't even read the introduction text, starting the experiment right away. Others said that they read the introduction text, but didn't really focus on the health aspect. One individual explained: "I just pick what I prefer to eat". Nevertheless, the individuals in the healthy group of the snacks did state that they found the health aspect of the snacks the most important factor, in contrast to the tasty group. This was also visible in the ranking for the snack products, since the top two of the healthy group consists of the sponge cake and the apple, which are both quite healthy.

The preferences of the healthy group and tasty group were significantly in concordance with one another for both sets of food products. However, the between-group concordance was slightly greater in the case of the general food products. This can again be explained by the fact that the individuals experienced more difficulty in deciding which snacks they preferred. The latter namely caused that the preferences of the individuals differed a lot from one another, which not only resulted in a lower coefficient of agreement but also in less concordance between the groups. Next to that, the results of the data study for the snack products showed that determining the between-group concordance only based on the rankings of the objects is quite difficult, since the rankings may deviate a lot from one another, but the groups might still be in concordance. Therefore, a hypothesis test, such as Kraemer's, is needed so that the significance of the between-group concordance can be confirmed.

In conclusion it can be stated that a pairwise comparison experiment is quite a reasonable method to use for modelling human food preferences. However, it does seem to be more appropriate to use for general food products, since the individuals made less circular triads in their preferences in that case. This means that a more definite ranking of the food products can be obtained, which is the primary objective. Furthermore, the individuals stated that it was really hard for them to choose between the snack products. Next to that, one individual stated they she found it peculiar that some products were presented on their own and not as part of a dish. She stated: "For example, I really like avocado when it's used in a dish, but I would never eat an avocado alone." In addition, the pairwise comparison experiments were received differently by the participants. Some participants found the experiment quite dull and had the feeling that they answered the same question again and again. Others, however, loved to fill in the questionnaire, stating that they could easily choose between the products since the number of options was limited down to two.

Besides, the advantage of applying Thurstone's Pairwise Comparison Model on the data sets is that the rankings of the products can be obtained in terms of tangible values and the matrix rankings can be verified. Because of the fact that the rankings are assigned values, it can also be determined how big the differences between the rankings of the products are. For example, in the case of the general food products, including intransitivities, the banana and grilled cheese sandwich ended in first and second place respectively based on the matrix rankings. However, the parameters of the scale values of both products turned out to be very close to one another, from which it can be concluded that the banana is only slightly more preferred than the grilled cheese sandwich. Next to that, Thurstone's model also yields extra information about the variability of the preferences of the individuals for certain products by means of the variances of the products. An example of this are the avocado and the marshmallows, both products are not quite preferred in both groups. However, in both groups, the variance of the marshmallows is a lot smaller than the variance of the avocado. This implies that, even though both products are generally not preferred, the judgments concerning the the avocado vary a lot more.

# 6 Conclusion

This thesis has investigated several properties of Thurstone's Pairwise Comparison Model and how it can be applied to model human food preferences.

In the simulation study it was discovered that in order for the estimations of the scale values, on which the rankings are based, to be accurate the distances between the means can not be too large. The maximum distance that may occur between the scale values in order for the estimates to be accurate depends on both the number of objects and the sample size. This not only holds for the most simple case of Thurstone's model, Case V, but was also discovered to be the case for the more complicated variants. In addition, the accuracy of the estimates decrease when the sample size decreases as well.

The two-step optimization process was found to be a suitable method to use for the estimations of the parameters for Case III. The variability in variance was found not to be of influence on the estimations of the parameters, since the means of the parameters were reasonably accurate. The variance of the object itself, however, is slightly of influence on the estimations of the parameters. In general, the larger the variance of an object, the more variability occurs in the estimations for the parameters. For Case II, however, the estimations for the two-step optimization turned out be inaccurate. This primarily had to do with the fact that the correlation was not captured adequately which resulted in the other parameters not being accurately estimated as well.

Next to that, the AIC turned out to be an appropriate method to determine the best case of Thurstone's Pairwise Comparison Model, provided that the sample size is large enough. A sufficiently large sample size is needed so that the parameters and thus rankings can be accurate. This was the case for both the two-step optimization process as the case when the parameters of the distribution are known.

The proposed alternative method for measuring between-group concordance when circular triads are present turned out be an unsuitable method. When the distance between the means was small, the alternative method never yielded a significant result. This was due to the fact that the proportion matrices are dissimilar to one another in the case of a small mean distance and sample size. When the mean distance increased, the alternative method began to perform better. Nevertheless, in order for the alternative method to always yield a significant result, the minimum distance between the smallest and largest mean had to be that large that accurate estimations for the parameters could not be obtained.

The data study performed discovered that Thurstone's Pairwise Comparison model is a reasonable model to use for modelling human food preferences. Of the two sets of food products used in the data study, general food products turned out be more suitable to use in pairwise comparison experiments. Participants in the snack experiment stated that they struggled judging the snacks side by side, since they practically liked all the snacks. The fact that they found it hard to decide which snacks to prefer, resulted in a higher average number of circular triads for the individuals in the snack experiment and more variability in the preferences of the individuals in the snack experiment. The latter was noticeable when the intransitive individuals were excluded from the sample: the rankings of the snacks changed in the groups of the snack experiment, while the rankings of the general food products remained intact.

In addition, the introduction text at the beginning of the experiment did not have any influence on the preferences of the individuals. For both sets of food products, there was no sign of a more significant preference for the individuals in the healthy group for the healthy products. This was also confirmed by Kraemer's test: the between-group concordance was found to be significant in both cases. Nevertheless, the between-group concordance was slightly more significant in the case of the general food products. This can be explained by the fact that the preferences of the individuals in the snack experiment varied a lot from one another, which resulted in a lower concordance between the groups. The results of the snack experiment also showed that one can not determine whether two groups are in concordance with one another just by looking at the rankings of the two groups. Even though the rankings of the snack products of the two groups differed a lot, Kraemer's test still yielded a significant outcome. This confirms that between-group concordance can only be determined by using a non-parametric test.

Finally, the advantage of fitting Thurstone's model on the pairwise comparison data is that the model, in contrast to determining the rankings of the objects by looking at the frequency matrix, yields information about how close the preferences of the individuals for the several objects are by assigning scale values to the food products. Next to that, the more complicated cases of Thurstone's model are also capable of giving more information about the variability in the preferences of the individuals by means of the variances.

# 7 Recommendations for future research

In this report, the parameters for the three cases of Thurstone's Pairwise Comparison Model were estimated by means of least squares minimization. For Case III and II of Thurstone's model, the minimization was performed by using a constrained optimization method. In this case, the cobyla function was utilized in the statistical software program R. However, the optimal solution of the cobyla depended heavily on the starting parameters, yielding different solutions for different initial conditions. For this, a two-step optimization was proposed, which turned out to be a favorable solution for Case III, but not for Case II. Further research may be conducted on the estimation for the parameters of Case II. For this, one might investigate a different software program for the implementation of the models, such as Matlab, which has more and other built-in functions that can be used for constrained optimization. Using another solving method, other than least squares, would probably not be an appropriate solution, since the resulting function that needs to be minimized would still be complicated.

The AIC proved to be a suitable method capable of determining the best case of Thurstone's Pairwise Comparison Model, as long as the sample size is sufficiently large. Nevertheless, in the simulation study, it was discovered that the AIC sometimes yielded a more complicated case as the best model for a data set drawn from a distribution having zero correlation and equal variances. This was explained by the fact that the AIC often prefers a more complicated model. Nevertheless, the AIC still performed quite accurately in the case of equal variances. Next to the AIC, the performance of other model selection criteria such as the Bayesian Information Criterion (BIC) or the Likelihood Ratio (LR), might be investigated as well to see if these are more appropriate.

Thurstone's Model furthermore assumes that the impressions of the objects that arise in the individuals are normally distributed. One might, however, wonder whether this assumption always holds. Other distributions, such as the logistic, exponential or Cauchy distribution, might be more appropriate in some cases. Next to that, it may also be possible for the objects to all have different types of distributions. The correctness of the normality assumption requires further investigation.

A disadvantage of using a pairwise comparison experiment is that individuals are prone to making intransitive judgments. In the case of food products, it turned out that the main cause of the occurrence of circular triads was that participants sometimes did not have a clear preference for one of the two products: the two products were equally liked or disliked. Moreover, a pairwise comparison experiment can take very long to complete, especially when the number of objects is large, since $\frac{n(n-1)}{2}$ comparisons need to be made. In this case, it might be more suitable to, instead of pairwise comparisons, use a ranking experiment in order to obtain the rankings of the food products. In a ranking experiment, an individual ranks the objects from most to least preferred. Furthermore, the ranking data can also be transformed into pairwise comparison data, so that Thurstone's Pairwise Comparison Model can still be applied to obtain the rankings of the objects based on the scale values [17]. In this way, the experiment can be performed quite fast, as only $n$ objects need to be ranked, and circular triads are avoided. In addition, between-group concordance can also be easily determined when a ranking experiment is used.

The introduction text was found not to be of influence on the preferences of the individuals. In this case, the individuals of the tasty and healthy group were independent of one another. Another research subject would be to examine the influence of the introduction text on the preferences of the individuals in only one group. For this, the individuals would have to judge the same set of objects twice, receiving a different introduction text each time before they start the experiment.

Concerning the between-group concordance, it was discovered that there does not exist a test for pairwise comparison experiments that can be used to determine whether the concordance between the groups is significant. A proposed alternative for testing between-group concordance in pairwise comparison experiments has been examined in this report, but unfortunately yielded negative results. Further research for an alternative method for testing between-group concordance in the presence of intransitivities is needed.

In the same simulation study, however, it was also discovered that Kraemer's test always gave a significant result for between-group concordance, even when the distance between the means and the sample size was small. This is peculiar, as two groups do not necessarily have to be in concordance with one another when the objects are identically distributed for the two groups. In the other parts of the simulation study it was discovered that less agreement occurs between the individuals when the mean distance and sample size is small implying that the between-group concordance may not always be significant. Therefore, one would expect that Kraemer's test would sometimes yield a non-significant result. The reason why this did not occur might be that the mean distance has to be really small in order for Kraemer's test to return a non-significant result. Further research on the power of Kraemer's test and what it means for two groups to be in concordance with one another is required.

# References


[1] Akaike Information Criterion. (n.d.). Retrieved from: `http://stanfordphd.com/AIC.html`

[2] Brown, T.C., & Peterson, G.L. (2009). *An enquiry into the method of paired comparison: reliability, scaling, and Thurstone's Law of Comparative Judgment.* Gen Tech. Rep. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 98 p.

[3] Brown, T.C., & Peterson, G.L. (1998). *Economic Valuation by the Method of Paired Comparison, with Emphasis on Evaluation of the Transitivity Axiom.* Land Economics, 74(2), 240-261.

[4] Burros, R.H., & Gibson, W.A. *A Solution for Case III of the Law of Comparative Judgment* Psychometrika (1954) 19: 57.

[5] Burros, R. H. *The application of the method of paired comparisons to the study of reaction potential.* Psychological Review, 1951, 58, 60-66

[6] Cooke, R. M. (1991). *Environmental ethics and science policy series. Experts in uncertainty: Opinion and subjective probability in science.* New York, NY, US: Oxford University Press.

[7] David, H.A. (1988). *The method of paired comparisons.* London: Griffin

[8] Harris, W. P. (1957). *A revised law of comparative judgment.* Psychometrika, 22, 189-198.

[9] Heldsinger, S., & Humphry, S. (2010). *Using the method of pairwise comparison to obtain reliable teacher assessments.* The Australian Educational Researcher, 37(2), 1-19.

[10] Hofacker, C.F. (2009). *Mathematical Marketing* New South Network Services

[11] Iida, Y. (2009). *The number of circular triads in a pairwise comparison matrix and a consistency test in AHP.* Journal of the Operations Research Society of Japan, 52, 174–185

[12] Kendall, M.G. (1975). *Rank Correlation Methods.* Charles Griffin, London.

[13] Kendall, M. G., & Babington-Smith, B. (1940). *On the method of paired comparisons.* Biometrika 31 324-345.

[14] Kendall, M. G., & Babington-Smith, B. (1939). *The Problem of m Rankings.* The Annals of Mathematical Statistics. 10 (3): 275–287.

[15] Kraemer, H. C. (1981). *Intergroup concordance: Definition and estimation.* Biometrika, 68(3), 641-646.

[16] Kurowicka, D., & Cooke, R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling.* John Wiley & Sons.

[17] Maydeu-Olivares, A. (2004). *Thurstone's case V model: A structural equations modeling perspective.* In Recent developments on structural equation models (pp. 41-67). Springer, Dordrecht.

[18] McKenna, S. P., Hunt, S. M., & McEwen, J. (1981). *Weighting the seriousness of perceived health problems using Thurstone's method of paired comparisons.* International journal of epidemiology, 10(1), 93-97.

[19] Mosteller, F. (1951). *Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations.* Psychometrika, 16, 3-9.

[20] Mosteller, F. (1951). *Remarks on the method of paired comparisons: III A test of significance when equal standard deviations and equal correlations are assumed.* Psychometrika 16, 207-218

[21] Pan, D., Lu, X., Liu, J., & Deng, Y. (2014). *A Ranking Procedure by Incomplete Pairwise Comparisons Using Information Entropy and Dempster-Shafer Evidence Theory.* The Scientific World Journal, 2014, 904596.

[22] Powell, M. J. D. (2007). *A view of algorithms for optimization without derivatives.* Cambridge University Technical Report DAMTP 2007.

[23] Serlin, R. C., & Marascuilo, L. A. (1983). *Planned and Post Hoc Comparisons in Tests of Concordance and Discordance for G Groups of Judges.* Journal of Educational Statistics, 8(3), 187-205.

[24] Thurstone, L. L. (1927). *A law of comparative judgment.* Psychological Review, 34(4), 273-286.

[25] Thurstone, L. L. (1928). *Attitudes can be measured.* American Journal of Sociology, 33, 529-554.

[26] Thurstone, L., & Jones, L. (1957). *The Rational Origin for Measuring Subjective Values.* Journal of the American Statistical Association, 52(280), 458-471.

[27] Torgerson, W.S. (1958). *Theory and Methods of Scaling.* John Wiley & Sons, Inc.

[28] Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences* New York: McGraw-Hill.

[29] Siraj, S., Mikhailov, L., & Keane, J. A. (2015). *Contribution of individual judgments toward inconsistency in pairwise comparisons.* European Journal of Operational Research, 242(2), 557–567.

[30] Vojnovic, M., & Yun, S. (2016). *Parameter Estimation for Generalized Thurstone Choice Models.* In Proceedings of ICML 2016, New York, NY.

# Appendices

## A Simulation Study

### A.1 Case V: The effect of both the spacing of the scale values and size of the common variance on the accuracy of the estimation.

In this report, the common variance of the objects has been assumed to be equal to 1, giving that the Law of Comparative Judgment is equal to:

$$S_i - S_j = z_{ij}\sqrt{2}$$

However, it is also possible to assume a different common variances. In that case the estimates of the scale values can still be obtained by the following formula:

$$\widehat{S}_i = \frac{\sqrt{2\sigma^2}}{n}\sum_{j=1}^{n} z_{ij}$$

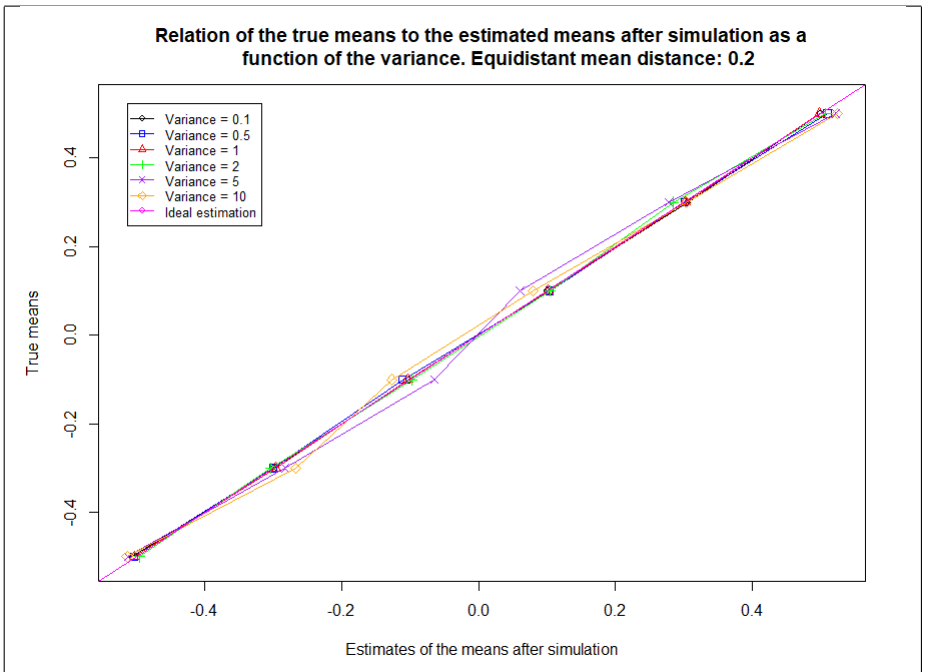
In this section it is investigated how the spacing of the means of the objects and the size of the common variance affects the precision of the estimates. For this, six different variances have been involved in the simulation, which is denoted by the vector $\boldsymbol{\sigma^2}$.

$$\boldsymbol{\sigma^2} = (0.1, 0.5, 1, 2, 5, 10)$$

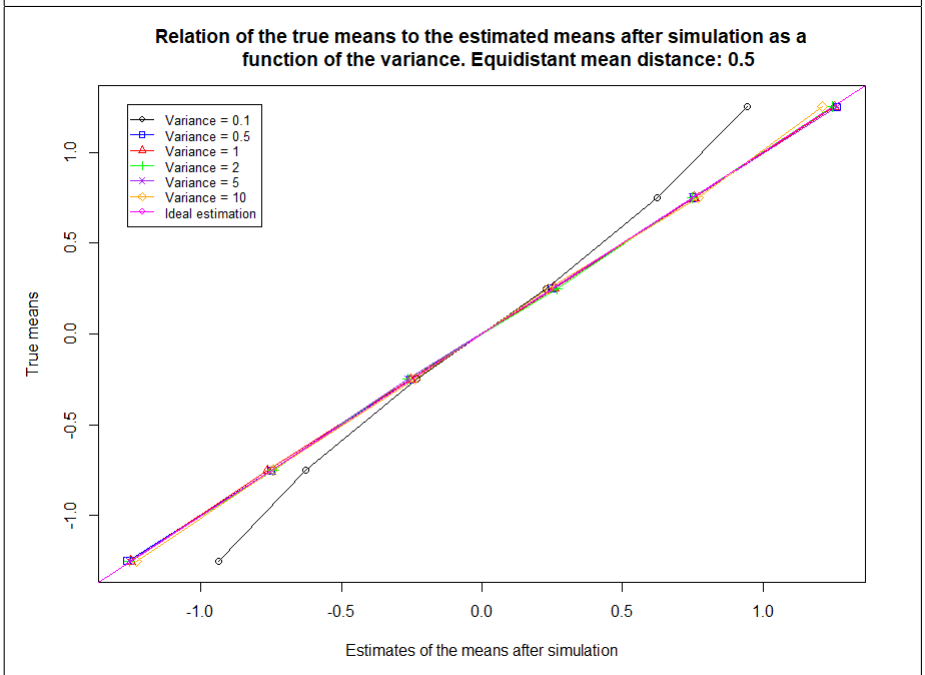
Furthermore, three different types of spacing have been considered: equidistantly spaced, randomly spaced and clustered spacing. The last one asking for some more elaboration.

Clustered spacing of scale values happens when the scale values tend to form groups. This phenomenon occurs when the objects can be divided into groups with common characteristics, where the clusters relate to one another in a certain order. The common characteristics evoke similar impressions within experts, resulting in similar scale values. An example of clustering is a pairwise comparison experiment about the choice and popularity of electronic devices where several mobile phones from the 90's, laptops from the 00's and the newest smartphones participate in the pairwise comparison experiment. One would expect the mobile phones from the 90's to be less popular than both the laptops and smartphones, however the scale values of the mobile phones may be quite close to one another as they stimulate approximately the same feelings within individuals therefore resulting in the scale values of the mobile phones being clustered together. The same can be said about the laptops and the smartphones. Another example of clustering occurrence is when one part of the objects evokes positive feelings while the other evokes negative feelings within the experts resulting in a clear segregation between the positive and negative judged objects.

For the investigation of the equidistantly placed means, the same five mean vectors given in table 3 have been employed. Then, a simulation was performed for the different mean vectors where different common variances were assumed. The sample size used for the simulation is again equal to 1000 and the simulation is repeated twenty-five times. For every equidistantly spaced mean vector, a plot is produced showing the estimation of the scale values with respect to the true means when the common variance is varied. These plots can be seen in figures 23 and 24.




Distance of 0.2 between the means




Distance of 0.5 between the means

Figure 23: The accuracy of the estimated means where the common variance is varied and the distances between the means are equal to 0.2 and 0.5 respectively.

From figure 23, it can be deduced that a small distance between the scale values leads to quite a good approximation of the scale values for almost all variances. When the distance is equal to 0.2, the scale values are estimated best when the variance is small, say 0.1 or 0.5. However, if the variance increases the approximation of the scale values worsens and becomes less stable. This is due to the fact that with increasing variance, more overlap between the distributions takes place resulting in more inconsistent choices which decreases the quality of the fit [29]. This can also be seen in figure 25, where the overlap between the six distributions is plotted for an equidistant mean distance of 0.2 and variances being equal

to 0.1 and 10 respectively.

So in the case of a mean distance as small as 0.2, a small variance is needed in order to obtain accurate estimates of the scale values.

When the mean distance increases to 0.5 all variances seem to produce decent estimates of the means, except when the variance is equal to 0.1. In that case, the means are underestimated especially at the ends of the scaling spectrum. Why is this happening? According to Brown and Peterson (2009) [2], scale values are over- or underestimated for objects that lie on the edges of the spectrum as the paired comparison data is less rich at the end of the scaling spectrum due to the lack of objects near the most and least preferred objects, thus resulting in a limited number of choices. The experts practically always choose the most preferred object over the other objects and the other way around for the least preferred object such that variety is almost absent. Variety in choices however are needed to construct the scale values for the impressions of the objects [19]. Over- or underestimating the scale values at the end of the spectrum is a general feature of pairwise comparisons [2].

Nevertheless, the lack of other objects is not the only cause of the underestimation of the estimates at the end of the spectrum. If it were the only case, one could wonder why it didn't occur that obvious in the earlier simulations. Looking at the previous simulations, it is indeed true that the estimates deviate a little bit more from the true means at the end of the spectrum, but not by that much. In this case, the deviations from the true means at the end of the spectrum are quite noticeable, implying that another unknown factor contributes to this phenomenon. This factor has to do with the fact that the interval on which the scale values are defined is too large to get an accurate estimation of the scale values, by the limit effect of the inverse of the standard normal deviate. When the common variance equals 0.1, the scale values of the objects are estimated by the following formula:

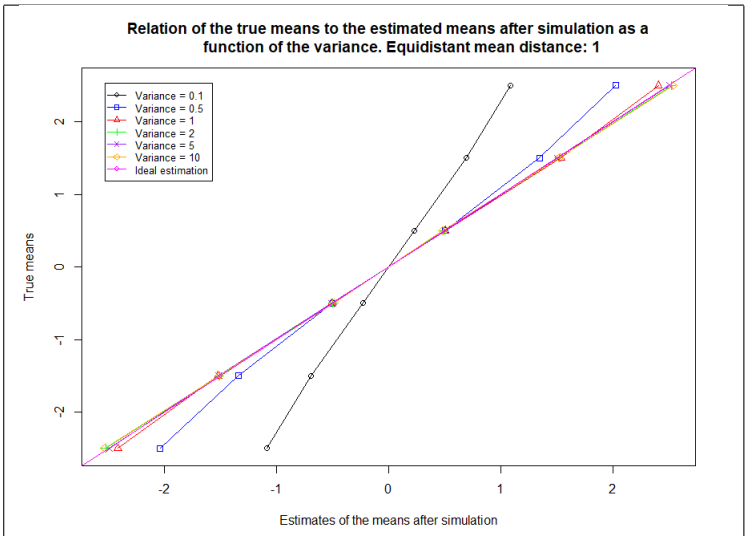
$$\widehat{S}_i = \frac{\sqrt{0.2}}{n} \sum_{j=1}^{n} z_{ij}.$$

As the sample size is equal to 1000 and six objects are involved, the maximum distance between the scale values of the most and least preferred objects that may occur such that estimates are still quite accurate equals:

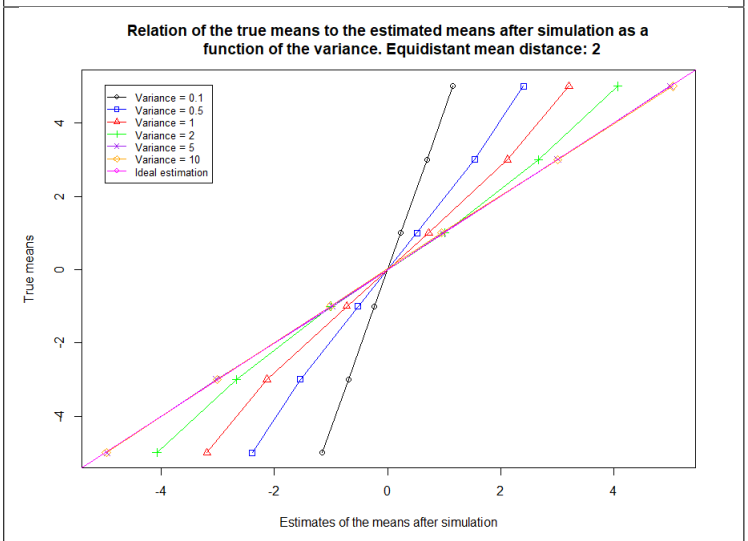
$$\sqrt{0.2}\left(\frac{5 \cdot \Phi^{-1}(0.999)}{6} - \frac{5 \cdot \Phi^{-1}(0.001)}{6}\right) \approx 2.303.$$

For the mean vector with equidistant mean spacing of 0.5, the distance between the scale values of the most and least preferred stimuli equals 2.5, which is more than the permissible distance, resulting in a decrease in the accuracy of the scale values.

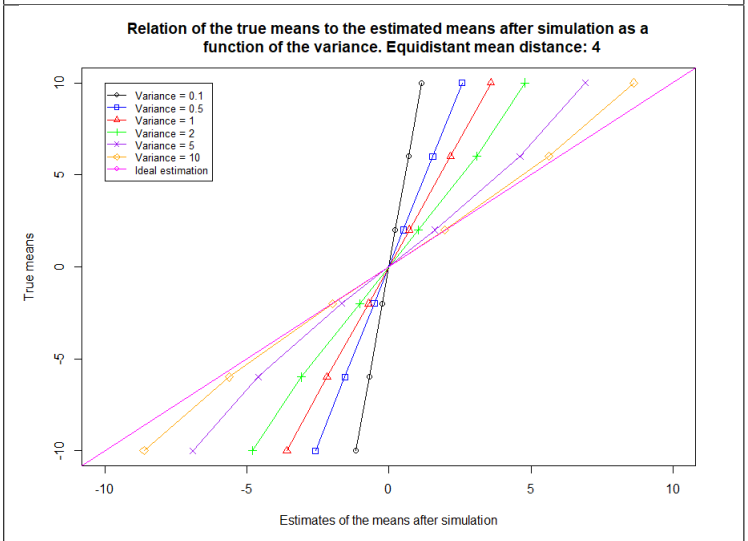
If the distance between the scale values equals 1, the accuracy of the estimates decreases when the variance is small (equal to 0.1 or 0.5) as the distance of the interval where the scale values are defined exceeds the permitted maximum distance of the scale values determined by the size of the variance. The distance between the scale values of the most and least preferred objects is equal to 5 in this case, whereas the maximum distances that are permitted such that reliable estimates can be obtained equals 2.303 when the variance is 0.1 and 7.284 when the variance equals 1. This explains the fact why the estimation of the scale values, when the means are equidistantly spaced with distance 1, is underestimated if the variance is small.
The last remark proves that the maximum allowed distances between the scale values at the end of the spectrum for different variances do not only depend on the number of objects and the sample size, but also on the size of the variance. Table 16 shows the maximum distance between the scale values that can be estimated when six objects are judged by a sample of 1000 experts for the six different variance sizes and the maximum equidistant spacing that can occur between the six stimuli such that the estimation of the means is still quite accurate.




Distance of 1 between the means

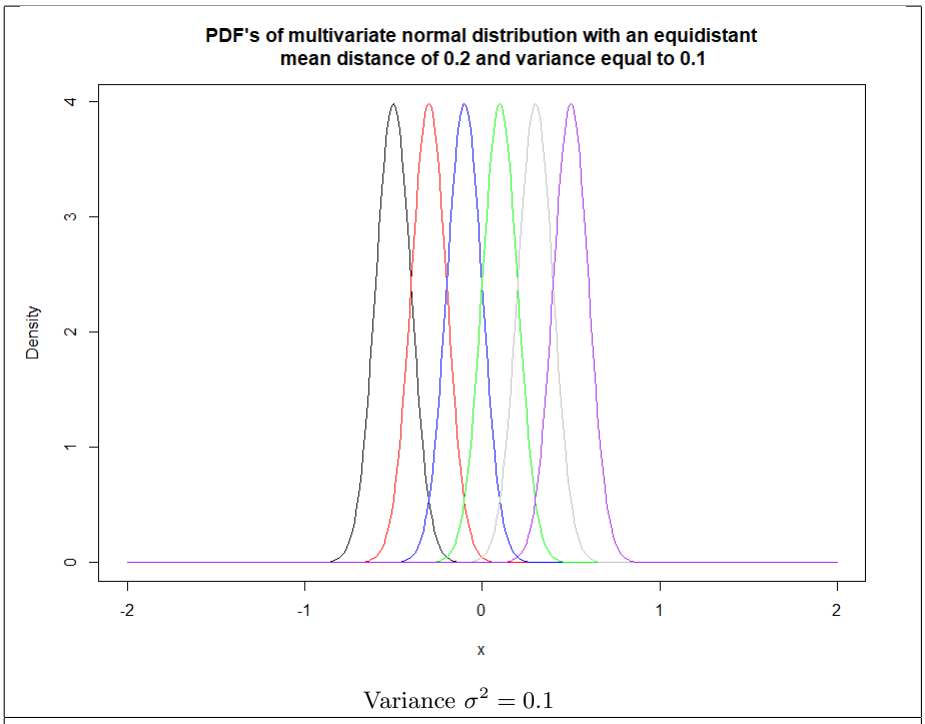



Distance of 2 between the means

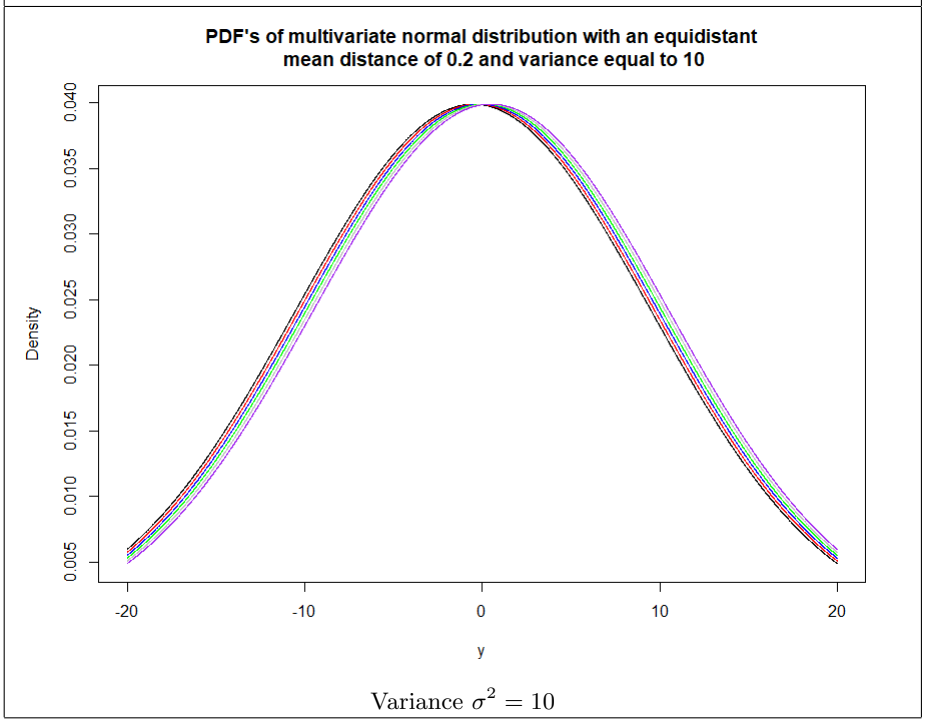



Distance of 4 between the means

Figure 24: The accuracy of the estimated means where the common variance is varied and the distances between the means are equal to 1, 2 and 4 respectively.




Variance $\sigma^2 = 0.1$




Variance $\sigma^2 = 10$

Figure 25: Plot of the densities of the sensations of the stimuli when variance is assumed to be equal to 0.1 (left) and 10 (right).

| Variance $\sigma^2$ | Maximum obtainable distance | Maximum equidistant spacing |
|---|---|---|
| $\sigma^2 = 0.1$ | 2.303 | 0.4606 |
| $\sigma^2 = 0.5$ | 5.150 | 1.03 |
| $\sigma^2 = 1$ | 7.284 | 1.457 |
| $\sigma^2 = 2$ | 10.301 | 2.006 |
| $\sigma^2 = 5$ | 16.287 | 3.257 |
| $\sigma^2 = 10$ | 23.033 | 4.607 |

Table 16: Maximum distance that can be obtained between the scale values depending on the variance $\sigma^2$, when the sample size is equal to 1000 and six stimuli are compared.

Furthermore, the limiting effect is also the reason why the estimation of the means becomes even worse for smaller variances when the distance between the true means increases to 2 or 4. The means are too far away spaced, such that reliable estimates can not be obtained.

One may wonder, however, why in the case of a large variance and small distance between the means the estimations of the scale values are still pretty accurate. Figure 25 shows that the distributions almost fully overlap when the variance is large. This implies that the number of inconsistent choices increases, which would decrease the quality of the fit [29]. The fact that this is not the case, might be explained by looking at the formula that is used for the estimation of the scale values. The scale values might be underestimated in the case of a larger variance, however the average of the $z_{ij}$'s are multiplied by the square root of two times the common variance which brings the estimates closer to the actual values.

For the investigation of the random place of the scale values, the same mean vectors as in table 5, have been used for simulation. The simulation results were almost exactly the same when compared to the case where scale values were equidistantly spaced and no new things were discovered. Therefore, it has been chosen not to elaborate on the results.

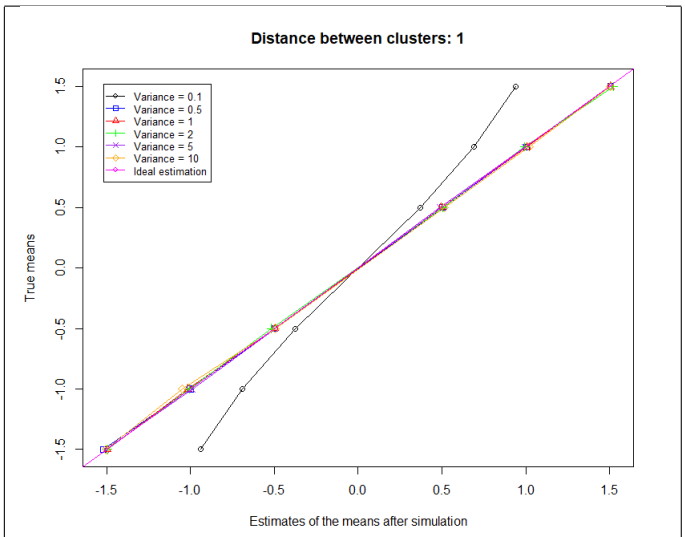
In order to examine the effect of clustering on the estimation of the scale values, the set of six objects is clustered into two groups of three. Two settings are considered regarding the clustering: equidistant clustering where the distance between the scale values within the two clusters are evenly spaced and random clustering where the scale values are randomly spaced within the two clusters.

The distance between the scale values equals 0.5 in all cases for the equidistant clustering and the distance between the clusters increases. The three mean vectors used to simulate are given in table 17

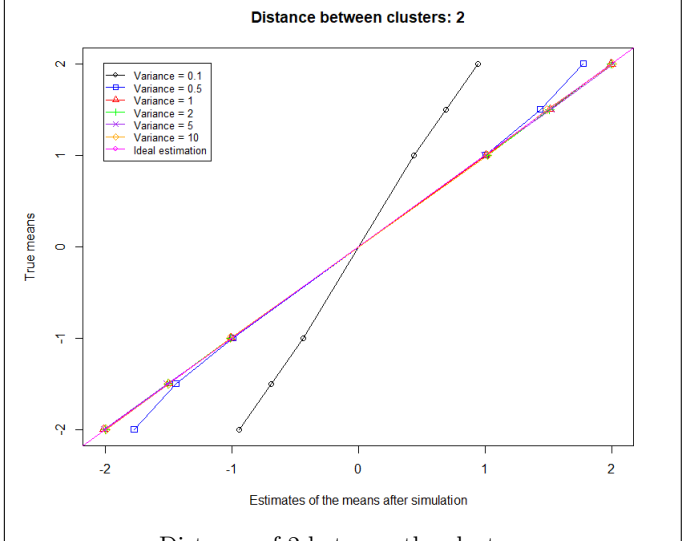
| **S** | Mean vector | Cluster distance |
|---|---|---|
| $\mathbf{S_9}$ | (-1.5, -1, -0.5, 0.5, 1, 1.5) | 1 |
| $\mathbf{S_{10}}$ | (-2, -1.5, -1, 1, 1.5, 2) | 2 |
| $\mathbf{S_{11}}$ | (-3, -2.5, -2, 2, 2.5, 3) | 4 |

Table 17: The three equidistantly clustered means vectors used in the simulation.

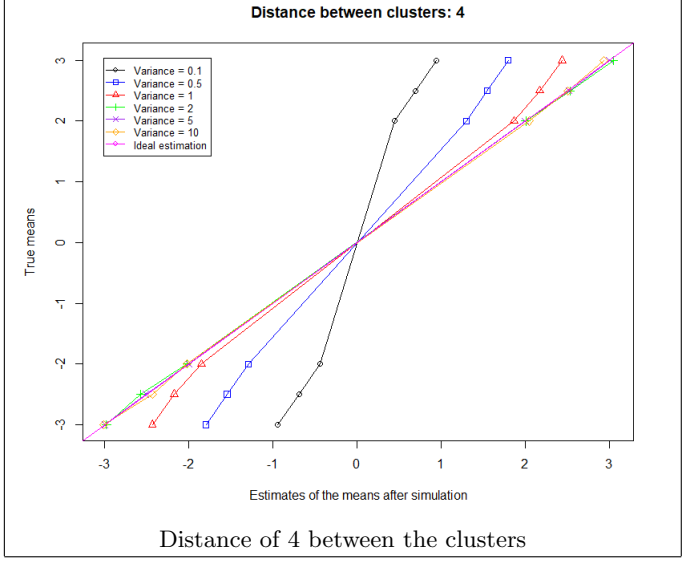
The results of the simulation can be seen in figure 26, where the estimates of the means are plotted against the true means. Equidistantly clustered scale values produce results similar to the case where the means are equidistantly or randomly spaced. The accuracy of the estimates decreases when the cluster distance increases and variance's small by the same reasoning that the true means can not be attained.




Distance of 1 between the clusters




Distance of 2 between the clusters




Distance of 4 between the clusters

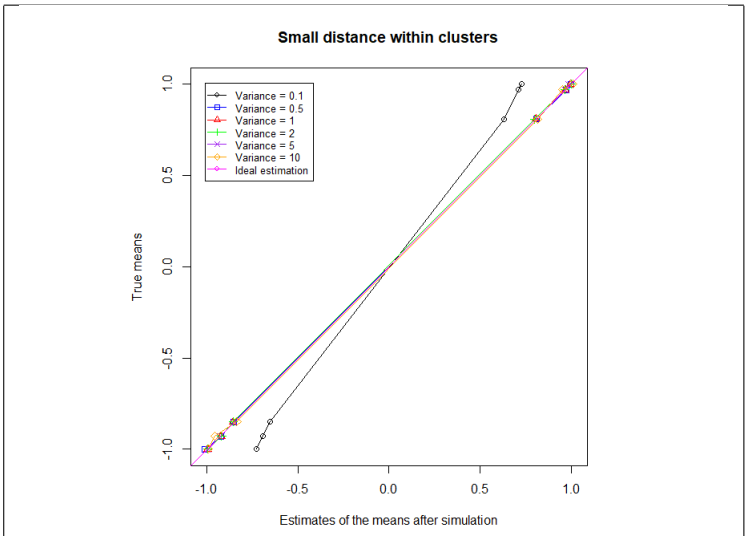
Figure 26: The influence of equidistantly clustered spacing and the size of the variance on the accuracy of the estimation of the scale values.

Next to that, two random clustered vectors have been defined, where the distance of the scale values within the clusters differs while keeping the interval on which the scales are defined constant. The two vectors used for simulation can be seen in table 18.

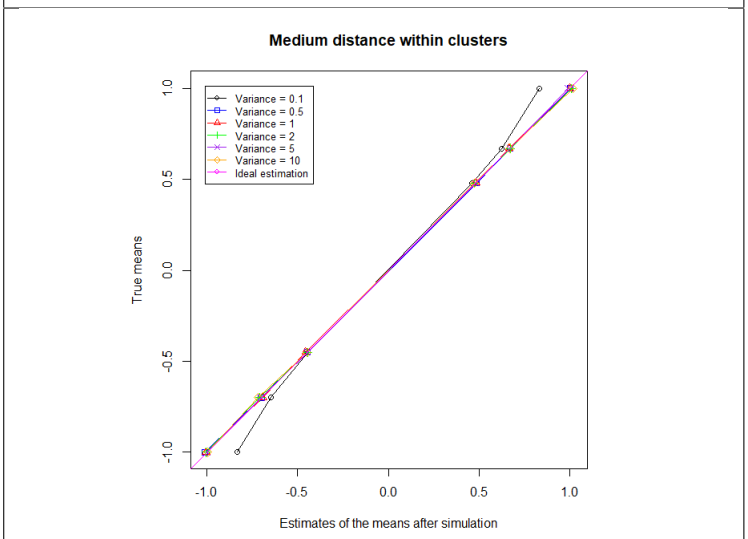
| **S** | Mean vector |
|---|---|
| $\mathbf{S_{12}}$ | (-1, -0.93, -0.85, 0.81, 0.97, 1) |
| $\mathbf{S_{13}}$ | (-1, -0.7, -0.45, 0.48, 0.67, 1) |

Table 18: Two randomly clustered vectors used for simulation.

The results of this simulation can be found in figure 27.




Small distance within the clusters




Medium distance within the clusters

Figure 27: The influence of the cluster distance and variance size on the accuracy of the estimated means.

Looking at the accuracy of the estimates in figure 27, the estimates are quite precise, except when the variance equals 0.1. This is because of the fact that the maximum interval distance of the scale values that is allowed is almost attained at that point. Furthermore, figure 27 also shows that in the case of a small variance, the scale values are better estimated if they are spaced further apart from one another. This makes sense as the distributions have less overlap when they are spaced further apart, resulting in a better estimation.

## A.2 Case II: Investigating the effect of correlation

Case II includes correlation between the objects, next to unequal variances and different means. Due to computational reasons, it is assumed that this correlation is constant for all pairs of objects [19]. This section explores the influence of the correlation and the variance on the sample's response. For this, two objects have been taken into consideration with the following means:

$$\begin{pmatrix} S_1 \\ S_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

The means will be held constant during the entire simulation.

In order to examine the influence of the variance and correlation on the sample's judgment, kernel density estimate plots will be produced by means of simulating from a bivariate normal distribution. Three different variance settings are used for the sampling of the bivariate normal distribution. The first setting assigns variance 1 to both objects, the second setting assigns variance 0.5 to the first object and variance 1.5 to the second while the third variance setting prescribes the variance exactly the other way around. Next to that, the correlation is varied as well so that the effect of the latter can be captured. Note that the correlation must be chosen between the interval $(-\frac{1}{n-1}, 1)$, if $n$ objects are compared, otherwise the covariance matrix won't be positive definite [16]. The following correlations have been used in the simulation:

$$\rho = (-0.5, -0.25, 0, 0.25, 0.5, 0.75)$$

Some of the produced kernel density estimate plots are listed in the figures 28, 29 and 30 below. These plots give an idea how the variance and correlation influences the choice between a pair of two objects. Not all of the kernel density estimates will be given as the other plots yield more or less the same outcome.

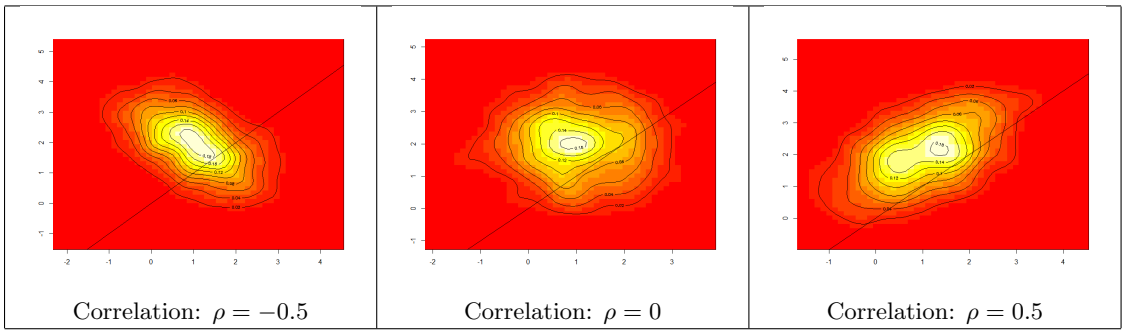



Figure 28: Kernel density estimate plots of the first variance setting: both stimuli have variance equal to 1.

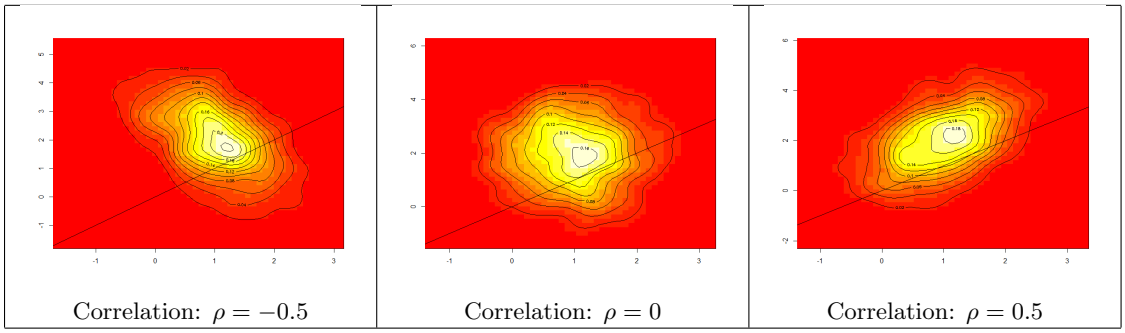



Figure 29: Kernel density estimate plots of the second variance setting: the first object having variance $\sigma_1^2 = 0.5$, the second object having variance $\sigma_2^2 = 1.5$.

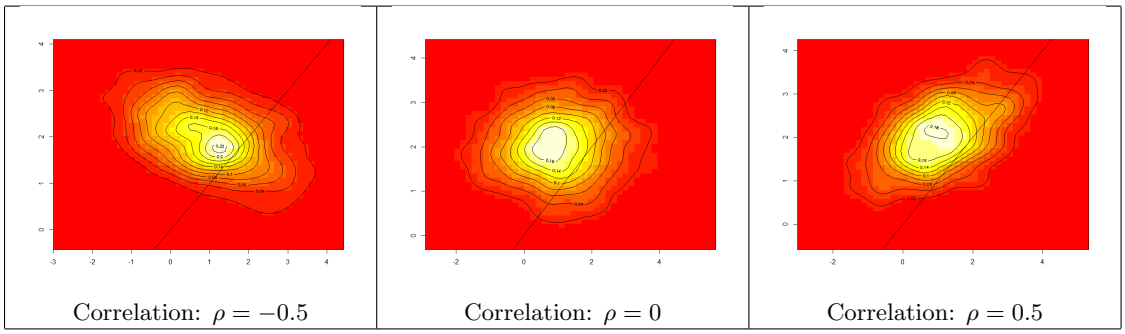



Figure 30: Kernel density estimate plots of the third variance setting: the first object having variance $\sigma_1^2 = 1.5$, the second object having variance $\sigma_2^2 = 0.5$.

The x-axis of the kernel density plot in figures 28, 29 and 30 represents the continuum on which the distribution of the first object, with mean $S_1 = 1$, is defined. The y-axis, on the other hand, is the interval where the distribution of the second object takes on its values. The kernel density estimate plots produce ellipse-shaped contours, where all the points that lie on the contour have the same probability of occurring. The color within the interior of two contours describes the magnitude of the probability: the more yellow the color within the interior the higher the probability. Furthermore, the line $y = x$ is drawn within the kernel density plot to display the division between the two objects. The area above the line is the area where the second object is judged greater than the first and the area below the line is the area where the first object is judged greater. The line itself is the boundary between the two objects: the two objects are then regarded equal.

From figures 28, 29 and 30, it can be deduced that in all three different cases the probability that the first object is judged greater than the second is highest when the correlation is negative. When the correlation is positive, the probability that the first object is greater than the second is the smallest. The probabilities that the first object is judged greater than the second are given in table 19:

| | $\sigma_1^2 = \sigma_2^2 = 1$ | $\sigma_1^2 = 0.5, \sigma_2^2 = 1.5$ | $\sigma_1^2 = 1.5, \sigma_2^2 = 0.5$ |
|---|---|---|---|
| $\rho = -0.5$ | 0.282 | 0.289 | 0.289 |
| $\rho = 0$ | 0.240 | 0.240 | 0.240 |
| $\rho = 0.5$ | 0.159 | 0.128 | 0.128 |

Table 19: Probability that the first object is judged greater than the second in the three cases.

## A.3 Determining the average number of circular triads depending on the number of objects and sample size for a sample having no preference

One of the questions to be answered in the data study is how both the similarity of food products and the introduction text of the experiment affect the degree of transitivity in judgments made by the individual. Section 3 has shown that it is possible to determine whether an individual is too random in his preference by looking at the number of circular triads which merely depends on the number of objects being compared. However, the question becomes what is the average number of triads an individual produces when his preference is completely random and how does this depend on the number of objects and sample size?

For this purpose, the average $\zeta$ value for a sample of $N$ individuals is computed, where the individuals have no preference. This means that the probability for one of the both objects to be preferred over the other is equal to 0.5 for both objects. This is modelled by simulating from a uniform $[0, 1]$ distribution.

Then, the coefficient of consistence $\zeta$ is computed for every individual in the sample and all the coefficients averaged. This results in an average coefficient of consistence for the sample. Moreover, the procedure described above is repeated twenty-five times in order to get an average of the average coefficients of consistence, since the coefficients may differ per simulation due to random errors. Furthermore, the minimum and maximum values of the average coefficient of consistence from the ten simulations are saved as well. The resulting average of the coefficient of consistence for a random judgment may serve as a guideline for the minimum rejection region. If the coefficient of consistence $\zeta$ for an individual is close to the average coefficient of consistence for a random preference, say between the minimum and maximum value of the coefficient of consistence, or even below the minimum value, the individual can be regarded as being random in his preference and thus dropped from the sample. Note that the outcome of this simulation is based on the worst case scenario, when the individual is totally random. This, however, may not be very realistic as it might be that even individuals with higher $\zeta$ scores, who are not totally random, still make too many intransitive choices and should therefore be excluded from the sample. Therefore it is advised to still use the method of Kendall and Babington [13] when one wants to test whether an individual's preference is too random.

The average $\zeta$ values obtained from the twenty-five simulations together with the minimum and maximum $\zeta$ values for three, five, seven and ten objects and a sample size varying between 10 and 250 are displayed in figure 31.

From figure 31 it can be seen that the average value for the coefficient of consistence when judgments are made at random remains approximately constant for all sample sizes, regardless the number of objects. The average coefficient of consistence is larger when the number of objects involved in the simulation is smaller. This makes sense, as the probability of making an intransitive choice within a pairwise comparison experiment increases when the number of objects increases [29]. Next to that, the minimum and maximum value of the coefficient of consistence seems to converge to the average coefficient of consistence when the sample size becomes larger. This is especially visible when the number of objects is equal to seven and ten. Thus besides the average coefficient of consistence for a random judgment being almost constant, the minimum and maximum value for the coefficient of consistence will stabilize as well when the sample size becomes larger.

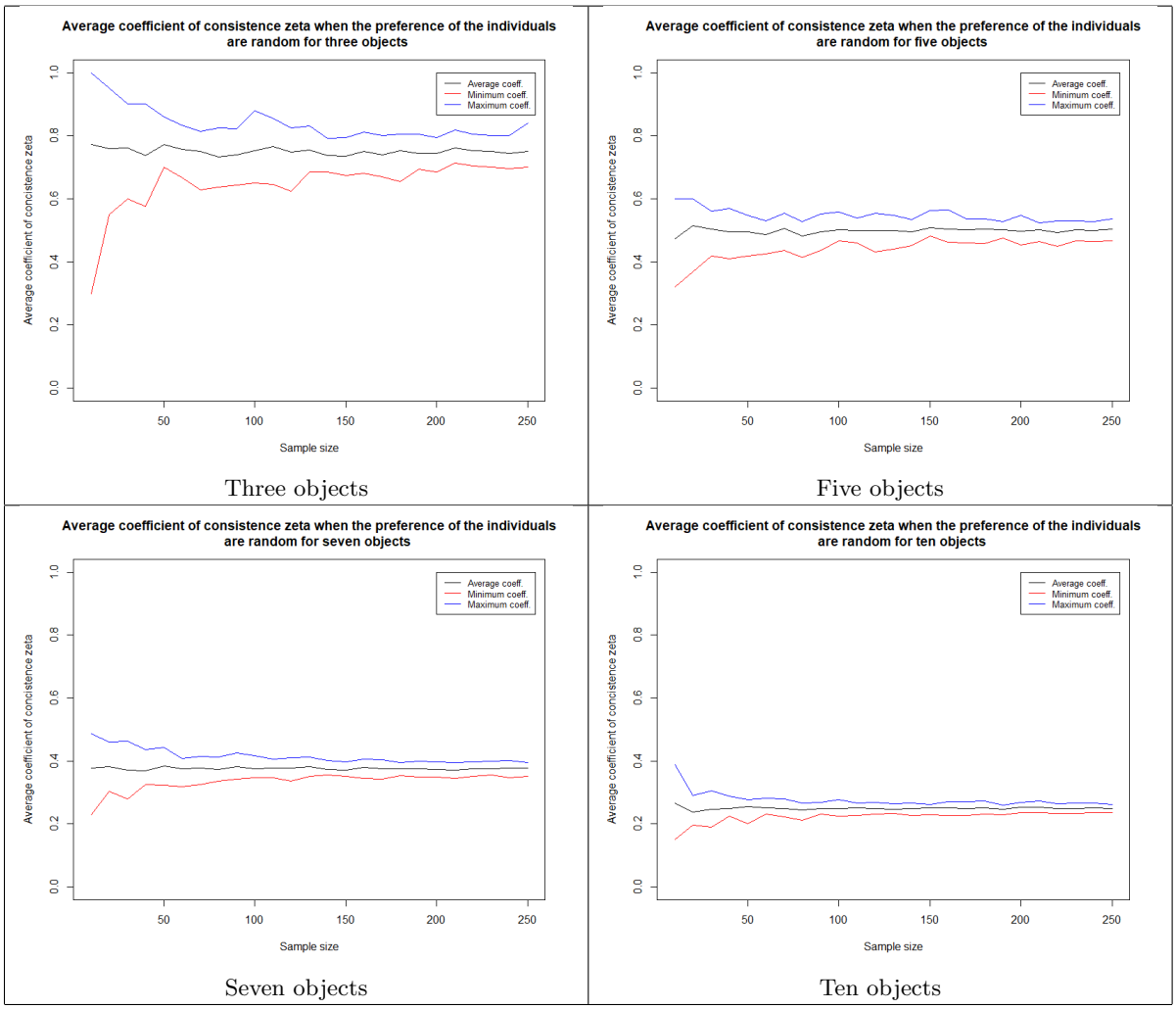



Figure 31: Average coefficient of consistence zeta for a random preference depending on the sample size and number of objects.

## A.4 Simulation study Case III

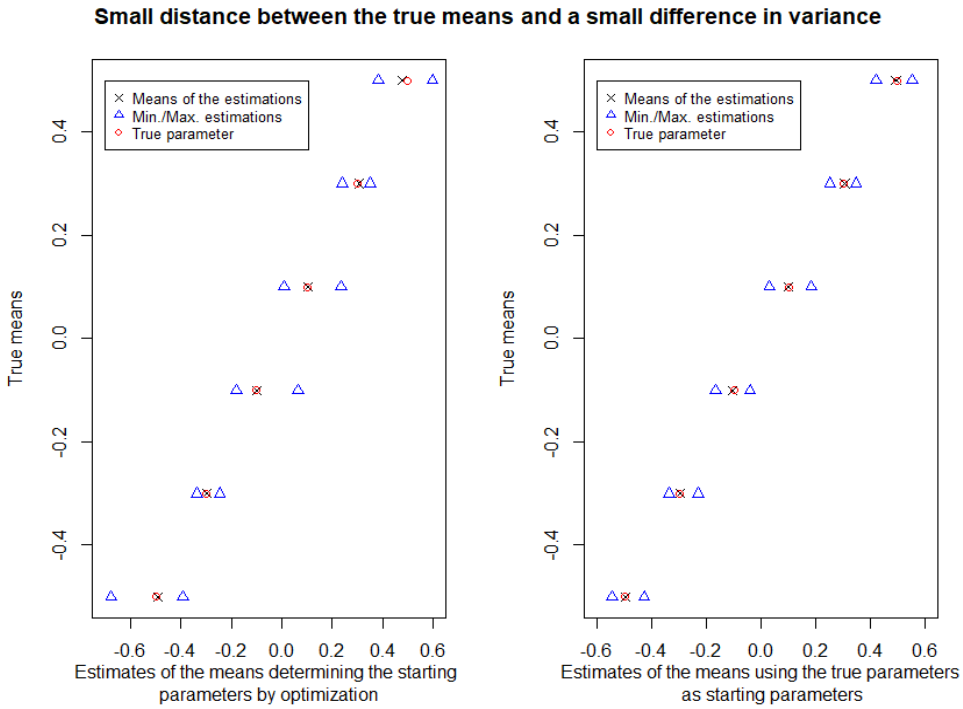



Figure 32: The accuracy of the estimations of the means for both optimization routines, when the distance between the means is small and the variability in variance is low.

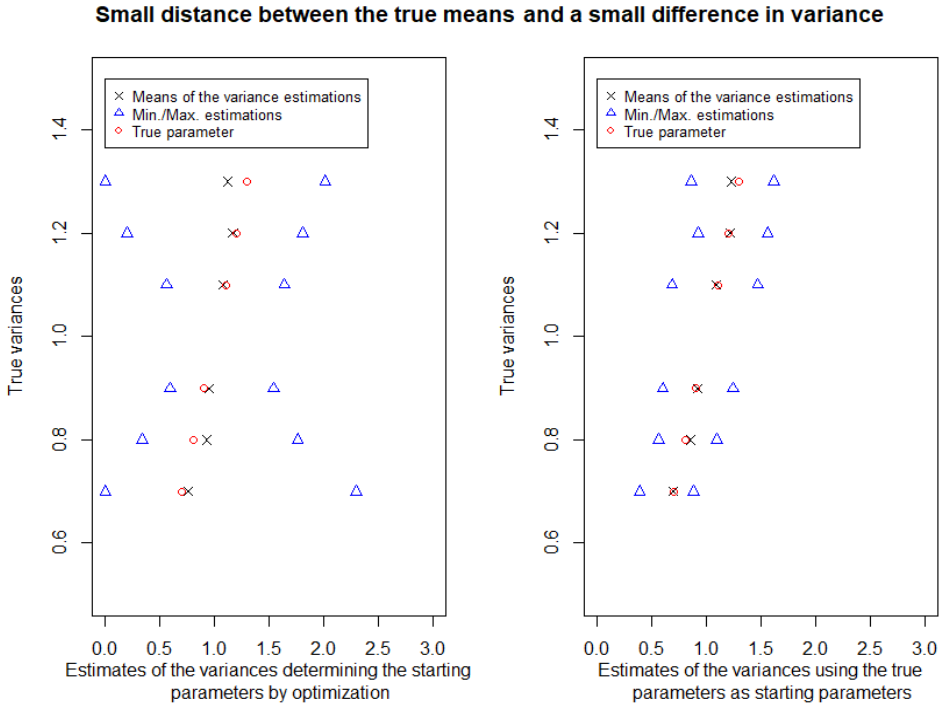



Figure 33: The accuracy of the estimations of the variances for both optimization routines, when the distance between the means is small and the variability in variance is low.

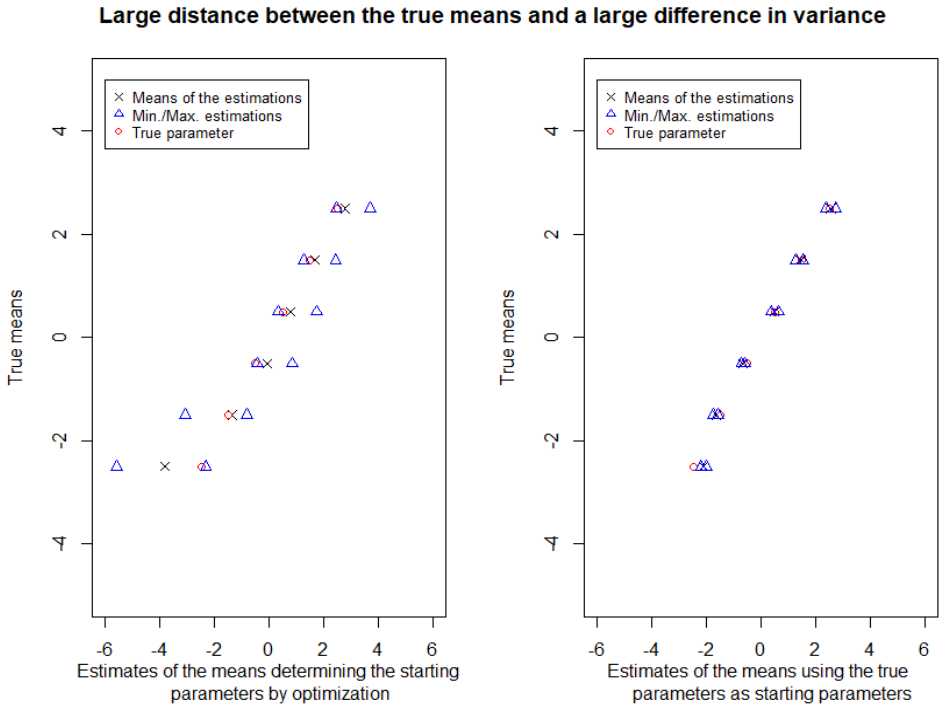



Figure 34: The accuracy of the estimations of the means for both optimization routines, when the distance between the means is large and the variability in variance is high.

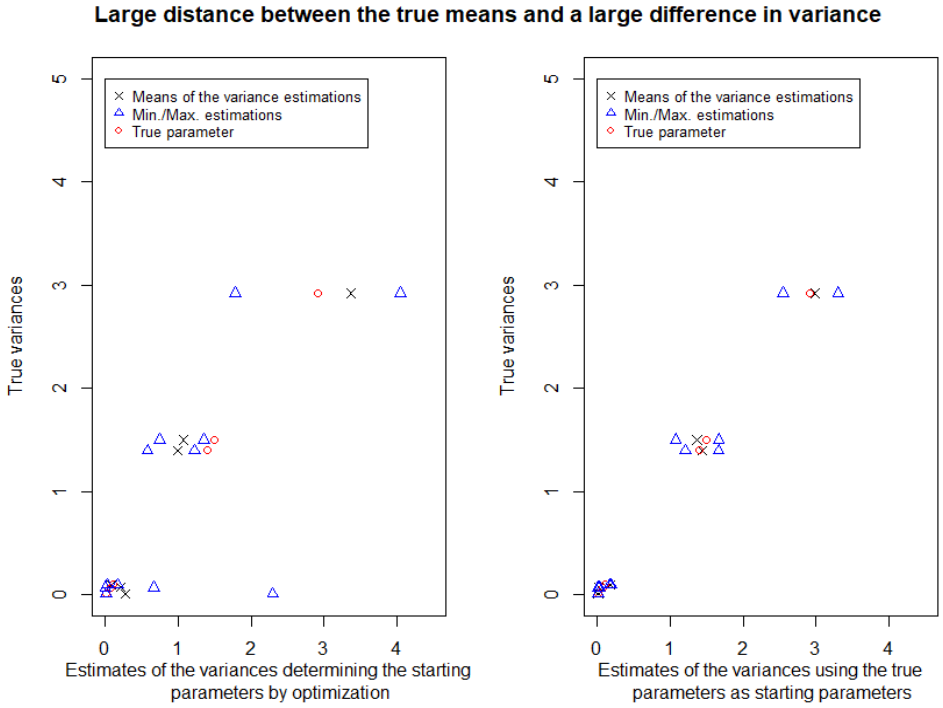



Figure 35: The accuracy of the estimations of the variances for both optimization routines, when the distance between the means is large and the variability in variance is high.

# B Data study

## B.1 Results of the general food products

| | Pizza | Avocado | Unseasoned nuts | Banana | Marshmallows | Grilled sand. | Turkey |
|---|---|---|---|---|---|---|---|
| Pizza | - | 34 | 29 | 20 | 49 | 25 | 35 |
| Avocado | 21 | - | 23 | 15 | 35 | 19 | 25 |
| Unseasoned nuts | 26 | 32 | - | 18 | 45 | 22 | 35 |
| Banana | 35 | 40 | 37 | - | 51 | 30 | 41 |
| Marshmallows | 6 | 20 | 10 | 4 | - | 3 | 10 |
| Grilled sand. | 30 | 36 | 33 | 25 | 52 | - | 39 |
| Turkey | 20 | 30 | 20 | 14 | 45 | 16 | - |

Table 20: Frequency matrix of the tasty group, including intransitivities (row elements are preferred over the column elements).

| | Pizza | Avocado | Unseasoned nuts | Banana | Marshmallows | Grilled sand. | Turkey |
|---|---|---|---|---|---|---|---|
| Pizza | - | 33 | 38 | 35 | 44 | 25 | 42 |
| Avocado | 17 | - | 21 | 20 | 29 | 14 | 28 |
| Unseasoned nuts | 12 | 29 | - | 18 | 40 | 12 | 35 |
| Banana | 15 | 30 | 32 | - | 42 | 16 | 39 |
| Marshmallows | 6 | 21 | 10 | 8 | - | 3 | 16 |
| Grilled sand. | 25 | 36 | 38 | 34 | 47 | - | 45 |
| Turkey | 8 | 22 | 15 | 11 | 34 | 5 | - |

Table 21: Frequency matrix of the healthy group, including intransitivities (row elements are preferred over the column elements).

| Food product | Sum tasty group | Sum healthy group | Matrix rank tasty | Matrix rank healthy |
|---|---|---|---|---|
| Pizza | 192 | 217 | 3 | 2 |
| Avocado | 138 | 129 | 6 | 5 |
| Unseasoned nuts | 178 | 146 | 4 | 4 |
| Banana | 234 | 174 | 1 | 3 |
| Marshmallows | 53 | 64 | 7 | 7 |
| Grilled sand. | 215 | 225 | 2 | 1 |
| Turkey | 145 | 95 | 5 | 6 |

Table 22: Number of times the food products were preferred over any other food product and resulting matrix rankings of the products for both groups, including intransitivities.

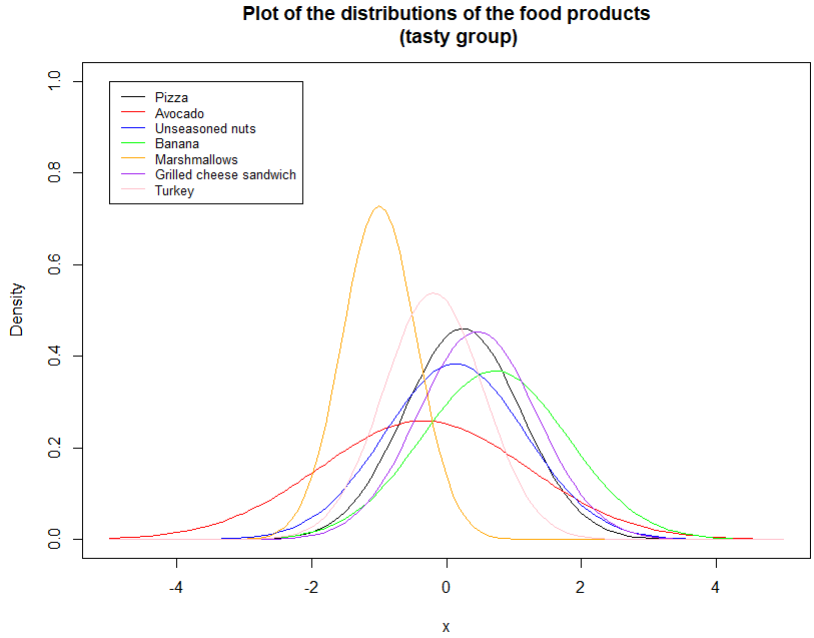



Figure 36: Distributions of the general food products for the tasty group, including intransitivities.

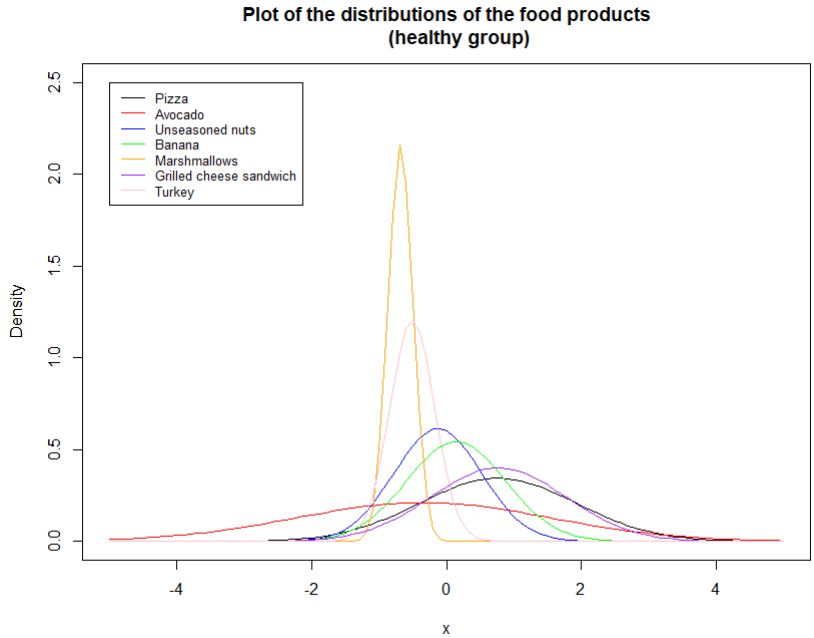



Figure 37: Distributions of the general food products for the healthy group, including intransitivities.

| | Pizza | Avocado | Unseasoned nuts | Banana | Marshmallows | Grilled sand. | Turkey |
|---|---|---|---|---|---|---|---|
| Pizza | - | 24 | 19 | 13 | 35 | 18 | 25 |
| Avocado | 16 | - | 16 | 10 | 25 | 16 | 19 |
| Unseasoned nuts | 21 | 24 | - | 14 | 35 | 16 | 23 |
| Banana | 27 | 30 | 26 | - | 37 | 23 | 30 |
| Marshmallows | 5 | 15 | 5 | 3 | - | 1 | 8 |
| Grilled sand. | 22 | 24 | 24 | 17 | 39 | - | 28 |
| Turkey | 15 | 21 | 17 | 10 | 32 | 12 | - |

Table 23: Frequency matrix of the tasty group, without intransitivities (row elements are preferred over the column elements).

| | Pizza | Avocado | Unseasoned nuts | Banana | Marshmallows | Grilled sand. | Turkey |
|---|---|---|---|---|---|---|---|
| Pizza | - | 27 | 26 | 26 | 34 | 17 | 32 |
| Avocado | 11 | - | 15 | 15 | 21 | 10 | 23 |
| Unseasoned nuts | 12 | 23 | - | 13 | 29 | 7 | 28 |
| Banana | 12 | 23 | 25 | - | 32 | 11 | 30 |
| Marshmallows | 4 | 17 | 9 | 6 | - | 2 | 12 |
| Grilled sand. | 21 | 28 | 31 | 27 | 36 | - | 35 |
| Turkey | 6 | 15 | 10 | 8 | 26 | 3 | - |

Table 24: Frequency matrix of the healthy group, without intransitivities (row elements are preferred over the column elements).

| Food product | Sum tasty group | Sum healthy group | Matrix rank tasty | Matrix rank healthy |
|---|---|---|---|---|
| Pizza | 134 | 162 | 3 | 2 |
| Avocado | 102 | 95 | 6 | 5 |
| Unseasoned nuts | 133 | 112 | 4 | 4 |
| Banana | 173 | 133 | 1 | 3 |
| Marshmallows | 37 | 50 | 7 | 7 |
| Grilled sand. | 154 | 178 | 2 | 1 |
| Turkey | 107 | 68 | 5 | 6 |

Table 25: Number of times the food products were preferred over any other food product and resulting matrix rankings of the products for both groups, without intransitivities.

| | Means tasty | Means healthy | Var. tasty | Var. healthy | Rank. tasty | Rank. healthy |
|---|---|---|---|---|---|---|
| Pizza | 0.169 | 0.763 | 0.697 | 1.615 | 3 | 2 |
| Avocado | -0.322 | -0.384 | 2.784 | 2.691 | 6 | 5 |
| Unseasoned nuts | 0.156 | -0.118 | 0.842 | 0.508 | 4 | 4 |
| Banana | 0.754 | 0.191 | 1.176 | 0.717 | 1 | 3 |
| Marshmallows | -0.980 | -0.843 | 0.229 | 0.390 | 7 | 7 |
| Grilled sand. | 0.391 | 0.955 | 0.513 | 1.056 | 2 | 1 |
| Turkey | -0.168 | -0.564 | 0.693 | 0.023 | 5 | 6 |

Table 26: Means, variances and rankings of the general food products for both groups, without intransitivities.

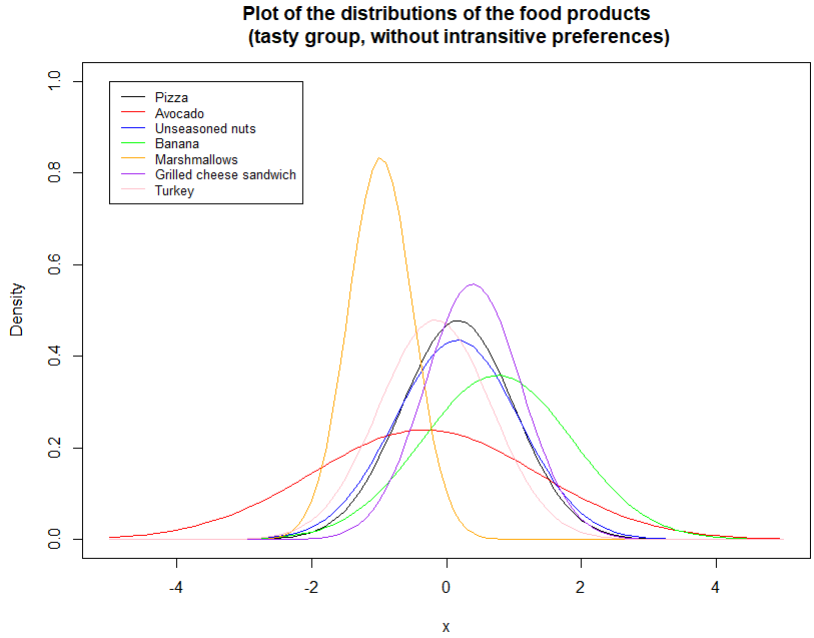



Figure 38: Distributions of the general food products for the tasty group, without intransitivities.

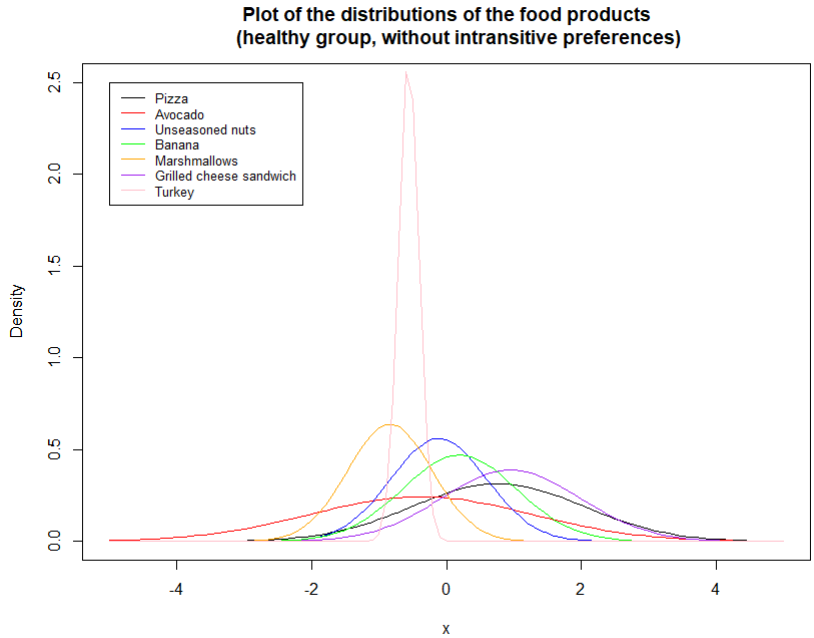



Figure 39: Distributions of the general food products for the healthy group, without intransitivities.

## B.2 Results of the similar products

| | Crisps | Almond cookie | Rice cookie | Snickers | Sponge cake | Muesli bar | Apple |
|---|---|---|---|---|---|---|---|
| Crisps | - | 30 | 44 | 30 | 28 | 34 | 31 |
| Almond cookie | 28 | - | 41 | 31 | 22 | 35 | 24 |
| Rice cookie | 14 | 17 | - | 17 | 11 | 25 | 7 |
| Snickers | 28 | 27 | 41 | - | 20 | 35 | 25 |
| Sponge cake | 30 | 36 | 47 | 38 | - | 41 | 27 |
| Muesli bar | 24 | 23 | 33 | 23 | 17 | - | 18 |
| Apple | 27 | 34 | 51 | 33 | 31 | 40 | - |

Table 27: Frequency matrix of the tasty group, including intransitivities (row elements are preferred over the column elements).

| | Crisps | Almond cookie | Rice cookie | Snickers | Sponge cake | Muesli bar | Apple |
|---|---|---|---|---|---|---|---|
| Crisps | - | 56 | 57 | 52 | 45 | 54 | 34 |
| Almond cookie | 23 | - | 38 | 31 | 20 | 33 | 15 |
| Rice cookie | 22 | 41 | - | 39 | 24 | 39 | 14 |
| Snickers | 27 | 48 | 40 | - | 27 | 41 | 24 |
| Sponge cake | 34 | 59 | 55 | 52 | - | 49 | 24 |
| Muesli bar | 25 | 46 | 40 | 38 | 30 | - | 20 |
| Apple | 45 | 64 | 65 | 55 | 55 | 59 | - |

Table 28: Frequency matrix of the healthy group, including intransitivities (row elements are preferred over the column elements).

| Food product | Sum tasty group | Sum healthy group | Matrix rank tasty | Matrix rank healthy |
|---|---|---|---|---|
| Crisps | 197 | 298 | 3 | 2 |
| Almond cookie | 181 | 160 | 4 | 7 |
| Rice cookie | 91 | 179 | 7 | 6 |
| Snickers | 176 | 207 | 5 | 4 |
| Sponge cake | 219 | 273 | 1 | 3 |
| Muesli bar | 138 | 199 | 6 | 5 |
| Apple | 216 | 343 | 2 | 1 |

Table 29: Number of times the snacks products were preferred over any other snack product and resulting matrix rankings of the products for both groups, including intransitivities.

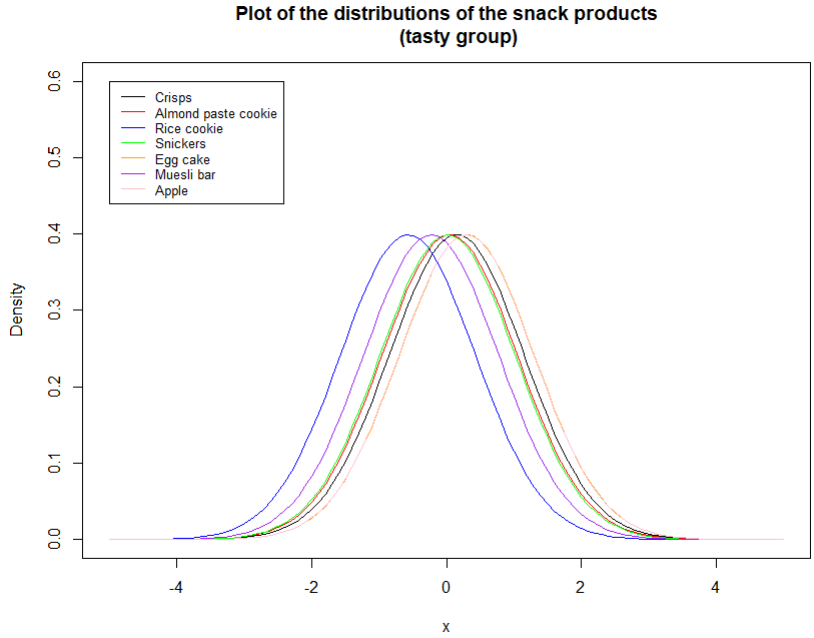



Figure 40: Distributions of the snack products for the tasty group, including intransitivities.

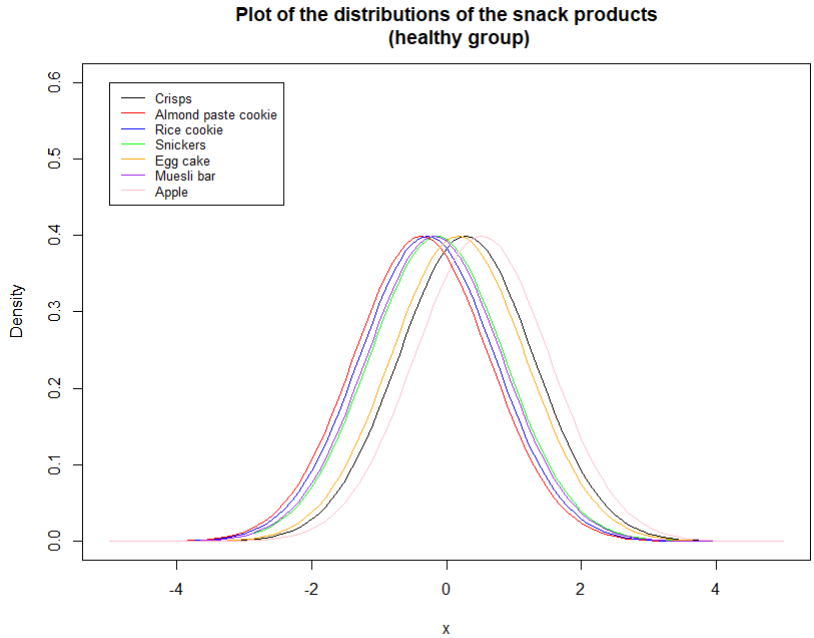



Figure 41: Distributions of the snack products for the healthy group, including intransitivities.

| | Crisps | Almond cookie | Rice cookie | Snickers | Sponge cake | Muesli bar | Apple |
|---|---|---|---|---|---|---|---|
| Crisps | - | 19 | 24 | 19 | 16 | 21 | 18 |
| Almond cookie | 16 | - | 23 | 17 | 10 | 21 | 10 |
| Rice cookie | 11 | 12 | - | 10 | 8 | 15 | 4 |
| Snickers | 16 | 18 | 25 | - | 10 | 20 | 14 |
| Sponge cake | 19 | 25 | 27 | 25 | - | 25 | 17 |
| Muesli bar | 14 | 14 | 20 | 15 | 10 | - | 9 |
| Apple | 17 | 25 | 31 | 21 | 18 | 26 | - |

Table 30: Frequency matrix of the tasty group, without intransitivities (row elements are preferred over the column elements).

| | Crisps | Almond cookie | Rice cookie | Snickers | Sponge cake | Muesli bar | Apple |
|---|---|---|---|---|---|---|---|
| Crisps | - | 42 | 40 | 40 | 33 | 40 | 25 |
| Almond cookie | 14 | - | 25 | 22 | 13 | 23 | 9 |
| Rice cookie | 16 | 31 | - | 29 | 18 | 31 | 9 |
| Snickers | 16 | 34 | 27 | - | 18 | 28 | 14 |
| Sponge cake | 23 | 43 | 38 | 38 | - | 36 | 17 |
| Muesli bar | 16 | 33 | 25 | 28 | 20 | - | 12 |
| Apple | 31 | 47 | 47 | 42 | 39 | 44 | - |

Table 31: Frequency matrix of the healthy group, without intransitivities (row elements are preferred over the column elements).

| Food product | Sum tasty group | Sum healthy group | Matrix rank tasty | Matrix rank healthy |
|---|---|---|---|---|
| Crisps | 117 | 220 | 3 | 2 |
| Almond cookie | 97 | 106 | 5 | 7 |
| Rice cookie | 60 | 134 | 7 | 5.5 |
| Snickers | 103 | 137 | 4 | 4 |
| Sponge cake | 138 | 195 | 1.5 | 3 |
| Muesli bar | 82 | 134 | 6 | 5.5 |
| Apple | 138 | 250 | 1.5 | 1 |

Table 32: Number of times the snacks products were preferred over any other snack product and resulting matrix rankings of the products for both groups, without intransitivities.

| | Means tasty | Means healthy | Rank. tasty | Rank. healthy |
|---|---|---|---|---|
| Crisps | 0.126 | 0.352 | 3 | 2 |
| Almond paste cookie | -0.088 | -0.433 | 5 | 7 |
| Rice cookie | -0.512 | -0.244 | 7 | 6 |
| Snickers | -0.021 | -0.211 | 4 | 4 |
| Sponge cake | 0.359 | 0.184 | 2 | 3 |
| Muesli bar | -0.246 | -0.233 | 6 | 5 |
| Apple | 0.382 | 0.585 | 1 | 1 |

Table 33: Means and rankings of the snacks for both groups.

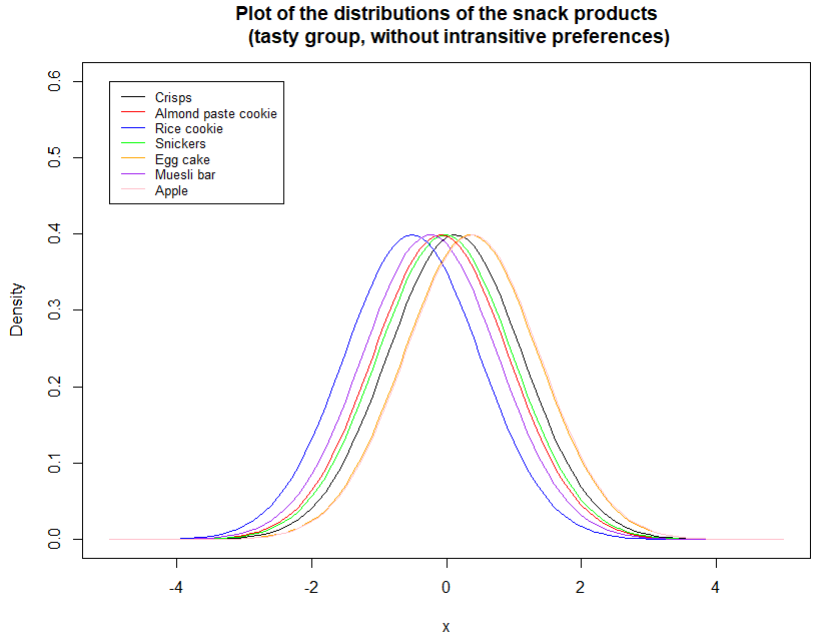



Figure 42: Distributions of the snack products for the tasty group, without intransitivities.

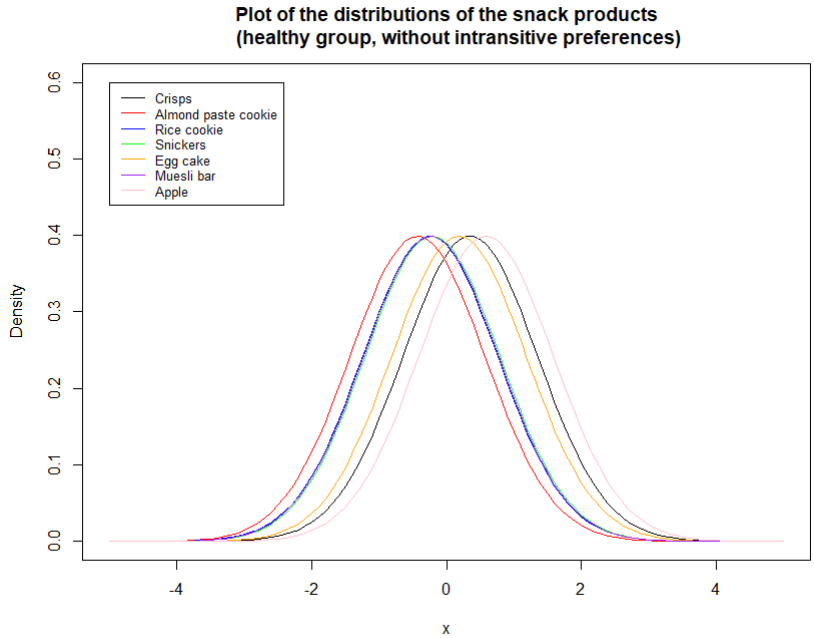



Figure 43: Distributions of the snack for the healthy group, without intransitivities.

# C R code

## C.1 Implementation of the least-squares method for the case when the distribution is known

```
library(nloptr)
library(MASS)
#Implementation of Case V, maybe not necessary as R has a standard Thurstone function
scaleest<-function(A){
  Invnorm<-qnorm(A)
  Dim <- dim(A)[1]
  scalevec<-numeric(Dim)
  for (i in 1:Dim){
    scalevec[i]<-(sum(Invnorm[,i])/Dim)*sqrt(2)
  }
  return(scalevec)
}

#Implementation of Case III and II using the true parameters as initial condition
#Case III
MinCase3go<-function(A,init){
  Normdev<-qnorm(A)
  Dim <- dim(A)[1]
  #Define objective function, sum of squares
  Obj<-function(x){
    f = c()
    for (i in (1:(Dim-1))){
      for (j in (i+1):Dim){
        fij <- (pnorm((x[i]-x[j])/sqrt(x[i+Dim]+x[j+Dim])) - A[i,j])^2
        f<-c(f,fij)
      }

    }
    return(sum(f))
  }
  #Define constraints, sum of scale values add up to 0, sum of variances are equal to n
  hinx<-function(x){
    h<-numeric(4)
    Dim<-0.5*length(x)
    y<-Dim+1
    z<-2*Dim
    h[1]<-sum(x[1:Dim])
    h[2]<- -1*sum(x[1:Dim])
    h[3]<-sum(x[y:z])-Dim
    h[4]<-Dim-sum(x[y:z])
    return(h)
  }
  #Minimize the function with constraints
  SolutionCase3.1<-cobyla(init,Obj,lower=c(rep(-100,Dim),rep(0,Dim)),
  upper=c(rep(100,2*Dim)),hin=hinx)
  return(SolutionCase3.1$par)
}

#Case II
MinCase2go<-function(A,init){
```

```
Normdev<-qnorm(A)
Dim <- dim(A)[1]
#Define objective function, sum of squares
Obj<-function(x){
  f = c()
  for (i in (1:(Dim-1))){
    for (j in (i+1):Dim){
      fij <- (pnorm((x[i]-x[j])/sqrt(x[i+Dim]+x[j+Dim]-
      2*x[2*Dim+1]*sqrt(x[i+Dim])*sqrt(x[j+Dim]))) - A[i,j])^2
      f<-c(f,fij)
    }

  }
  return(sum(f))
}
#Define constraints, sum of scale values add up to 0, sum of variances are equal to n
hinx<-function(x){
  h<-numeric(4)
  Dim<-0.5*(length(x)-1)
  y<-Dim+1
  z<-2*Dim
  constr<-numeric(4)
  h[1]<-sum(x[1:Dim])
  h[2]<- -1*sum(x[1:Dim])
  h[3]<-sum(x[y:z])-Dim
  h[4]<-Dim-sum(x[y:z])
  return(h)}
#Minimize the objective function
SolutionCase2.3<-cobyla(init,Obj,lower=c(rep(-Inf,Dim),rep(0,Dim),(-1/Dim),
upper=c(rep(Inf,2*Dim),1),hin=hinx)
return(SolutionCase2.3$par)
}
```

## C.2 Implementation of the least-squares method for the two step optimization for Case III and II

```
library(nloptr)
library(MASS)
MinCase3go2<-function(A){
  Normdev<-qnorm(A)
  Dim <- dim(A)[1]
  #Define objective function, sum of squares
  Obj<-function(x){
    f = c()
    for (i in (1:(Dim-1))){
      for (j in (i+1):Dim){
        fij <- (pnorm((x[i]-x[j])/sqrt(x[i+Dim]+x[j+Dim])) - A[i,j])^2
        f<-c(f,fij)
      }

    }
    return(sum(f))
  }
  #Define constraints, sum of scale values add up to 0, sum of variances are equal to n
  hinx<-function(x){
    h<-numeric(4)
    Dim<-0.5*length(x)
    y<-Dim+1
    z<-2*Dim
    h[1]<-sum(x[1:Dim])
    h[2]<- -1*sum(x[1:Dim])
    h[3]<-sum(x[y:z])-Dim
    h[4]<-Dim-sum(x[y:z])
    return(h)
  }
  #Use Quasi-Newton to determine the start parameters
  start<-nlminb(start=c(rep(0,Dim),rep(1,Dim)),objective=Obj,
                lower=c(rep(-10,Dim),rep(0,Dim)),upper=c(rep(10,2*Dim)))
  startpar<-start$par
  #Minimize the constrained objective function
  SolutionCase3.1<-cobyla(startpar, Obj, lower=c(rep(-10,Dim),rep(0,Dim)),
  upper=c(rep(10,2*Dim)),hin=hinx)
  return(SolutionCase3.1)
}


MinCase2go2<-function(A){
  Normdev<-qnorm(A)
  Dim <- dim(A)[1]
  #Define objective function, sum of squares
  Obj<-function(x){
    f = c()
    for (i in (1:(Dim-1))){
      for (j in (i+1):Dim){
        fij <- (pnorm((x[i]-x[j])/(sqrt(abs(x[i+Dim]+x[j+Dim]-
        2*x[2*Dim+1]*sqrt(x[i+Dim])*sqrt(x[j+Dim]))))) - A[i,j])^2
        f<-c(f,fij)
      }
```

```
    }
    return(sum(f))
  }
  #Define constraints, sum of scale values add up to 0, sum of variances are equal to n
  hinx<-function(x){
    h<-numeric(4)
    Dim<-0.5*(length(x)-1)
    y<-Dim+1
    z<-2*Dim
    constr<-numeric(4)
    h[1]<-sum(x[1:Dim])
    h[2]<- -1*sum(x[1:Dim])
    h[3]<-sum(x[y:z])-Dim
    h[4]<-Dim-sum(x[y:z])
    return(h)}
  #Use Quasi-Newton to determine the start parameters
  start<-nlminb(start=c(rep(0,Dim),rep(1,Dim),0),objective=Obj,
                lower=c(rep(-10,Dim),rep(0,Dim),(-1/Dim)),upper=c(rep(10,2*Dim),1))
  startpar<-start$par
  #Minimize the constrained objective function
  SolutionCase2.3<-cobyla(startpar,Obj,lower=c(rep(-10,Dim),rep(0,Dim),(-1/Dim)),
  upper=c(rep(10,2*Dim),1),hin=hinx)
  return(SolutionCase2.3)
}
```

## C.3 Simulation study Case V

```
#Simulation study
library(MASS)
#Case V, assuming equal variances and different means

#Simulates proportion matrix from multivariate normal distribution
simmatrixV<-function(mu,sigma,N){
  stimuli<-length(mu)
  #Define frequency matrix
  Q<-matrix(0,nrow=stimuli,ncol=stimuli)
  #Simulate N judgment vectors
  for (i in 1:N){
    sim<-mvrnorm(1,mu,sigma)
    #Define preference matrix for one judge
    prefmatrix<-matrix(0,nrow=stimuli,ncol=stimuli)
    #Assign 1-0 encoding to the matrix according to the outcome of the mvrnorm
    for (j in 1:(stimuli-1)){
      for (k in (j+1):stimuli){
        if (sim[j]>sim[k]){prefmatrix[j,k]=1}
        else {prefmatrix[k,j]=1}}
    }
    for (l in 1:stimuli){
      prefmatrix[l,l]=0.5}
    Q = Q + prefmatrix
  }
  #Get the proportion matrix by dividing the frequency matrix by the number of individuals
  propmatrix<-Q/N
  #Take care of 0's and 1's in the matrix by replacing them with a small number close to 0 or 1.
  #So that an estimation can be made
  for (i in 1:stimuli){
    for (j in 1:stimuli){
      if (propmatrix[i,j] ==0){propmatrix[i,j]=1/N}
      else if (propmatrix[i,j] ==1){propmatrix[i,j]=1-(1/N)}
    }
  }
  return(propmatrix)
}

#Defining different equidistant mean vectors (with differing distance)
mu1<-c(-0.5,-0.3,-0.1,0.1,0.3,0.5)
mu2<-c(-1.25,-0.75,-0.25,0.25,0.75,1.25)
mu3<-c(-2.5,-1.5,-0.5,0.5,1.5,2.5)
mu4<-c(-5,-3,-1,1,3,5)
mu5<-c(-10,-6,-2,2,6,10)
N<-1000
n<-25
#Defining different randomly spaced mean vectors
mu6<-c(-0.5,-0.3,-0.2,-0.1,0.3,0.8)
mu7<-c(-2.3,-1.5,-0.6,-0.1,1.8,2.7)
mu8<-c(-5.3,-3.5,-0.8,0.9,3.8,4.9)

#Assuming variance = 0.5 Look at the effect of different spacing of
the means (equidistantly or random)
\\
```

```
Simulation0CaseV<-function(mu,N,n){
  parameters<-c(rep(0,length(mu)))
  y<-rep(1,length(mu))
  sigma<-diag(y)
  for (i in 1:n){
    prop0<-simmatrixV(mu,sigma,N)
    parameters<-parameters+scaleest(t(prop0))}
  parameters<-parameters/n
  return(parameters)
}

#Simulate the parameters for the five different equidistant mean vectors
par1<-Simulation0CaseV(mu1,N,n)
par2<-Simulation0CaseV(mu2,N,n)
par3<-Simulation0CaseV(mu3,N,n)
par4<-Simulation0CaseV(mu4,N,n)
par5<-Simulation0CaseV(mu5,N,n)
parameterequi<-rbind(par1,par2,par3,par4,par5)


muequi<-rbind(mu1,mu2,mu3,mu4,mu5)
#Define a Goodness of fit function which measures the residual sum of squares (distance)
GoodnessFitMeasure5.0<-function(par,mu){
  diff<-abs(par-mu)
  ssq<-sum(diff^2)
  return(ssq)
}

#This can also be seen by transforming the parameters to a [0,1] interval
and plot them simultaneously
transformedparameters<-matrix(0,ncol=6,nrow=5)
for (i in 1:5){
  maxi<-max(muequi[i,])
  mini<-min(muequi[i,])
  for (j in 1:6){
    transformedparameters[i,j]<-(parameterequi[i,j]-mini)/(maxi-mini)
  }


  SSQ<-c()
  for (i in 1:5)
    SSQ[i]<-GoodnessFitMeasure5.0(transformedparameters[i,],sequence)
  Equivec<-c(0.2,0.5,1,2,4)
  plot(Equivec,SSQ,type="o",xlab="Distance between the means",
  ylab="Sum of squares of the differences of the true and estimated means",
       main="The influence of the distance between the means on the accuracy
       of the estimated means")

}
#Plot the transformed parameters
sequence<-seq(0,1,0.2)
minimum<-min(transformedparameters,sequence)
maximum<-max(transformedparameters,sequence)
plot(transformedparameters[1,],sequence,xlim<- c(minimum,maximum),
xlab="Estimated means after simulation",ylab="True means",
```

```
     main="Influence of the distance between the means on the
     accuracy of the estimated means",type="o")
lines(transformedparameters[2,],sequence,type="o",pch=0,col="blue")
lines(transformedparameters[3,],sequence,type="o",pch=2,col="red")
lines(transformedparameters[4,],sequence,type="o",pch=3,col="green")
lines(transformedparameters[5,],sequence,type="o",pch=4,col="yellow")
lines(sequence,sequence,type="o",pch=5,col="purple")
legend(minimum, maximum, legend=c("Distance = 0.2", "Distance = 0.5", "Distance = 1",
"Distance = 2","Distance = 4" ,"Ideal estimation"),
       col=c("black", "blue", "red", "green","yellow","purple"), lty=c(rep(1,6)),
       pch = c(1,0,2,3,4,5), cex=0.8)

samplesize<-c()
maxdist<-c()
for (i in 1:35){
  x<-10^{i}
  samplesize <- c(samplesize, x)
  dist<-(-10/6)*qnorm(1/x)
  maxdist<-c(maxdist,dist)
}
plot(samplesize,maxdist,xlab="Sample size",ylab="Maximum distance between the scale values
of the most and least preferred objects",type="o",main="Maximum distance between the
scale values of the most and least preferred stimuli in relation to the sample size")

#Randomly spaced means
par6<-Simulation0CaseV(mu6,1000,25) #small distance between the outer means
par7<-Simulation0CaseV(mu7,1000,25) #medium distance between the outer means
par8<-Simulation0CaseV(mu8,1000,25) #large distance between the outer means
```

## C.4 Simulation study of the two-step optimization process for Case III and II

```
#Case III simulation, investigating the performance of the two-step optimization
mu1<-c(-0.5,-0.3,-0.1,0.1,0.3,0.5)
mu3<-c(-2.5,-1.5,-0.5,0.5,1.5,2.5)
var1<-c(0.7,0.8,0.9,1.1,1.2,1.3)
var3<-c(0.01,0.07,0.1,1.5,1.4,2.92)
#Define function that produces a matrix of the estimates
Case3Sim2<-function(mu,var,N,n){
  parametermatrix<-matrix(0,nrow=n,ncol=2*length(mu))
  parametermatrix2<-matrix(0,nrow=n,ncol=2*length(mu))
  for (i in 1:n){
    simulation<-simmatrix(mu,var,0,N)[[1]]
    parameter<-MinCase3go2(simulation)$par
    parametermatrix[i,]<-parameter
    parameter2<-MinCase3go(simulation,c(mu,var))[,1]
    parametermatrix2[i,]<-parameter2
  }
  minmaxmatrix<-matrix(0,nrow=2*length(mu),ncol=2)
  minmaxmatrix2<-matrix(0,nrow=2*length(mu),ncol=2)
  for (i in 1:(2*length(mu))){
    minmaxmatrix[i,1]<-min(parametermatrix[,i])
    minmaxmatrix[i,2]<-max(parametermatrix[,i])
    minmaxmatrix2[i,1]<-min(parametermatrix2[,i])
    minmaxmatrix2[i,2]<-max(parametermatrix2[,i])
  }
  meanvec<-c()
  meanvec2<-c()
  for (i in 1:(2*length(mu))){
    meanvec[i]<-mean(parametermatrix[,i])
    meanvec2[i]<-mean(parametermatrix2[,i])
  }
  return(list(cbind(minmaxmatrix,meanvec,c(mu,var)),cbind(minmaxmatrix2,meanvec2,c(mu,var)),
              parametermatrix,parametermatrix2))
}

Sim2equismall1<-Case3Sim2(mu1,var1,1000,25)
Sim2equismall3<-Case3Sim2(mu1,var3,1000,25)

Sim2equilarge1<-Case3Sim2(mu3,var1,1000,25)
Sim2equilarge3<-Case3Sim2(mu3,var3,1000,25)

#Plots of the results
#Small variability in variance
#Means
par(mfrow=c(1,2))
plot(Sim2equismall1[[1]][1:6,3],mu1,pch=4,xlab="Estimates of the means determining the starting
parameters by optimization", ylab="True means", xlim=c(-0.7,0.6),
     ylim=c(-0.5,0.5), main="Small distance between the true means")
points(Sim2equismall1[[1]][1:6,1],mu1,pch=2,col="blue")
points(Sim2equismall1[[1]][1:6,2],mu1,pch=2,col="blue")
points(mu1,mu1,pch=1, col="red")
legend(-0.7, 0.5, legend=c("Means of the estimations", "Min./Max. estimations",
"True parameter" ),col=c("black", "blue", "red"), lty=c(rep(0,3)),
```

```
       pch = c(4,2,1), cex=0.8)
plot(Sim2equismall1[[2]][1:6,3],mu1, pch=4,xlab="Estimates of the means using the true
parameters as starting parameters", ylab="True means", xlim=c(-0.6,0.6),
     ylim=c(-0.5,0.5), main="and a small difference in variance")
points(Sim2equismall1[[2]][1:6,1],mu1,pch=2,col="blue")
points(Sim2equismall1[[2]][1:6,2],mu1,pch=2,col="blue")
points(mu1,mu1,pch=1, col="red")
legend(-0.6, 0.5, legend=c("Means of the estimations", "Min./Max. estimations",
"True parameter" ), col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)


plot(Sim2equilarge1[[1]][1:6,3],mu3,pch=4,xlab="Estimates of the means determining the starting
     parameters by optimization", ylab="True means", xlim=c(-4,3.5),ylim=c(-2.5,2.5)
     ,main = "Large distance between the true means")
points(mu3,mu3, pch=1,col="red")
points(Sim2equilarge1[[1]][1:6,1],mu3,pch=2,col="blue")
points(Sim2equilarge1[[1]][1:6,2],mu3,pch=2,col="blue")
legend(-4, 2.5, legend=c("Means of the estimations", "Min./Max. estimations","True parameter" ),
       col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)
plot(Sim2equilarge1[[2]][1:6,3],mu3, pch=4,xlab="Estimates of the means using the true
parameters as starting parameters", ylab="True means", xlim=c(-4,3.5),ylim=c(-2.5,2.5)
     ,main = "and a small difference in variance")
points(mu3,mu3, pch=1,col="red")
points(Sim2equilarge1[[2]][1:6,1],mu3,pch=2,col="blue")
points(Sim2equilarge1[[2]][1:6,2],mu3,pch=2,col="blue")
legend(-4, 2.5, legend=c("Means of the estimations", "Min./Max. estimations","True parameter" ),
       col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)

#Variances
plot(Sim2equismall1[[1]][7:12,3],var1, pch=4,xlab="Estimates of the variances determining
the starting parameters by optimization", ylab="True variances", xlim=c(0,3),
     ylim=c(0.5,1.5), main="Small distance between the true means")
points(var1,var1,pch=1, col="red")
points(Sim2equismall1[[1]][7:12,1],var1,pch=2,col="blue")
points(Sim2equismall1[[1]][7:12,2],var1,pch=2,col="blue")
legend(0, 1.5, legend=c("Means of the variance estimations", "Min./Max. estimations",
"True parameter" ), col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)
plot(Sim2equismall1[[2]][7:12,3],var1, pch=4,xlab="Estimates of the variances using the true
     parameters as starting parameters", ylab="True variances", xlim=c(0,3),
     ylim=c(0.5,1.5), main="and a small difference in variance")
points(var1,var1,pch=1, col="red")
points(Sim2equismall1[[2]][7:12,1],var1,pch=2,col="blue")
points(Sim2equismall1[[2]][7:12,2],var1,pch=2,col="blue")
legend(0, 1.5, legend=c("Means of the variance estimations", "Min./Max. estimations",
"True parameter" ), col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)


plot(Sim2equilarge1[[1]][7:12,3],var1, pch=4,xlab="Estimates of the variances determining the
starting parameters by optimization", ylab="True variances", xlim=c(0,3.5),ylim=c(0,2)
     ,main = "Large distance between the true means")
```

```
points(var1,var1, pch=1,col="red")
points(Sim2equilarge1[[1]][7:12,1],var1,pch=2,col="blue")
points(Sim2equilarge1[[1]][7:12,2],var1,pch=2,col="blue")
legend(0, 2, legend=c("Means of the variance estimations", "Min./Max. estimations",
"True parameter" ),col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)
plot(Sim2equilarge1[[2]][7:12,3],var1, pch=4,xlab="Estimates of the variances using the true
     parameters as starting parameters", ylab="True variances", xlim=c(0,3.5),ylim=c(0,2)
     ,main = "and a small difference in variance")
points(var1,var1, pch=1,col="red")
points(Sim2equilarge1[[2]][7:12,1],var1,pch=2,col="blue")
points(Sim2equilarge1[[2]][7:12,2],var1,pch=2,col="blue")
legend(0, 2, legend=c("Means of the variance estimations", "Min./Max. estimations",
"True parameter" ), col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)

#Large variability in variance
#Means
plot(Sim2equismall3[[1]][1:6,3],mu1, pch=4,xlab="Estimates of the means determining the starting
     parameters by optimization", ylab="True means", xlim=c(-1,0.6),
     ylim=c(-0.5,0.5), main="Small distance between the true means")
points(mu1,mu1,pch=1, col="red")
points(Sim2equismall3[[1]][1:6,1],mu1,pch=2,col="blue")
points(Sim2equismall3[[1]][1:6,2],mu1,pch=2,col="blue")
legend(-1, 0.5, legend=c("Means of the estimations", "Min./Max. estimations","True parameter" ),
       col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)
plot(Sim2equismall3[[2]][1:6,3],mu1, pch=4,xlab="Estimates of the means using the true
     parameters as starting parameters", ylab="True means", xlim=c(-1,0.6),
     ylim=c(-0.5,0.5), main="and a large difference in variance")
points(mu1,mu1,pch=1, col="red")
points(Sim2equismall3[[2]][1:6,1],mu1,pch=2,col="blue")
points(Sim2equismall3[[2]][1:6,2],mu1,pch=2,col="blue")
legend(-1, 0.5, legend=c("Means of the estimations", "Min./Max. estimations","True parameter" ),
       col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)

plot(Sim2equilarge3[[1]][1:6,3],mu3, pch=4,xlab="Estimates of the means determining the starting
     parameters by optimization", ylab="True means", xlim=c(-6,6),
     ylim=c(-5,5), main="Large distance between the true means")
points(mu3,mu3,pch=1, col="red")
points(Sim2equilarge3[[1]][1:6,1],mu3,pch=2,col="blue")
points(Sim2equilarge3[[1]][1:6,2],mu3,pch=2,col="blue")
legend(-6, 5, legend=c("Means of the estimations", "Min./Max. estimations","True parameter" ),
       col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)
plot(Sim2equilarge3[[2]][1:6,3],mu3, pch=4,xlab="Estimates of the means using the true
     parameters as starting parameters", ylab="True means", xlim=c(-6,6),
     ylim=c(-5,5), main="and a large difference in variance")
points(mu3,mu3,pch=1, col="red")
points(Sim2equilarge3[[2]][1:6,1],mu3,pch=2,col="blue")
points(Sim2equilarge3[[2]][1:6,2],mu3,pch=2,col="blue")
legend(-6, 5, legend=c("Means of the estimations", "Min./Max. estimations","True parameter" ),
       col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)
```

```
#Variances
plot(Sim2equismall3[[1]][7:12,3],var3, pch=4,xlab="Estimates of the variances determining the
starting parameters by optimization", ylab="True variances", xlim=c(0,3.5),
     ylim=c(0,5), main="Small distance between the true means")
points(var3,var3,pch=1, col="red")
points(Sim2equismall3[[1]][7:12,1],var3,pch=2,col="blue")
points(Sim2equismall3[[1]][7:12,2],var3,pch=2,col="blue")
legend(0, 5, legend=c("Means of the variance estimations", "Min./Max. estimations",
"True parameter" ), col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)
plot(Sim2equismall3[[2]][7:12,3],var3, pch=4,xlab="Estimates of the variances using the true
     parameters as starting parameters", ylab="True variances", xlim=c(0,3.5),
     ylim=c(0,5), main="and a large difference in variance")
points(var3,var3,pch=1, col="red")
points(Sim2equismall3[[2]][7:12,1],var3,pch=2,col="blue")
points(Sim2equismall3[[2]][7:12,2],var3,pch=2,col="blue")
legend(0, 5, legend=c("Means of the variance estimations", "Min./Max. estimations",
"True parameter" ), col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)

plot(Sim2equilarge3[[1]][7:12,3],var3,pch=4,xlab="Estimates of the variances determining
the starting parameters by optimization", ylab="True variances", xlim=c(0,4.5),ylim=c(0,5)
     ,main = "Large distance between the true means")
points(var3,var3, pch=1,col="red")
points(Sim2equilarge3[[1]][7:12,1],var3,pch=2,col="blue")
points(Sim2equilarge3[[1]][7:12,2],var3,pch=2,col="blue")
legend(0,5, legend=c("Means of the variance estimations", "Min./Max. estimations",
"True parameter" ), col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)
plot(Sim2equilarge3[[2]][7:12,3],var3, pch=4,xlab="Estimates of the variances using the true
     parameters as starting parameters", ylab="True variances", xlim=c(0,4.5),ylim=c(0,5)
     ,main = "and a large difference in variance")
points(var3,var3, pch=1,col="red")
points(Sim2equilarge3[[2]][7:12,1],var3,pch=2,col="blue")
points(Sim2equilarge3[[2]][7:12,2],var3,pch=2,col="blue")
legend(0,5, legend=c("Means of the variance estimations", "Min./Max. estimations",
"True parameter" ), col=c("black", "blue", "red"), lty=c(rep(0,3)),
       pch = c(4,2,1), cex=0.8)

#Case II
Case2Sim<-function(mu,var,rho,N,n){
  parametermatrix<-matrix(0,nrow=n,ncol=2*length(mu)+1)
  parametermatrix2<-matrix(0,nrow=n,ncol=2*length(mu)+1)
  for (i in 1:n){
    simulation<-simmatrix(mu,var,rho,N)[[1]]
    parameter<-MinCase2go2(simulation)$par
    parametermatrix[i,]<-parameter
    parameter2<-MinCase2go(simulation,c(mu,var,rho))[,1]
    parametermatrix2[i,]<-parameter2
  }
  minmaxmatrix<-matrix(0,nrow=2*length(mu)+1,ncol=2)
  minmaxmatrix2<-matrix(0,nrow=2*length(mu)+1,ncol=2)
  for (i in 1:(2*length(mu)+1)){
```

```
    minmaxmatrix[i,1]<-min(parametermatrix[,i])
    minmaxmatrix[i,2]<-max(parametermatrix[,i])
    minmaxmatrix2[i,1]<-min(parametermatrix2[,i])
    minmaxmatrix2[i,2]<-max(parametermatrix2[,i])
  }
  meanvec<-c()
  meanvec2<-c()
  for (i in 1:(2*length(mu)+1)){
    meanvec[i]<-mean(parametermatrix[,i])
    meanvec2[i]<-mean(parametermatrix2[,i])
  }
  return(list(cbind(minmaxmatrix,meanvec,c(mu,var,rho)),cbind(minmaxmatrix2,meanvec2,
  c(mu,var,rho))))
}

Case2Sim(mu1,var1,-0.1,100,25)
Case2Sim(mu1,var1,0.25,100,25)
Case2Sim(mu1,var1,0.5,100,25)
Case2Sim(mu1,var1,0.9,100,25)
```

## C.5 AIC implementation for the case when the distribution is known

```
#AIC study using the true parameters as initial condition
#Implementing the algorithm
library(nloptr)
library(MASS)
library(lavaan)
#Simulate matrices from multivar norm distribution including correlation
simmatrix<-function(mu,var,rho,N){
  Dim<-length(mu)
  sigma<-matrix(rho,ncol=Dim,nrow=Dim)
  diag(sigma)<-var
  covmatrix<-cor2cov(sigma,sqrt(var))
  diag(covmatrix)<-var
  stimuli<-length(mu)
  #Define frequency matrix
  Q<-matrix(0,nrow=stimuli,ncol=stimuli)
  #Define number of circ. triads vector
  numcirc<-numeric(N)
  #Simulate N judgment vectors
  for (i in 1:N){
    sim<-mvrnorm(1,mu,covmatrix)
    #Define preference matrix for one judge
    prefmatrix<-matrix(0,nrow=stimuli,ncol=stimuli)
    #Assign 1-0 encoding to the matrix according to the outcome of the mvrnorm
    for (j in 1:(stimuli-1)){
      for (k in (j+1):stimuli){
        if (sim[j]>sim[k]){prefmatrix[j,k]=1}
        else {prefmatrix[k,j]=1}}
    }
    numcirc[i]<-numcirctriads(prefmatrix)
    for (l in 1:stimuli){
      prefmatrix[l,l]=0.5}
    Q = Q + prefmatrix
  }
  #Get the proportion matrix by dividing the frequency matrix by the number of individuals
  propmatrix<-Q/N
  #Take care of 0's and 1's in the matrix by replacing them with a small number close to 0 or 1.
  #So that an estimation can be made
  for (i in 1:stimuli){
    for (j in 1:stimuli){
      if (propmatrix[i,j] ==0){propmatrix[i,j]=1/N}
      else if (propmatrix[i,j] ==1){propmatrix[i,j]=1-(1/N)}
    }
  }
  return(list(propmatrix,numcirc))
}
#Return the ranks of the matrix
Rankmatrix<-function(A){
  Dim<-dim(A)[1]
  prerank<-numeric(Dim)
  for (i in 1:Dim){
    prerank[i]<-sum(A[i,])
  }
  x<-rank(prerank)
```

```
  return(x)
}

#Return the MSE of the ranks
Devrank<-function(rank,Dim){
  truerank<-seq(1,Dim,1)
  deviation<-(1/Dim)*(sum((rank-truerank)^2))
  return(deviation)
}

#Determine the best model in the case when the true parameters are used as initial condition
AICmodel2<-function(mu,var,rho,N){
  Dim<-length(mu)
  simulatie<-simmatrix(mu,var,rho,N)
  propmatrix<-simulatie[[1]]
  numcirc<-simulatie[[2]]
  totmatrix<-N*propmatrix
  coeffagree<-Coeffagree(totmatrix,N,Dim)
  rankmatrix<-Rankmatrix(totmatrix)
  parcase5<-scaleest(t(propmatrix))
  parcase3<-MinCase3go(propmatrix,c(mu,var))[,1]
  parcase2<-MinCase2go(propmatrix,c(mu,var,rho))[,1]
  Case5matrix<-matrix(0,ncol=Dim,nrow=Dim)
  diag(Case5matrix)<-c(rep(0.5,Dim))
  for (i in 1:Dim-1){
    for (j in (i+1):Dim){
      Case5matrix[i,j] = pnorm((1/sqrt(2))*parcase5[i]-parcase5[j])
      Case5matrix[j,i] = 1-Case5matrix[i,j]
    }
  }
  Case3matrix<-matrix(0,ncol=Dim,nrow=Dim)
  diag(Case3matrix)<-c(rep(0.5,Dim))
  for (i in 1:Dim-1){
    for (j in (i+1):Dim){
      Case3matrix[i,j] = pnorm((parcase3[i]-parcase3[j])/(sqrt(parcase3[i+Dim]+parcase3[j+Dim])))
      Case3matrix[j,i] = 1-Case3matrix[i,j]
    }
  }
  Case2matrix<-matrix(0,ncol=Dim,nrow=Dim)
  diag(Case2matrix)<-c(rep(0.5,Dim))
  for (i in 1:Dim-1){
    for (j in (i+1):Dim){
      Case2matrix[i,j] = pnorm((parcase2[i]-parcase2[j])/(sqrt(parcase2[i+Dim]+parcase2[j+Dim]
    -2*sqrt(parcase2[i+Dim])*sqrt(parcase2[j+Dim])*parcase2[(2*Dim)+1])))
      Case2matrix[j,i] = 1-Case2matrix[i,j]}}
  pij5<-c()
  pij3<-c()
  pij2<-c()
  for (i in 1:Dim-1){
    for (j in (i+1):Dim){
      p5<-(N*propmatrix[i,j]*log(Case5matrix[i,j]))+
      ((N-N*propmatrix[i,j])*log(1-Case5matrix[i,j]))
      p3<-(N*propmatrix[i,j]*log(Case3matrix[i,j]))+
      ((N-N*propmatrix[i,j])*log(1-Case3matrix[i,j]))
      p2<-(N*propmatrix[i,j]*log(Case2matrix[i,j]))+
```

```
      ((N-N*propmatrix[i,j])*log(1-Case2matrix[i,j]))
      pij5<-c(pij5,p5)
      pij3<-c(pij3,p3)
      pij2<-c(pij2,p2)
    }
  }
  L5<-sum(pij5)
  L3<-sum(pij3)
  L2<-sum(pij2)
  AIC5<-2*Dim-2*L5
  AIC3<-4*Dim-2*L3
  AIC2<-2*(2*Dim+1)-2*L2
  outcome<-c(AIC5,AIC3,AIC2)
  minimum<-which.min(outcome)
  parametersAIC<-c()
  if (minimum ==1){
    parametersAIC<-parcase5   }
  if(minimum==2){
    parametersAIC<-parcase3 }
  if (minimum==3){
    parametersAIC<-parcase2}
  return(list(minimum,rank(parametersAIC[1:Dim]),rankmatrix,numcirc,
  coeffagree,outcome,propmatrix,parametersAIC))}

#Repeat the procedure described above n times
totalfunction<-function(mu,var,rho,N,n){
  Dim<-length(mu)
  coeffagree<-numeric(n)
  devrankmatrix<-numeric(n)
  devrankpar<-numeric(n)
  bestmodel<-numeric(n)
  totnumcirc<-numeric(n)
  AICvalues<-matrix(0,ncol=3,nrow=n)
  for (i in 1:n){
    outcome<-AICmodel2(mu,var,rho,N)
    bestmodel[i]<-outcome[[1]]
    devrankmatrix[i]<-Devrank(outcome[[3]],Dim)
    devrankpar[i]<-Devrank(outcome[[2]],Dim)
    coeffagree[i]<-outcome[[5]]
    totnumcirc[i]<-sum(outcome[[4]])
    AICvalues[i,]<-outcome[[6]]
  }
  return(list(outcome,bestmodel,devrankmatrix,devrankpar,coeffagree,totnumcirc,AICvalues))
}

#Computes how many times the AIC proposed cases V, III and II
howmany<-function(vec){
  n<-length(vec)
  howmany1<-0
  howmany2<-0
  howmany3<-0
  for(i in 1:n){
    if(vec[i]==1){howmany1=howmany1+1}
    else if(vec[i]==2){howmany2=howmany2+1}
    else{howmany3=howmany3+1}
```

```
  }
  return(list(howmany1,howmany2,howmany3))
}

#Computes how many times ranking of the AIC was right in the three different cases
howmanyrightpar<-function(uitkomst,model){
  total0<-0
  indexvec<-which(uitkomst[[2]]==model)
  for (i in 1:length(indexvec)){
    if(uitkomst[[4]][indexvec[i]]==0){total0<-total0+1}
  }
  return(total0)
}
#Computes how many times the rankings of the matrix coincided with the real rankings
howmanyrightmat<-function(uitkomst,model){
  total0<-0
  indexvec<-which(uitkomst[[2]]==model)
  for (i in 1:length(indexvec)){
    if(uitkomst[[3]][indexvec[i]]==0){total0<-total0+1}
  }
  return(total0)
}
```

## C.6 AIC implementation for the two-step optimization process

```
#Determine the best model using a two-step optimization
AICmodel3<-function(mu,var,rho,N){
  Dim<-length(mu)
  simulatie<-simmatrix(mu,var,rho,N)
  propmatrix<-simulatie[[1]]
  numcirc<-simulatie[[2]]
  totmatrix<-N*propmatrix
  coeffagree<-Coeffagree(totmatrix,N,Dim)
  rankmatrix<-Rankmatrix(totmatrix)
  parcase5<-scaleest(t(propmatrix))
  parcase3<-MinCase3go2(propmatrix)$par
  Case5matrix<-matrix(0,ncol=Dim,nrow=Dim)
  diag(Case5matrix)<-c(rep(0.5,Dim))
  for (i in 1:Dim-1){
    for (j in (i+1):Dim){
      Case5matrix[i,j] = pnorm((1/sqrt(2))*parcase5[i]-parcase5[j])
      Case5matrix[j,i] = 1-Case5matrix[i,j]
    }
  }
  Case3matrix<-matrix(0,ncol=Dim,nrow=Dim)
  diag(Case3matrix)<-c(rep(0.5,Dim))
  for (i in 1:Dim-1){
    for (j in (i+1):Dim){
      Case3matrix[i,j] = pnorm((parcase3[i]-parcase3[j])/(sqrt(parcase3[i+Dim]+parcase3[j+Dim])))
      Case3matrix[j,i] = 1-Case3matrix[i,j]
    }
  }
  pij5<-c()
  pij3<-c()
  for (i in 1:Dim-1){
    for (j in (i+1):Dim){
      p5<-(N*propmatrix[i,j]*log(Case5matrix[i,j]))+
      ((N-N*propmatrix[i,j])*log(1-Case5matrix[i,j]))
      p3<-(N*propmatrix[i,j]*log(Case3matrix[i,j]))+
      ((N-N*propmatrix[i,j])*log(1-Case3matrix[i,j]))
      pij5<-c(pij5,p5)
      pij3<-c(pij3,p3)
    }
  }
  L5<-sum(pij5)
  L3<-sum(pij3)
  AIC5<-2*Dim-2*L5
  AIC3<-4*Dim-2*L3
  outcome<-c(AIC5,AIC3)
  minimum<-which.min(outcome)
  parametersAIC<-c()
  parametersfout<-c()
  if (minimum ==1){
    parametersAIC<-parcase5
    parametersfout<-parcase3}
  if(minimum==2){
    parametersAIC<-parcase3
    parametersfout<-parcase5}
```

```
  return(list(minimum,rank(parametersAIC[1:Dim]),rankmatrix,numcirc,coeffagree,
  outcome,propmatrix,parametersAIC,rank(parametersfout[1:Dim])))}

#Repeat the procedure described above n times
totalfunction2<-function(mu,var,rho,N,n){
  Dim<-length(mu)
  coeffagree<-numeric(n)
  devrankmatrix<-numeric(n)
  devrankpar<-numeric(n)
  devrankfout<-numeric(n)
  bestmodel<-numeric(n)
  totnumcirc<-numeric(n)
  AICvalues<-matrix(0,ncol=2,nrow=n)
  for (i in 1:n){
    outcome<-AICmodel3(mu,var,rho,N)
    bestmodel[i]<-outcome[[1]]
    devrankmatrix[i]<-Devrank(outcome[[3]],Dim)
    devrankpar[i]<-Devrank(outcome[[2]],Dim)
    devrankfout[i]<-Devrank(outcome[[9]],Dim)
    coeffagree[i]<-outcome[[5]][[2]]
    totnumcirc[i]<-sum(outcome[[4]])
    AICvalues[i,]<-outcome[[6]]}
  return(list(outcome,bestmodel,devrankmatrix,devrankpar,coeffagree,
  totnumcirc,AICvalues,devrankfout))}
```

## C.7 Simulation study AIC for the two-step optimization process

```
#Six objects
mu6<-c(-1.1, -0.6, -0.1, 0.3, 0.6, 0.9)
mu7<-c(-0.9, -0.5, -0.1, 0.2, 0.5, 0.8)
vareq6<-c(1,1,1,1,1,1)
varuneq6.1<-c(0.5,1.5,0.8,1.2,1.1,0.9)

#Seven objects
muzeven<-c(-1.0,-0.8,-0.3,0.1,0.3,0.7,1)
varuneq7<-c(0.8,1.3,1.9,0.1,1.2,0.7,1)
vareq7<-c(rep(1,7))

#Ten objects
mutien<-c(-1,-0.9,-0.75,-0.45,-0.1,0.3,0.45,0.6,0.85,1)
varuneq10<-c(0.4,1.6,0.35,0.55,1,1.65,1.45,1,0.1,1.9)
vareq10<-c(rep(1,10))

#Fifteen objects
muvijftien<-c(-1.0,-0.85,-0.65,-0.4,-0.3,-0.15,-0.05,0.05,0.1,0.25,0.3,0.4,0.55,0.75,1)
varuneq15<-c(0.2,1.8,1,0.05,1.95,0.6,1.4,0.35,1.65,0.9,1.1,1.2,0.8,0.3,1.7)
vareq15<-c(rep(1,15))

#Six objects
sixobjmu1varuneqsamp10<-totalfunction2(mu6,varuneq6.1,0,10,200)
sixobjmu1varuneqsamp30<-totalfunction2(mu6,varuneq6.1,0,30,200)
sixobjmu1varuneqsamp50<-totalfunction2(mu6,varuneq6.1,0,50,200)
sixobjmu1varuneqsamp100<-totalfunction2(mu6,varuneq6.1,0,100,200)
sixobjmu1varuneqsamp200<-totalfunction2(mu6,varuneq6.1,0,200,200)

sixobjmu1vareqsamp10<-totalfunction2(mu6,vareq6,0,10,200)
sixobjmu1vareqsamp50<-totalfunction2(mu6,vareq6,0,50,200)
sixobjmu1vareqsamp100<-totalfunction2(mu6,vareq6,0,100,200)
sixobjmu1vareqsamp200<-totalfunction2(mu6,vareq6,0,200,200)

sixobjmu2varuneqsamp10<-totalfunction2(mu7,varuneq6.1,0,10,200)
sixobjmu2varuneqsamp50<-totalfunction2(mu7,varuneq6.1,0,50,200)
sixobjmu2varuneqsamp100<-totalfunction2(mu7,varuneq6.1,0,100,200)
sixobjmu2varuneqsamp200<-totalfunction2(mu7,varuneq6.1,0,200,200)

sixobjmu2vareqsamp10<-totalfunction2(mu7,vareq6,0,10,200)
sixobjmu2vareqsamp50<-totalfunction2(mu7,vareq6,0,50,200)
sixobjmu2vareqsamp100<-totalfunction2(mu7,vareq6,0,100,200)
sixobjmu2vareqsamp200<-totalfunction2(mu7,vareq6,0,200,200)

#Six objects large mean distance unequal variances
sixobjlargemusamp10<-totalfunction2(mu3,varuneq6.1,0,10,200)
sixobjlargemusamp50<-totalfunction2(mu3,varuneq6.1,0,50,200)
sixobjlargemusamp100<-totalfunction2(mu3,varuneq6.1,0,100,200)
sixobjlargemusamp200<-totalfunction2(mu3,varuneq6.1,0,200,200)

#Seven objects
sevenobjmu1varuneqsamp10<-totalfunction2(muzeven,varuneq7,0,10,200)
sevenobjmu1varuneqsamp50<-totalfunction2(muzeven,varuneq7,0,50,200)
sevenobjmu1varuneqsamp100<-totalfunction2(muzeven,varuneq7,0,100,200)
```

```
sevenobjmu1varuneqsamp200<-totalfunction2(muzeven,varuneq7,0,200,200)

sevenobjmu1vareqsamp10<-totalfunction2(muzeven,vareq7,0,10,200)
sevenobjmu1vareqsamp50<-totalfunction2(muzeven,vareq7,0,50,200)
sevenobjmu1vareqsamp100<-totalfunction2(muzeven,vareq7,0,100,200)
sevenobjmu1vareqsamp200<-totalfunction2(muzeven,vareq7,0,200,200)

#Ten objects
tenobjmu1varuneqsamp10<-totalfunction2(mutien,varuneq10,0,10,200)
tenobjmu1varuneqsamp50<-totalfunction2(mutien,varuneq10,0,50,200)
tenobjmu1varuneqsamp100<-totalfunction2(mutien,varuneq10,0,100,200)
tenobjmu1varuneqsamp200<-totalfunction2(mutien,varuneq10,0,200,200)
tenobjmu1varuneqsamp500<-totalfunction2(mutien,varuneq10,0,500,200)
tenobjmu1varuneqsamp1000<-totalfunction2(mutien,varuneq10,0,1000,200)

tenobjmu1vareqsamp10<-totalfunction2(mutien,vareq10,0,10,200)
tenobjmu1vareqsamp50<-totalfunction2(mutien,vareq10,0,50,200)
tenobjmu1vareqsamp100<-totalfunction2(mutien,vareq10,0,100,200)
tenobjmu1vareqsamp200<-totalfunction2(mutien,vareq10,0,200,200)
tenobjmu1vareqsamp500<-totalfunction2(mutien,vareq10,0,500,200)
tenobjmu1vareqsamp1000<-totalfunction2(mutien,vareq10,0,1000,200)

#Fifteen objects
fifteenobjmu1varuneqsamp10<-totalfunction2(muvijftien,varuneq15,0,10,200)
fifteenobjmu1varuneqsamp50<-totalfunction2(muvijftien,varuneq15,0,50,200)
fifteenobjmu1varuneqsamp100<-totalfunction2(muvijftien,varuneq15,0,100,200)
fifteenobjmu1varuneqsamp200<-totalfunction2(muvijftien,varuneq15,0,200,200)

fifteenobjmu1vareqsamp10<-totalfunction2(muvijftien,vareq15,0,10,200)
fifteenobjmu1vareqsamp50<-totalfunction2(muvijftien,vareq15,0,50,200)
fifteenobjmu1vareqsamp100<-totalfunction2(muvijftien,vareq15,0,100,200)
fifteenobjmu1vareqsamp200<-totalfunction2(muvijftien,vareq15,0,200,200)

#Plot relevant figures
#Sample size 10 MSE
hist(sixobjmu1varuneqsamp10[[4]],main="MSE's of the parameter rankings",
xlab="Six objects, first mean vector, unequal variances, sample size equal to 10")
hist(sixobjmu1varuneqsamp10[[3]],main="MSE's of the matrix rankings",
xlab="Six objects, first mean vector, unequal variances, sample size equal to 10")
#KS test
ks.test(sixobjmu1varuneqsamp10[[4]],sixobjmu1varuneqsamp10[[3]])

#Sample size 50
par(mfrow=c(2,2))
indexVsamp50<-which(sixobjmu1varuneqsamp50[[2]]==1)
indexIIIsamp50<-which(sixobjmu1varuneqsamp50[[2]]==2)
hist(sixobjmu1varuneqsamp50[[4]][indexVsamp50], main="MSE's parameter rankings AIC",
xlab="Six objects, first mean vector, unequal variances, sample size equal to 50, AIC: Case V")
hist(sixobjmu1varuneqsamp50[[8]][indexVsamp50], main="MSE's parameter rankings other model",
xlab = "Six objects, first mean vector, unequal variances, sample size equal to 50, AIC: Case V")
hist(sixobjmu1varuneqsamp50[[4]][indexIIIsamp50], main="MSE's parameter rankings AIC",
xlab="Six objects, first mean vector, unequal variances, sample size equal to 50, AIC: Case III")
hist(sixobjmu1varuneqsamp50[[8]][indexIIIsamp50], main="MSE's parameter rankings other model",
xlab="Six objects, first mean vector, unequal variances, sample size equal to 50, AIC: Case III")
```

```
#Sample size 100
par(mfrow=c(1,2))
hist(sixobjmu1varuneqsamp100[[4]],main="MSE's of the parameter rankings",
xlab="Six objects, first mean vector, unequal variances, sample size equal to 100")
hist(sixobjmu1varuneqsamp200[[4]],main="MSE's of the parameter rankings",
xlab="Six objects, first mean vector, unequal variances, sample size equal to 200")

#Ks.test
ks.test(sixobjmu1varuneqsamp50[[4]],sixobjmu1varuneqsamp50[[3]])

a1<-length(which(sixobjmu1varuneqsamp10[[4]]-sixobjmu1varuneqsamp10[[3]]==0))
a2<-length(which(sixobjmu1varuneqsamp50[[4]]-sixobjmu1varuneqsamp50[[3]]==0))
a3<-length(which(sixobjmu1varuneqsamp100[[4]]-sixobjmu1varuneqsamp100[[3]]==0))
a4<-length(which(sixobjmu1varuneqsamp200[[4]]-sixobjmu1varuneqsamp200[[3]]==0))

plot(c(10,50,100,200),c(a1,a2,a3,a4),type="l",xlim=c(0,200),xlab="Sample size",
     ylab="Number of times the rankings of parameter and matrix coincided",
     main="Plot showing the relationship between the sample size
     and number of times the rankings of parameters and matrices coincided")

par(mfrow=c(1,3))
hist(sixobjmu1vareqsamp50[[4]],main="MSE's of the parameter rankings",
xlab="Six objects, first mean vector, equal variances, sample size equal to 50")
hist(sixobjmu1vareqsamp100[[4]],main="MSE's of the parameter rankings",
xlab="Six objects, first mean vector, equal variances, sample size equal to 100")
hist(sixobjmu1vareqsamp200[[4]],main="MSE's of the parameter rankings",
xlab="Six objects, first mean vector, equal variances, sample size equal to 200")

#Ten objects
par(mfrow=c(2,2))
hist(tenobjmu1varuneqsamp10[[4]],main="MSE's of the parameter rankings",xlab="Ten objects,
unequal variances, sample size equal to 10")
hist(tenobjmu1varuneqsamp50[[4]],main="MSE's of the parameter rankings",xlab="Ten objects,
     unequal variances, sample size equal to 50")
hist(tenobjmu1varuneqsamp100[[4]],main="MSE's of the parameter rankings",xlab="Ten objects,
     unequal variances, sample size equal to 100")
hist(tenobjmu1varuneqsamp200[[4]],main="MSE's of the parameter rankings",xlab="Ten objects,
unequal variances, sample size equal to 200")

#Larger sample size
par(mfrow=c(1,2))
hist(tenobjmu1varuneqsamp500[[4]],main="MSE's of the parameter rankings",xlab="Ten objects,
     unequal variances, sample size equal to 500")
hist(tenobjmu1varuneqsamp1000[[4]],main="MSE's of the parameter rankings",xlab="Ten objects,
     unequal variances, sample size equal to 1000")

#Coefficient of agreement
coeffagree2<-data.frame(cbind(sixobjmu1varuneqsamp10[[5]],
sixobjmu1varuneqsamp50[[5]],
sixobjmu1varuneqsamp100[[5]],sixobjmu1varuneqsamp200[[5]]))
boxplot(coeffagree2,names=c("Sample 10", "Sample 50", "Sample 100", "Sample 200"),
main="Distribution of the coefficient of agreement for six objects having unequal
        variances and different sample sizes")

#Average coefficient of agreement
```

```
par(mfrow=c(1,2))
mean(sixobjmu1varuneqsamp50[[5]][which(sixobjmu1varuneqsamp50[[2]]==2)])
mean(sixobjmu1varuneqsamp50[[5]][which(sixobjmu1varuneqsamp50[[2]]==1)])
par(mfrow=c(1,2))
boxplot(sixobjmu1varuneqsamp50[[5]][which(sixobjmu1varuneqsamp50[[2]]==1)],
xlab="Six objects, first mean vector,
unequal variances and a sample size equal to 50",
main="Case V recommended by the AIC",ylim=c(0.15,0.4))
boxplot(sixobjmu1varuneqsamp50[[5]][which(sixobjmu1varuneqsamp50[[2]]==2)],
xlab="Six objects, first mean vector,
unequal variances and a sample size equal to 50",
main="Case III recommended by the AIC",ylim=c(0.15,0.4))
```

## C.8 Implementation of the proposed alternative method for testing between-group concordance in the case of intransitivities

```
library(Matching)
library(RVAideMemoire)

simmatrix<-function(mu,var,rho,N){
  Dim<-length(mu)
  sigma<-matrix(rho,ncol=Dim,nrow=Dim)
  diag(sigma)<-var
  covmatrix<-cor2cov(sigma,sqrt(var))
  diag(covmatrix)<-var
  stimuli<-length(mu)
  #Define frequency matrix
  Q<-matrix(0,nrow=stimuli,ncol=stimuli)
  #Define number of circ. triads vector
  numcirc<-numeric(N)
  prefmatrixlist<-list()
  #Simulate N judgment vectors
  for (i in 1:N){
    sim<-mvrnorm(1,mu,covmatrix)
    #Define preference matrix for one judge
    prefmatrix<-matrix(0,nrow=stimuli,ncol=stimuli)
    #Assign 1-0 encoding to the matrix according to the outcome of the mvrnorm
    for (j in 1:(stimuli-1)){
      for (k in (j+1):stimuli){
        if (sim[j]>sim[k]){prefmatrix[j,k]=1}
        else {prefmatrix[k,j]=1}}
    }
    numcirc[i]<-numcirctriads(prefmatrix)
    prefmatrixlist[[i]]<-prefmatrix
    for (l in 1:stimuli){
      prefmatrix[l,l]=0.5}
    Q = Q + prefmatrix
  }
  #Get the proportion matrix by dividing the frequency matrix by the number of individuals
  propmatrix<-Q/N
  #Take care of 0's and 1's in the matrix by replacing them with a small number close to 0 or 1.
  #So that an estimation can be made
  for (i in 1:stimuli){
    for (j in 1:stimuli){
      if (propmatrix[i,j] ==0){propmatrix[i,j]=1/N}
      else if (propmatrix[i,j] ==1){propmatrix[i,j]=1-(1/N)}
    }
  }
  return(list(propmatrix,numcirc,prefmatrixlist))
}

comparemsematrix<-function(A1,A2,N1,n){
  Dim<-dim(A1)[1]
  Bestmodel<-BestmodelmatrixAIC(A1,N1)
  parametersAIC<-Bestmodel[[1]]
  lengte<-length(parametersAIC)
  MSEA1<-numeric(n)
  MSEA2<-numeric(n)
```

```
  if (lengte == Dim){
    mu<-parametersAIC
    for (i in 1:n){
      simulatiematrix<-simmatrix(mu,c(rep(1,Dim)),0,N1)[[1]]
      verschilA1<-(simulatiematrix-A1)^2
      verschilA2<-(simulatiematrix-A2)^2
      MSEA1[i]<-(1/Dim^2)*sum(verschilA1)
      MSEA2[i]<-(1/Dim^2)*sum(verschilA2)
    }
  }
  else if (lengte==2*Dim){
    mu<-parametersAIC[1:Dim]
    var<-parametersAIC[(Dim+1):(2*Dim)]
    for (i in 1:n){
      simulatiematrix<-simmatrix(mu,var,0,N1)[[1]]
      verschilA1<-(simulatiematrix-A1)^2
      verschilA2<-(simulatiematrix-A2)^2
      MSEA1[i]<-(1/Dim^2)*sum(verschilA1)
      MSEA2[i]<-(1/Dim^2)*sum(verschilA2)}
  }
  #Use ks.boot to allow for ties
  ksboot<-ks.boot(MSEA1,MSEA2)
  return(ksboot$ks.boot.pvalue)
}

Checkfunctie<-function(mu,var,rho,samp1,samp2,N){
  Dim<-length(mu)
  pvaluenewmethod<-numeric(N)
  pvaluekraemer<-numeric(N)
  for(i in 1:N){
    A1<-simmatrix(mu,var,rho,samp1)
    A2<-simmatrix(mu,var,rho,samp2)
    newmethod<-comparemsematrix(A1[[1]],A2[[1]],samp1,500)
    if(newmethod>0.05){
      pvaluenewmethod[i]<-1
    }
    kraemer<-Kraemertest(A1[[3]],A2[[3]],samp1,samp2,Dim)[[5]]
    if(kraemer<0.05){
      pvaluekraemer[i]<-1
    }
  }
  return(list(pvaluekraemer,pvaluenewmethod,sum(pvaluekraemer),sum(pvaluenewmethod)))
}
muzeven3<-c(-2.5,-1.75,-0.75,0,0.75,1.75,2.5)

#Test mean vector close difference, two variances
test1<- Checkfunctie(muzeven,varuneq7,0,50,50,25)
test2<- Checkfunctie(muzeven,vareq7,0,50,50,25)
test1samp100<- Checkfunctie(muzeven,varuneq7,0,100,100,25)
test2samp100<- Checkfunctie(muzeven,vareq7,0,100,100,25)
test1samp500<- Checkfunctie(muzeven,varuneq7,0,500,500,25)
test2samp500<- Checkfunctie(muzeven,vareq7,0,500,500,25)
test1samp250<- Checkfunctie(muzeven,varuneq7,0,250,250,25)
test2samp250<- Checkfunctie(muzeven,vareq7,0,250,250,25)
```

```
#Test mean vector large difference, two variances
test1largediff<- Checkfunctie(muzeven3,varuneq7,0,50,50,25)
test2largediff<- Checkfunctie(muzeven3,vareq7,0,50,50,25)
test1samp100largediff<- Checkfunctie(muzeven3,varuneq7,0,100,100,25)
test2samp100largediff<- Checkfunctie(muzeven3,vareq7,0,100,100,25)
test1samp250largediff<- Checkfunctie(muzeven3,varuneq7,0,250,250,25)
test2samp250largediff<- Checkfunctie(muzeven3,vareq7,0,250,250,25)
test1samp500largediff<- Checkfunctie(muzeven3,varuneq7,0,500,500,25)
test2samp500largediff<- Checkfunctie(muzeven3,vareq7,0,500,500,25)

#Larger difference
test1largediff2<- Checkfunctie(muzeven3*3,varuneq7,0,50,50,25)
test2largediff2<- Checkfunctie(muzeven3*3,vareq7,0,50,50,25)
test1samp100largediff2<- Checkfunctie(muzeven3*3,varuneq7,0,100,100,25)
test2samp100largediff2<- Checkfunctie(muzeven3*3,vareq7,0,100,100,25)
test1samp250largediff2<- Checkfunctie(muzeven3*3,varuneq7,0,250,250,25)
test2samp250largediff2<- Checkfunctie(muzeven3*3,vareq7,0,250,250,25)
test1samp500largediff2<- Checkfunctie(muzeven3*3,varuneq7,0,500,500,25)
test2samp500largediff2<- Checkfunctie(muzeven3*3,vareq7,0,500,500,25)

#Different sample sizes
test1diffsamp<- Checkfunctie(muzeven3*3,varuneq7,0,50,75,25)
test2diffsamp<- Checkfunctie(muzeven3*3,vareq7,0,50,75,25)
test1samp100diffsamp<- Checkfunctie(muzeven3*3,varuneq7,0,100,75,25)
test2samp100diffsamp<- Checkfunctie(muzeven3*3,vareq7,0,100,75,25)
test1samp500diffsamp<- Checkfunctie(muzeven3*3,varuneq7,0,500,100,25)
test2samp500diffsamp<- Checkfunctie(muzeven3*3,vareq7,0,500,100,25)

#Other number of objects, six
testsixobjuneqvar<-Checkfunctie(mu6,varuneq6.1,0,50,50,25)
testsixobjeqvar<-Checkfunctie(mu6,vareq6,0,50,50,25)
testsixobjuneqvarlargediff<-Checkfunctie(mu3,varuneq6.1,0,50,50,25)
testsixobjeqvarlargediff<-Checkfunctie(mu3,vareq6,0,50,50,25)
testsixobjuneqvarlargediff2<-Checkfunctie(mu5,varuneq6.1,0,50,50,25)
testsixobjeqvarlargediff2<-Checkfunctie(mu5,vareq6,0,50,50,25)

#Ten objects
testtenobjuneqvar<-Checkfunctie(mutien,varuneq10,0,50,50,25)
testtenobjeqvar<-Checkfunctie(mutien,vareq10,0,50,50,25)
testtenobjuneqvarlargediff<-Checkfunctie(mutien*3,varuneq10,0,50,50,25)
testtenobjeqvarlargediff<-Checkfunctie(mutien*3,vareq10,0,50,50,25)
testtenobjuneqvarlargedif2f<-Checkfunctie(mutien*6,varuneq10,0,50,50,25)
testtenobjeqvarlargediff2<-Checkfunctie(mutien*6,vareq10,0,50,50,25)

#Sample 1000
test1samp1000<- Checkfunctie(muzeven,varuneq7,0,1000,1000,25)
test2samp1000<- Checkfunctie(muzeven,vareq7,0,1000,1000,25)
test1samp1000largediff<- Checkfunctie(muzeven3,varuneq7,0,1000,1000,25)
test2samp1000largediff<- Checkfunctie(muzeven3,vareq7,0,1000,1000,25)
test1samp1000largediff2<- Checkfunctie(muzeven3*3,varuneq7,0,1000,1000,25)
test2samp1000largediff2<- Checkfunctie(muzeven3*3,vareq7,0,1000,1000,25)

#Test for influence of correlation
test1cor0.25<- Checkfunctie(muzeven3,varuneq7,0.25,50,50,25)
test2cor0.25<- Checkfunctie(muzeven3,vareq7,0.25,50,50,25)
```

```
test1cor0.5<- Checkfunctie(muzeven3,varuneq7,0.5,50,50,25)
test2cor0.5<- Checkfunctie(muzeven3,vareq7,0.5,50,50,25)
test1cor0.9<- Checkfunctie(muzeven3,varuneq7,0.9,50,50,25)
test2cor0.9<- Checkfunctie(muzeven3,vareq7,0.9,50,50,25)
test1samp100cor0.25<- Checkfunctie(muzeven3,varuneq7,0.25,100,100,25)
test2samp100cor0.25<- Checkfunctie(muzeven3,vareq7,0.25,100,100,25)
test1samp100cor0.5<- Checkfunctie(muzeven3,varuneq7,0.5,100,100,25)
test2samp100cor0.5<- Checkfunctie(muzeven3,vareq7,0.5,100,100,25)
test1samp100cor0.9<- Checkfunctie(muzeven3,varuneq7,0.9,100,100,25)
test2samp100cor0.9<- Checkfunctie(muzeven3,vareq7,0.9,100,100,25)
```

## C.9 Data study

```
library(gdata)
makeprefmatrix<-function(datavec,Dim){
  prefmatrix<-matrix(0,Dim,Dim)
  upperTriangle(prefmatrix,byrow=TRUE)<-as.numeric(datavec)
  for (j in 1:(Dim-1)){
    for (k in (j+1):Dim){
      if(prefmatrix[j,k]==1){
        prefmatrix[k,j]=0}
      else(prefmatrix[k,j]=1)
       }
   }
  return(prefmatrix)
}

prefmatrixlist<-function(dataframe,Dim){
  aantal<-length(dataframe)
  lijstprefmatrix<-list()
  Q<-matrix(0,Dim,Dim)
  for (i in 1:aantal){
    datavec<-dataframe[i,]
    prefmatrix<-makeprefmatrix(datavec,Dim)
    lijstprefmatrix[[i]]<-prefmatrix
    Q<-Q+prefmatrix
  }
  propmatrix<-Q/aantal
  diag(propmatrix)<-0.5
  return(list(lijstprefmatrix,Q,propmatrix))
}


Coeffagree<-function(A,N,n){
  prefvec<-c()
  for (i in 1:(n-1)){
    for (j in (i+1):n){
      prefvec<-c(prefvec,choose(A[i,j],2))
    }
  }
  for (i in 1:(n-1)){
    for (j in (i+1):n){
      prefvec<-c(prefvec,choose(A[j,i],2))
    }
  }
  som<-sum(prefvec)
  u<-((2*som)/(choose(n,2)*choose(N,2)))-1
  chi<-4/(N-2)*(som-0.5*choose(N,2)*choose(n,2)*((N-3)/(N-2)))
  df<-(choose(n,2))*((N*(N-1))/((N-2)^2))
  spvalue<-pchisq(chi,df,lower.tail=FALSE)
  return(list(som,u,pvalue))
}

numcirctriads<-function(A){
  Dim<-dim(A)[1]
  prefscore<-numeric(Dim)
```

```
  for (l in 1:Dim){
    prefscore[l]<-sum(A[l,])
  }
  adot<-0.5*(Dim-1)
  diff<-(prefscore-adot)
  num<-(Dim/24)*(Dim^2-1)-0.5*sum(diff^2)
  return(num)
}

numberofcirctriads<-function(dataframe,Dim){
  aantal<-nrow(dataframe)
  triadstot<-numeric(aantal)
  for(i in 1:aantal){
    prefmatrix<-prefmatrixlist(dataframe,Dim)[[1]][[i]]
    circtriads<-numcirctriads(prefmatrix)
    triadstot[i]<-circtriads
   }
  return(triadstot)}


BestmodelmatrixAIC<-function(propmatrix,N){
  Dim<-dim(propmatrix)[1]
  parcase5<-scaleest(t(propmatrix))
  parcase3<-MinCase3go2(propmatrix)$par
  Case5matrix<-matrix(0,ncol=Dim,nrow=Dim)
  diag(Case5matrix)<-c(rep(0.5,Dim))
  for (i in 1:Dim-1){
    for (j in (i+1):Dim){
      Case5matrix[i,j] = pnorm((1/sqrt(2))*parcase5[i]-parcase5[j])
      Case5matrix[j,i] = 1-Case5matrix[i,j]
    }
  }
  Case3matrix<-matrix(0,ncol=Dim,nrow=Dim)
  diag(Case3matrix)<-c(rep(0.5,Dim))
  for (i in 1:Dim-1){
    for (j in (i+1):Dim){
      Case3matrix[i,j] = pnorm((parcase3[i]-parcase3[j])/(sqrt(parcase3[i+Dim]+parcase3[j+Dim])))
      Case3matrix[j,i] = 1-Case3matrix[i,j]
    }
  }
  pij5<-c()
  pij3<-c()
  for (i in 1:Dim-1){
    for (j in (i+1):Dim){
      p5<-(N*propmatrix[i,j]*log(Case5matrix[i,j]))+
      ((N-N*propmatrix[i,j])*log(1-Case5matrix[i,j]))
      p3<-(N*propmatrix[i,j]*log(Case3matrix[i,j]))+
      ((N-N*propmatrix[i,j])*log(1-Case3matrix[i,j]))
      pij5<-c(pij5,p5)
      pij3<-c(pij3,p3)

    }
  }
  L5<-sum(pij5)
  L3<-sum(pij3)
```

```
  AIC5<-2*Dim-2*L5
  AIC3<-4*Dim-2*L3
  outcome<-c(AIC5,AIC3)
  minimum<-which.min(outcome)
  parametersAIC<-c()
  if (minimum ==1){
    parametersAIC<-parcase5   }
  if(minimum==2){
    parametersAIC<-parcase3 }
  return(list(parametersAIC,outcome))
}

whichcirctriads<-function(prefmatrix){
  Dim<-dim(prefmatrix)[1]
  tottriads<-choose(Dim,3)
  whichtriadvec<-c()
  for(i in 1:(Dim-2)){
    for(j in (i+1):(Dim-1)){
      for(k in (j+1):(Dim)){
        if((prefmatrix[i,j]==1)&(prefmatrix[j,k]==1)&(prefmatrix[k,i]==1))
        {whichtriadvec<-c(whichtriadvec,1)
        }
        else if((prefmatrix[i,j]==0)&(prefmatrix[j,k]==0)&(prefmatrix[k,i]==0))
        {whichtriadvec<-c(whichtriadvec,1)
        }
        else{whichtriadvec<-c(whichtriadvec,0)
        }
      }
    }
  }
  return(whichtriadvec)}

#Food preferences, version without healthy introduction text
datavoedingsnormaal<-read.csv("Voedingsmiddelencoderingnormaalonly10.csv",FALSE,";")
numcircnormaal<-numberofcirctriads(datavoedingsnormaal,7)
aantal1<-nrow(datavoedingsnormaal)
prefmatrixlijstnormaal<-prefmatrixlist(datavoedingsnormaal,7)[[1]]
totmatrixnormaal<-prefmatrixlist(datavoedingsnormaal,7)[[2]]
propmatrixnormaal<-prefmatrixlist(datavoedingsnormaal,7)[[3]]
coeffagreenormaal<-Coeffagree(totmatrixnormaal,aantal1,7)

#Food preferences, version with healthy introduction text
datavoedingsgezond<-read.csv("Voedingsmiddelencoderinggezond.csv",FALSE,";")
numcircgezond<-numberofcirctriads(datavoedingsgezond,7)
aantal2<-nrow(datavoedingsgezond)
prefmatrixlijstgezond<-prefmatrixlist(datavoedingsgezond,7)[[1]]
totmatrixgezond<-prefmatrixlist(datavoedingsgezond,7)[[2]]
propmatrixgezond<-prefmatrixlist(datavoedingsgezond,7)[[3]]
coeffagreegezond<-Coeffagree(totmatrixgezond,aantal2,7)

#Snack preferences, version without healthy introduction text
datasnacksnormaal<-read.csv("Datasnacksnormaalonly10.csv",FALSE,";")
numcircsnacksnormaal<-numberofcirctriads(datasnacksnormaal,7)
aantal3<-nrow(datasnacksnormaal)
prefmatrixlijstsnacksnormaal<-prefmatrixlist(datasnacksnormaal,7)[[1]]
```

```
totmatrixsnacksnormaal<-prefmatrixlist(datasnacksnormaal,7)[[2]]
propmatrixsnacksnormaal<-prefmatrixlist(datasnacksnormaal,7)[[3]]
coeffagreesnacksnormaal<-Coeffagree(totmatrixsnacksnormaal,aantal3,7)

#Snack preferences, version with healthy introduction text
datasnacksgezond<-read.csv("Datasnacksgezondonly10.csv",FALSE,";")
numcircsnacksgezond<-numberofcirctriads(datasnacksgezond,7)
aantal4<-nrow(datasnacksgezond)
prefmatrixlijstsnacksgezond<-prefmatrixlist(datasnacksgezond,7)[[1]]
totmatrixsnacksgezond<-prefmatrixlist(datasnacksgezond,7)[[2]]
propmatrixsnacksgezond<-prefmatrixlist(datasnacksgezond,7)[[3]]
coeffagreesnacksgezond<-Coeffagree(totmatrixsnacksgezond,aantal4,7)

#which sets of triads are judged most inconsistentsnacksly?
whichmostinconsist<-function(lijstprefmatrix){
  Dim<-dim(lijstprefmatrix[[1]])[1]
  len<-length(lijstprefmatrix)
  tottriads<-choose(Dim,3)
  totwhichtriadvec<-numeric(tottriads)
  for(i in 1:len){
    indtriadvec<-whichcirctriads(lijstprefmatrix[[i]])
    totwhichtriadvec<-totwhichtriadvec+indtriadvec}
  perctotwhichtriadvec<-totwhichtriadvec/len
  return(list(totwhichtriadvec,perctotwhichtriadvec))
}

inconsistentsnacksnormaal<-whichmostinconsist(prefmatrixlijstnormaal)
inconsistentsnacksgezond<-whichmostinconsist(prefmatrixlijstgezond)
inconsistentsnackssnacksnormaal<-whichmostinconsist(prefmatrixlijstsnacksnormaal)
inconsistentsnackssnacksgezond<-whichmostinconsist(prefmatrixlijstsnacksgezond)

#Examine how many circular triads in which one product is involved and the total number of it
#Regular version (food)
#Product 1: Pizza
sum(inconsistentnormaal[[1]][seq(1,15,1)])
length(which(inconsistentnormaal[[1]][seq(1,15,1)]>0))
#Product 2: Avocado
sum(inconsistentnormaal[[1]][c(seq(1,5,1),seq(16,25,1))])
length(which(inconsistentnormaal[[1]][c(seq(1,5,1),seq(16,25,1))]>0))
#Product 3: Nuts
sum(inconsistentnormaal[[1]][c(1,seq(6,9,1),seq(16,19,1),seq(26,31,1))])
length(which(inconsistentnormaal[[1]][c(1,seq(6,9,1),seq(16,19,1),seq(26,31,1))]>0))
#Product 4: Banana
sum(inconsistentnormaal[[1]][c(2,6,seq(10,12,1),16,seq(20,22,1),seq(26,28,1),seq(32,34,1))])
length(which(inconsistentnormaal[[1]][c(2,6,seq(10,12,1),16,
seq(20,22,1),seq(26,28,1),seq(32,34,1))]>0))
#Product 5: Marshmallows
sum(inconsistentnormaal[[1]][c(3,7,10,13,14,17,20,23,24,26,29,30,32,33,35)])
length(which(inconsistentnormaal[[1]][c(3,7,10,13,14,17,20,23,24,26,29,30,32,33,35)]>0))
#Product 6: Grilled cheese sandwich
sum(inconsistentnormaal[[1]][c(4,8,11,13,15,18,21,23,25,27,29,31,32,34,35)])
length(which(inconsistentnormaal[[1]][c(4,8,11,13,15,18,21,23,25,27,29,31,32,34,35)]>0))
#Product 7: Turkey
sum(inconsistentnormaal[[1]][c(5,9,12,14,15,19,22,24,25,28,30,31,33,34,35)])
length(which(inconsistentnormaal[[1]][c(5,9,12,14,15,19,22,24,25,28,30,31,33,34,35)]>0))
```

```
#Healthy version (food)
#Product 1: Pizza
sum(inconsistentgezond[[1]][seq(1,15,1)])
length(which(inconsistentgezond[[1]][seq(1,15,1)]>0))
#Product 2: Avocado
sum(inconsistentgezond[[1]][c(seq(1,5,1),seq(16,25,1))])
length(which(inconsistentgezond[[1]][c(seq(1,5,1),seq(16,25,1))]>0))
#Product 3: Nuts
sum(inconsistentgezond[[1]][c(1,seq(6,9,1),seq(16,19,1),seq(26,31,1))])
length(which(inconsistentgezond[[1]][c(1,seq(6,9,1),seq(16,19,1),seq(26,31,1))]>0))
#Product 4: Banana
sum(inconsistentgezond[[1]][c(2,6,seq(10,12,1),16,seq(20,22,1),seq(26,28,1),seq(32,34,1))])
length(which(inconsistentgezond[[1]][c(2,6,seq(10,12,1),16,seq(20,22,1),
seq(26,28,1),seq(32,34,1))]>0))
#Product 5: Marshmallows
sum(inconsistentgezond[[1]][c(3,7,10,13,14,17,20,23,24,26,29,30,32,33,35)])
length(which(inconsistentgezond[[1]][c(3,7,10,13,14,17,20,23,24,26,29,30,32,33,35)]>0))
#Product 6: Grilled cheese sandwich
sum(inconsistentgezond[[1]][c(4,8,11,13,15,18,21,23,25,27,29,31,32,34,35)])
length(which(inconsistentgezond[[1]][c(4,8,11,13,15,18,21,23,25,27,29,31,32,34,35)]>0))
#Product 7: Turkey
sum(inconsistentgezond[[1]][c(5,9,12,14,15,19,22,24,25,28,30,31,33,34,35)])
length(which(inconsistentgezond[[1]][c(5,9,12,14,15,19,22,24,25,28,30,31,33,34,35)]>0))

#Regular version (snacks)
#Product 1: Chips
sum(inconsistentsnacksnormaal[[1]][seq(1,15,1)])
length(which(inconsistentsnacksnormaal[[1]][seq(1,15,1)]>0))
#Product 2: Filled cookie
sum(inconsistentsnacksnormaal[[1]][c(seq(1,5,1),seq(16,25,1))])
length(which(inconsistentsnacksnormaal[[1]][c(seq(1,5,1),seq(16,25,1))]>0))
#Product 3: Rice cookie
sum(inconsistentsnacksnormaal[[1]][c(1,seq(6,9,1),seq(16,19,1),seq(26,31,1))])
length(which(inconsistentsnacksnormaal[[1]][c(1,seq(6,9,1),seq(16,19,1),seq(26,31,1))]>0))
#Product 4: Snickers
sum(inconsistentsnacksnormaal[[1]][c(2,6,seq(10,12,1),16,seq(20,22,1),seq(26,28,1),seq(32,34,1))])
length(which(inconsistentsnacksnormaal[[1]][c(2,6,seq(10,12,1),16
,seq(20,22,1),seq(26,28,1),seq(32,34,1))]>0))
#Product 5: Sponge cake
sum(inconsistentsnacksnormaal[[1]][c(3,7,10,13,14,17,20,23,24,26,29,30,32,33,35)])
length(which(inconsistentsnacksnormaal[[1]][c(3,7,10,13,14,17,20,23,24,26,29,30,32,33,35)]>0))
#Product 6: Muesli bar
sum(inconsistentsnacksnormaal[[1]][c(4,8,11,13,15,18,21,23,25,27,29,31,32,34,35)])
length(which(inconsistentsnacksnormaal[[1]][c(4,8,11,13,15,18,21,23,25,27,29,31,32,34,35)]>0))
#Product 7: Apple
sum(inconsistentsnacksnormaal[[1]][c(5,9,12,14,15,19,22,24,25,28,30,31,33,34,35)])
length(which(inconsistentsnacksnormaal[[1]][c(5,9,12,14,15,19,22,24,25,28,30,31,33,34,35)]>0))

#Healthy version (snacks)
#Product 1: Chips
sum(inconsistentsnacksgezond[[1]][seq(1,15,1)])
length(which(inconsistentsnacksgezond[[1]][seq(1,15,1)]>0))
#Product 2: Filled cookie
sum(inconsistentsnacksgezond[[1]][c(seq(1,5,1),seq(16,25,1))])
```

```
length(which(inconsistentsnacksgezond[[1]][c(seq(1,5,1),seq(16,25,1))]>0))
#Product 3: Rice cookie
sum(inconsistentsnacksgezond[[1]][c(1,seq(6,9,1),seq(16,19,1),seq(26,31,1))])
length(which(inconsistentsnacksgezond[[1]][c(1,seq(6,9,1),seq(16,19,1),seq(26,31,1))]>0))
#Product 4: Snickers
sum(inconsistentsnacksgezond[[1]][c(2,6,seq(10,12,1),16,seq(20,22,1),
seq(26,28,1),seq(32,34,1))])
length(which(inconsistentsnacksgezond[[1]][c(2,6,seq(10,12,1),16,
seq(20,22,1),seq(26,28,1),seq(32,34,1))]>0))
#Product 5: Sponge cake
sum(inconsistentsnacksgezond[[1]][c(3,7,10,13,14,17,20,23,24,26,29,30,32,33,35)])
length(which(inconsistentsnacksgezond[[1]][c(3,7,10,13,14,17,20,23,24,26,29,30,32,33,35)]>0))
#Product 6: Mueslibar
sum(inconsistentsnacksgezond[[1]][c(4,8,11,13,15,18,21,23,25,27,29,31,32,34,35)])
length(which(inconsistentsnacksgezond[[1]][c(4,8,11,13,15,18,21,23,25,27,29,31,32,34,35)]>0))
#Product 7: Apple
sum(inconsistentsnacksgezond[[1]][c(5,9,12,14,15,19,22,24,25,28,30,31,33,34,35)])
length(which(inconsistentsnacksgezond[[1]][c(5,9,12,14,15,19,22,24,25,28,30,31,33,34,35)]>0))

#Determine the best model for the data by making use of the AIC
BestmodelmatrixAIC(propmatrixnormaal,aantal1)
BestmodelmatrixAIC(propmatrixgezond,aantal2)
BestmodelmatrixAIC(propmatrixsnacksnormaal,aantal3)
BestmodelmatrixAIC(propmatrixsnacksgezond,aantal4)

#Matrix rankings of the different data sets
Rankmatrix(totmatrixnormaal)
Rankmatrix(totmatrixgezond)
Rankmatrix(totmatrixsnacksnormaal)
Rankmatrix(totmatrixsnacksgezond)

#Remove persons with circular triads
prefmatrixlijstnormaalwithouttri<-prefmatrixlijstnormaal[c(which(numcircnormaal==0))]
prefmatrixlijstgezondwithouttri<-prefmatrixlijstgezond[c(which(numcircgezond==0))]
prefmatrixlijstsnacksnormaalwithouttri<-prefmatrixlijstsnacksnormaal
[c(which(numcircsnacksnormaal==0))]
prefmatrixlijstsnacksgezondwithouttri<-prefmatrixlijstsnacksgezond
[c(which(numcircsnacksgezond==0))]

#Make new proportion and total matrices excluding individuals with intransitive preferences
makenewproptotmat<-function(lijst){
  len<-length(lijst)
  Dim<-dim(lijst[[1]])[1]
  totmatrix<-matrix(0,Dim,Dim)
  for(i in 1:len){
    totmatrix<-totmatrix+lijst[[i]]
  }
  propmatrix<-totmatrix/len
  diag(propmatrix)<-0.5
  return(list(totmatrix,propmatrix))
}

aantal1wotri<-length(prefmatrixlijstnormaalwithouttri)
propmatrixnormaalwithouttri<-makenewproptotmat(prefmatrixlijstnormaalwithouttri)[[2]]
totmatrixnormaalwithouttri<-makenewproptotmat(prefmatrixlijstnormaalwithouttri)[[1]]
```

```
aantal2wotri<-length(prefmatrixlijstgezondwithouttri)
propmatrixgezondwithouttri<-makenewproptotmat(prefmatrixlijstgezondwithouttri)[[2]]
totmatrixgezondwithouttri<-makenewproptotmat(prefmatrixlijstgezondwithouttri)[[1]]

aantal3wotri<-length(prefmatrixlijstsnacksnormaalwithouttri)
propmatrixsnacksnormaalwithouttri<-makenewproptotmat(prefmatrixlijstsnacksnormaalwithouttri)[[2]]
totmatrixsnacksnormaalwithouttri<-makenewproptotmat(prefmatrixlijstsnacksnormaalwithouttri)[[1]]

aantal4wotri<-length(prefmatrixlijstsnacksgezondwithouttri)
propmatrixsnacksgezondwithouttri<-makenewproptotmat(prefmatrixlijstsnacksgezondwithouttri)[[2]]
totmatrixsnacksgezondwithouttri<-makenewproptotmat(prefmatrixlijstsnacksgezondwithouttri)[[1]]

#Determine the best model for the new datasets making use of the AIC
BestmodelmatrixAIC(propmatrixnormaalwithouttri,aantal1wotri)
BestmodelmatrixAIC(propmatrixgezondwithouttri,aantal2wotri)
BestmodelmatrixAIC(propmatrixsnacksnormaalwithouttri,aantal3wotri)
BestmodelmatrixAIC(propmatrixsnacksgezondwithouttri,aantal4wotri)

#Coefficient of agreement
coeffagreenormaalwithouttri<-Coeffagree(totmatrixnormaalwithouttri,aantal1wotri,7)
coeffagreegezondwithouttri<-Coeffagree(totmatrixgezondwithouttri,aantal2wotri,7)
coeffagreesnacksnormaalwithouttri<-Coeffagree(totmatrixsnacksnormaalwithouttri,aantal3wotri,7)
coeffagreesnacksgezondwithouttri<-Coeffagree(totmatrixsnacksgezondwithouttri,aantal4wotri,7)

#Matrix ranks
Rankmatrix(totmatrixnormaalwithouttri)
Rankmatrix(totmatrixgezondwithouttri)
Rankmatrix(totmatrixsnacksnormaalwithouttri)
Rankmatrix(totmatrixsnacksgezondwithouttri)

#Implementing Kraemer's method
Rankingpref<-function(A){
  Dim<-dim(A)[1]
  prerank<-numeric(Dim)
  for (i in 1:Dim){
    prerank[i]<-sum(A[i,])
  }
  x<-rank(-prerank)
  return(x)
}

KendallsW<-function(datalijst,sampsize,nrobj){
  len<-length(datalijst)
  Dim<-dim(datalijst[[1]])[1]
  totalrank<-c(rep(0,nrobj))
  for(i in 1:len){
    indrank<-Rankingpref(datalijst[[i]])
    totalrank<-totalrank+indrank
  }
  meanrank<-mean(totalrank)
  diff<-totalrank-meanrank
  S<-sum(diff^2)
  W<-(12*S)/((sampsize^2)*(nrobj^3-nrobj))
  return(list(W,S))
```

```
}

Kraemertest<-function(datalijst1,datalijst2,samp1,samp2,nrobj){
  W1<-KendallsW(datalijst1,samp1,nrobj)[[1]]
  W2<-KendallsW(datalijst2,samp2,nrobj)[[1]]
  totsamp<-samp1+samp2
  WW<-(samp1*W1+samp2*W2)/(totsamp)
  datatotaal<-c(datalijst1,datalijst2)
  WT<-KendallsW(datatotaal,totsamp,7)
  WB<-(WT)/(WW)
  eta<-(WB)/(1-WB)
  pvalue<-pf(eta,nrobj-1,nrobj-1,lower.tail=FALSE)
  return(list(WW,WT,WB,eta,pvalue))
}

#Search individuals with transit
for(i in 1:length(prefmatrixlijstsnacksnormaal)){
  a<-c()
  if(whichcirctriads(prefmatrixlijstsnacksnormaal[[i]])[5]==1){print(i)}
}

prefmatrixlijstsnacksnormaaltri<-prefmatrixlijstsnacksnormaal[c(which(numcircsnacksnormaal>0))]
propmatrixsnacksnormaaltri<-makenewproptotmat(prefmatrixlijstsnacksnormaaltri)
Rankingpref(propmatrixsnacksnormaaltri[[1]])

#Are the groups the same?
voedingcoincide<-Kraemertest(prefmatrixlijstnormaalwithouttri,
prefmatrixlijstgezondwithouttri,aantal1wotri,aantal2wotri,7)
#WW = 0.2791, WT = 0.2654, WB = 0.9508, eta = 19.362, pvalue = 0.00109.
Significant at level 0.05, groups are concordant

Snackscoincide<-Kraemertest(prefmatrixlijstsnacksnormaalwithouttri,
prefmatrixlijstsnacksgezondwithouttri,
aantal3wotri,aantal4wotri,7)
#WW = 0.3653, WT = 0.1570, WB = 0.4297, eta = 0.7534, pvalue = 0.6301.
#Not significant at level 0.05,groups are not concordant


#Make plots of distributions
#Without intransitivities
#tasty
x<-seq(-5,5,0.1)
plot(x,dnorm(x,0.239,sqrt(0.75)),type="l",ylim=c(0,1),ylab="Density",
main="Plot of the distributions of the food products
     (tasty group)")
lines(x,dnorm(x,-0.342,sqrt(2.373)),type="l",col="red")
lines(x,dnorm(x,0.126,sqrt(1.085)),type="l",col="blue")
lines(x,dnorm(x,0.712,sqrt(1.176)),type="l",col="green")
lines(x,dnorm(x,-1,sqrt(0.3)),type="l",col="orange")
lines(x,dnorm(x,0.457,sqrt(0.771)),type="l",col="purple")
lines(x,dnorm(x,-0.192,sqrt(0.549)),type="l",col="pink")
legend(-5,1,legend=c("Pizza", "Avocado","Unseasoned nuts","Banana","Marshmallows",
"Grilled cheese sandwich","Turkey"),
       col=c("black", "red","blue","green","orange","purple","pink"), lty=c(rep(1,7)), cex=0.8)
```

```
#Healthy
x<-seq(-5,5,0.1)
plot(x,dnorm(x,0.762,sqrt(1.344)),type="l",ylim=c(0,2.5),ylab="Density",
main="Plot of the distributions of the food products
     (healthy group)")
lines(x,dnorm(x,-0.344,sqrt(3.560)),type="l",col="red")
lines(x,dnorm(x,-0.131,sqrt(0.422)),type="l",col="blue")
lines(x,dnorm(x,0.141,sqrt(0.540)),type="l",col="green")
lines(x,dnorm(x,-0.685,sqrt(0.034)),type="l",col="orange")
lines(x,dnorm(x,0.767,sqrt(0.990)),type="l",col="purple")
lines(x,dnorm(x,-0.510,sqrt(0.110)),type="l",col="pink")
legend(-5,2.5,legend=c("Pizza", "Avocado","Unseasoned nuts","Banana","Marshmallows",
"Grilled cheese sandwich","Turkey"),
       col=c("black", "red","blue","green","orange","purple","pink"), lty=c(rep(1,7)), cex=0.8)

#Snacks
#tasty
x<-seq(-5,5,0.1)
plot(x,dnorm(x,0.15,1),type="l",ylim=c(0,0.6),ylab="Density",
main="Plot of the distributions of the snack products
     (tasty group)")
lines(x,dnorm(x,0.046,1),type="l",col="red")
lines(x,dnorm(x,-0.574,1),type="l",col="blue")
lines(x,dnorm(x,0.015,1),type="l",col="green")
lines(x,dnorm(x,0.298,1),type="l",col="orange")
lines(x,dnorm(x,-0.230,1),type="l",col="purple")
lines(x,dnorm(x,0.293,1),type="l",col="pink")
legend(-5,0.6,legend=c("Crisps", "Almond paste cookie","Rice cookie",
"Snickers","Egg cake","Muesli bar","Apple"),
       col=c("black", "red","blue","green","orange","purple","pink"), lty=c(rep(1,7)), cex=0.8)

#Healthy
x<-seq(-5,5,0.1)
plot(x,dnorm(x,0.289,1),type="l",ylim=c(0,0.6),ylab="Density",
main="Plot of the distributions of the snack products
     (healthy group)")
lines(x,dnorm(x,-0.374,1),type="l",col="red")
lines(x,dnorm(x,-0.287,1),type="l",col="blue")
lines(x,dnorm(x,-0.142,1),type="l",col="green")
lines(x,dnorm(x,0.172,1),type="l",col="orange")
lines(x,dnorm(x,-0.182,1),type="l",col="purple")
lines(x,dnorm(x,0.524,1),type="l",col="pink")
legend(-5,0.6,legend=c("Crisps", "Almond paste cookie","Rice cookie",
"Snickers","Egg cake","Muesli bar","Apple"),
       col=c("black", "red","blue","green","orange","purple","pink"), lty=c(rep(1,7)), cex=0.8)


#Plots of distributions without intransitivities

x<-seq(-5,5,0.1)
plot(x,dnorm(x,0.169,sqrt(0.697)),type="l",ylim=c(0,1),ylab="Density",
main="Plot of the distributions of the food products
     (tasty group, without intransitive preferences)")
lines(x,dnorm(x,-0.322,sqrt(2.784)),type="l",col="red")
lines(x,dnorm(x,0.156,sqrt(0.842)),type="l",col="blue")
```

```
lines(x,dnorm(x,0.754,sqrt(1.241)),type="l",col="green")
lines(x,dnorm(x,-0.98,sqrt(0.229)),type="l",col="orange")
lines(x,dnorm(x,0.391,sqrt(0.513)),type="l",col="purple")
lines(x,dnorm(x,-0.168,sqrt(0.693)),type="l",col="pink")
legend(-5,1,legend=c("Pizza", "Avocado","Unseasoned nuts","Banana","Marshmallows",
"Grilled cheese sandwich","Turkey"),
       col=c("black", "red","blue","green","orange","purple","pink"), lty=c(rep(1,7)), cex=0.8)

#Healthy
x<-seq(-5,5,0.1)
plot(x,dnorm(x,0.763,sqrt(1.615)),type="l",ylim=c(0,2.5),ylab="Density",
main="Plot of the distributions of the food products
     (healthy group, without intransitive preferences)")
lines(x,dnorm(x,-0.384,sqrt(2.691)),type="l",col="red")
lines(x,dnorm(x,-0.118,sqrt(0.508)),type="l",col="blue")
lines(x,dnorm(x,0.191,sqrt(0.717)),type="l",col="green")
lines(x,dnorm(x,-0.843,sqrt(0.390)),type="l",col="orange")
lines(x,dnorm(x,0.955,sqrt(1.056)),type="l",col="purple")
lines(x,dnorm(x,-0.564,sqrt(0.023)),type="l",col="pink")
legend(-5,2.5,legend=c("Pizza", "Avocado","Unseasoned nuts","Banana","Marshmallows",
"Grilled cheese sandwich","Turkey"),
       col=c("black", "red","blue","green","orange","purple","pink"), lty=c(rep(1,7)), cex=0.8)

#Snacks
#tasty
x<-seq(-5,5,0.1)
plot(x,dnorm(x,0.126,1),type="l",ylim=c(0,0.6),ylab="Density",
main="Plot of the distributions of the snack products
     (tasty group, without intransitive preferences)")
lines(x,dnorm(x,-0.088,1),type="l",col="red")
lines(x,dnorm(x,-0.512,1),type="l",col="blue")
lines(x,dnorm(x,-0.021,1),type="l",col="green")
lines(x,dnorm(x,0.359,1),type="l",col="orange")
lines(x,dnorm(x,-0.246,1),type="l",col="purple")
lines(x,dnorm(x,0.382,1),type="l",col="pink")
legend(-5,0.6,legend=c("Crisps", "Almond paste cookie","Rice cookie","Snickers",
"Egg cake","Muesli bar","Apple"),
       col=c("black", "red","blue","green","orange","purple","pink"), lty=c(rep(1,7)), cex=0.8)

#Healthy
x<-seq(-5,5,0.1)
plot(x,dnorm(x,0.352,1),type="l",ylim=c(0,0.6),ylab="Density",
main="Plot of the distributions of the snack products
     (healthy group, without intransitive preferences)")
lines(x,dnorm(x,-0.433,1),type="l",col="red")
lines(x,dnorm(x,-0.244,1),type="l",col="blue")
lines(x,dnorm(x,-0.211,1),type="l",col="green")
lines(x,dnorm(x,0.184,1),type="l",col="orange")
lines(x,dnorm(x,-0.233,1),type="l",col="purple")
lines(x,dnorm(x,0.585,1),type="l",col="pink")
legend(-5,0.6,legend=c("Crisps", "Almond paste cookie","Rice cookie","Snickers",
"Egg cake","Muesli bar","Apple"),
       col=c("black", "red","blue","green","orange","purple","pink"), lty=c(rep(1,7)), cex=0.8)
```