# Comparison of Heterogeneous Probability Models for Ranking Data

Marcus, P.

# Comparison of Heterogeneous Probability Models for Ranking Data

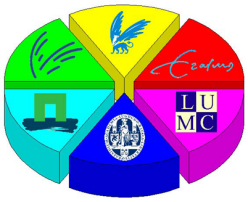
**Master Thesis**

Leiden University
Mathematics
Specialisation: Statistical Science

**Pieter Marcus**

Thesis Supervisors:
Prof. dr. W. J. Heiser
Dr. A. D'Ambrosio

**Defended on April 26, 2013**

# Contents

# 1 Introduction

Ranking data arise when a group of individuals is asked to rank a fixed set of objects according to their preferences. For example, if you want to know the preferences of members about the future president of a society.

In this thesis, we will look into partitioning methods for ranking data. Partitioning methods presuppose that a population of individual decision makers, called judges, can be decomposed into several components. Groups, components and clusters are used interchangeably, but have the same meaning. They imply that all judges in a data set can be grouped into a defined number of clusters, wherein judges rank objects more similarly than judges do in other clusters. Defining the interrelation between judges is based on a distance measure that indicates the dissimilarity between their rankings.

Cluster analysis is a statistical technique whereby groups are discovered solely based on the structure and geometry of the data at hand. It is a form of an explorative, unsupervised learning technique where no criterion measure is available. The outcome is a label or ranking that describes the group, associated by a spread parameter that takes the variability of the rankings of the group into account. It may well be that a population of judges consists of different groups of judges. The aim is to identify the appropriate number of clusters hidden in the data. Furthermore, each cluster is part of the whole population indicated by a probability of belonging to that population.

We will look at two classification methods that have been proposed for the decomposition of a heterogeneous population into a defined number of homogeneous groups. The first method $K$-Median Cluster Component Analysis (CCA), proposed by Heiser & D'Ambrosio (in press), is a clustering method where rankings are assigned with probabilities to all clusters. The second is a mixture of distance-based models (DBM) and was proposed by Murphy & Martin (2003). It is the extension of Mallows' $\phi$-model (Mallows, 1957).

We will answer the following research question in a simulation study: which clustering method is most suitable for recovering the centers. The recovery of the cluster centers is measured as an external validation criterium. In addition, we will examine which model identifies the appropriate number of clusters based on real data sets.

As of yet, only (weighted) mixtures of distance-based models based on Kendall's correlation coefficient have been implemented in the statistical package `R` (R Development Core Team, 2012) in packages developed by Lee & Yu (2011) and Gregory (2012). The aforementioned methods have been implemented in `R` and the code can be send upon request.

The structure of this thesis is as follows. The next section introduces the structure of ranking data. Section 3 describes the most used distance measures for ranking data. In section 4, summary statistics for ranking data are described. The methods for clustering ranking data are described in section 5. Section 6 gives an overview of techniques for cluster validation. In section 7 the methods are applied to simulated and real data. The discussion, to conclude with, is found in section 8.

## 2 Structure of ranking data

The collected data are listed by individuals called judges that order the set of $m$ objects with integer values from 1 to $m$. The process of ranking can be seen as the assignment of a specific permutation of these integers. Usually, we deal with $n$ independent judges, where an individual judge's ranking is denoted by $y_i$ where $i = 1, \ldots, n$. The data matrix has dimensions $n \times m$ and can be reduced by only taking the unique rankings, associated with a weight vector $w_i$ that corresponds with the frequency of ranking $y_i$.

### 2.1 Rankings and orderings

The two representations of a ranking are the rank vector and the order vector. The rank vector lists the ranks given to the objects, where 1 denotes the best rank and the value of $m$ denotes the worst rank. It is a permutation of the set of integers and presumes the objects are listed in a pre-specified order. The rank vector is denoted in between brackets. More formally, $y_i = \big(y_i(1), y_i(2), \ldots, y_i(m)\big)$ such that $y_i(r)$ is the rank given to object $r$. The order vector lists the true order of objects in order from best to worst. It is denoted in between triangular brackets, where the object labels (here, each object is given a letter) given to the object represent the order in which the objects are ordered. More formally, $y_i^{-1} = <y_i^{-1}(1), y_i^{-1}(2), \ldots, y_i^{-1}(m)>$ such that $y_i^{-1}(r)$ is the object assigned to rank $r$. The rank vector representation is used to calculate distances between pairs of rankings and to list the observed rankings, whereas the order representation is easily interpreted as summary statistic.

In this section and the next two, a small example of ranking data will be used. It concerns three movies (objects) that have to be watched with three friends (judges) during an evening. The movies are: $a$ = Into the Wild (2007), an adventure movie, $b$ = Old School (2003), a comedy, and $c$ = Stand van de Sterren, a documentary (2010). Suppose that person one prefers Into the Wild, person two prefers to watch Stand van de Sterren, while the last person is indifferent between watching Into the Wild and Old School and ranks them tied. Their preferences are listed in Table 1.

Table 1: Preferences of the friends example.

| Person | Ranking $a\ b\ c$ | Ordering |
|---|---|---|
| 1 | (1 2 3) | $<a\ b\ c>$ |
| 2 | (2 3 1) | $<c\ a\ b>$ |
| 3 | (2 1 1) | $<b$–$c\ a>$ |

### 2.2 Types of rankings

When a judge assigns distinct integer values from 1 to $m$ to all $m$ objects this is called a complete ranking, linear or full ordering. Whenever a judge fails to distinguish between two (or more) objects and assigns them equally, the literature calls this a tied ranking or a weak ordering. Allowing ties enlarges the freedom of the judges, but complicates the analysis, as we will see later on. A tied ranking can be interpreted as a positive statement of agreement or as a statement of indifference between those objects. Explanations and interpretations of tied rankings can be found in Kendall (1948, Chapter 3) and Emond

& Mason (2002, p. 24). The ordering of the third person in is a weak ordering (notice the hyphen between objects $b$ and $c$, where he cannot distinguish movies $b$ and $c$.

A further extension of complete and tied rankings is with partial and incomplete rankings, where a best subset of $q$ of $m$ objects is listed and $q < m$. Partial rankings occur when judges are asked to rank a specific subset of the entire set of objects. An example of partial rankings is the study in which people were asked to specify their top three out of five named parts of marriage (Critchlow, 1985, p. 1). With incomplete rankings there many different subsets with $m$ objects possible. The obtained data contains rankings of different object lengths. An example is the 'APA subset' data set (to be discussed in section 7.2), where 64% of the 15,449 psychologists ranked a subset of the five candidates.

If there is a single object missing, its rank can uniquely be determined. When two or more objects are not being ranked, it becomes more complicated. Baggerly (1995) suggested to treat the missing rankings as tied at the last position. By doing this, all objects are now being ranked and the obtained ranking is located on the sample space (to be discussed in the next subsection). Critchlow (1985, Chapter 3) extended the group-theoretic approach to partial rankings, using coset spaces of the sample size. Busse et al. (2007) also extended Mallows' $\phi$–model to fit it with top$-q$ rankings based on a maximum entropy model. Emond & Mason (2002, p. 23) suggested to insert a value of zero in equation (3.3) in the score matrix $a'_{rs}$ when the object is not ranked. The value of zero represents absence of information regarding that pair of objects. We do not consider partial and incomplete rankings in the rest of this thesis, because they are not located on the sample space that we start from.

## 2.3 Sample space of ranking data

With $m$ objects there are $m$ factorial possible complete rankings. When including ties this number gets even larger. Gross (1962) showed that by including tied rankings the number of possible rankings approximates $\frac{1}{2}(\frac{1}{\log(2)})^{m+1}m!$. The set $\Omega_m$ is defined as the collection of all possible permutations of $m$ objects and embraces complete and tied rankings. Any complete or tied ranking $y_i$ is an element of $\Omega_m$. The total number of rankings for up to ten objects is given in Table 2.

Table 2: Number of rankings for up to 10 objects.

| $m$ | Number of rankings |
|---|---|
| 2 | 3 |
| 3 | 13 |
| 4 | 75 |
| 5 | 541 |
| 6 | 4,683 |
| 7 | 47,293 |
| 8 | 545,835 |
| 9 | 7,087,261 |
| 10 | 102,247,563 |

## 2.4 Geometry of ranking data

The sample space of $m$ objects can be shown in a $m-1$ dimensional hyperplane. This space is called the permutation polytope and is a convex hull on the points $y_i \in \Omega_m \subset \mathbb{R}^m$, where the complete rankings form the vertices. The sample space of three objects is a hexagon given in Figure 1. Moving across the polytope goes by pairwise transposition of two adjacent objects. For example, going from ordering <$a$ $b$ $c$> to <$a$ $c$ $b$> the objects $b$ and $c$ are transposed. On the edge between these complete rankings, the weak ordering <$a$ $b$–$c$>, with a tie between objects $b$ and $c$, is located. The two new edges formed this way have the same length. In the center of the permutation polytope, the all-ties ranking is located. Every ranking, except the all-ties ranking, can be reversed. The reversal of any ranking is located at the opposite side of the polytope.

The sample space of four objects with the complete rankings is given in Figure 2. It has the shape of a truncated octahedron. It is a combination of six squares and eight hexagons, where each square connects to four hexagons. When looking at the rankings with a tie at the first or last position, these rankings form form a truncated tetrahedron. The rankings that have a tie in the middle, "are the intersection of a cube and an octahedron, forming a cuboctahedron" (Heiser & D'Ambrosio, in press). The ranking with two ties are located in the center of the squares or hexagons. The center of the squares has the ranking where the first and last two objects are tied. When looking at the six rankings that have the first and last two objects tied forming an octahedron. The rankings in the middle of the hexagons have either the first or the last object in common, the other objects are tied. The four rankings with a tie-block at the first or last position form each a tetrahedron. More about the graphical representation of ranking data can be found in Thompson (1993), Heiser (2004) and Heiser & D'Ambrosio (in press).

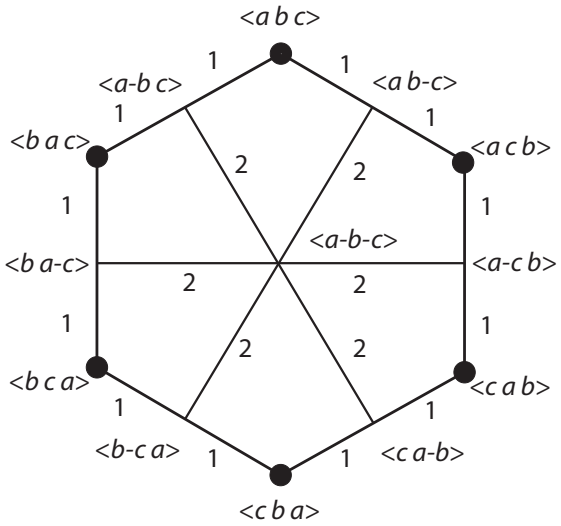



Figure 1: Sample space of 3 objects.

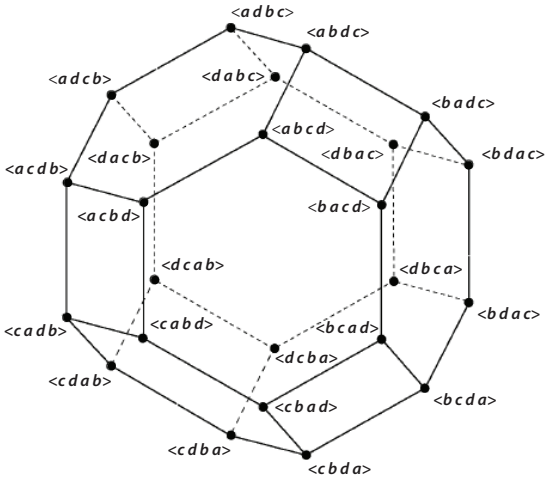



Figure 2: Sample space of 4 objects.

# 3 Distance measures for ranking data

In order to classify a heterogeneous population of $n$ judges into $K$ homogeneous clusters, we need to have a dissimilarity or distance measure defined on all rankings. Several distance measures have been proposed for ranking data. Extensive overviews are given in Critchlow (1985), Diaconis (1988) and Marden (1996). We restrict ourselves here to Kendall's and Emond & Mason's correlation coefficients, the Kemeny and the Spearman distance.

The reversal of a ranking (except the all-ties ranking) is located at the opposite side of the polytope. The reversal of a ranking has to be taken into account by a maximum distance or negative correlation. A distance measure associates to any pair of rankings $y$ and $y^*$ a distance $d(y, y^*)$. To be a valid distance measure, it needs to satisfy the following properties:

1. Reflexivity: $d(y, y) = 0$,
2. Positivity: $d(y^*, y) \geq 0$ if $y^*$ and $y \in \Omega_m$,
3. Symmetry: $d(y, y^*) = d(y^*, y) \geq 0$ if $y \neq y^*$, and
4. Triangle inequality: $d(y^*, y) \leq d(y^*, z) + d(z, y)$.

These properties are called axioms by Kemeny & Snell (1972, Chapter 2). A distance measure is said to be metric when it satisfies the triangle inequality. A label-invariant distance guarantees that the distance between two rankings remains the same even if the labels of the objects are permuted, which is a standard assumption when dealing with ranking data.

## 3.1 Kendall's correlation coefficient $\tau_b$

Kendall's correlation coefficient is probably the best known measure for ranking data (Kendall, 1948). Kendall presented multiple versions of its correlation coefficient $\tau$, where $\tau_b$ should be used when tied rankings are involved. It can be calculated in two ways. One is via the difference of the sum of the number of concordant and disconcordant pairs divided by $m(m-1)/2$ for any pair of judges. The other is by creating a score matrix of a ranking. The second alternative is applied here. A rank vector $y_i$ with $m$ objects can be transformed into a symmetric $m \times m$ score matrix $a_{rs}$, where its elements are defined by:

$$a_{rs} = \begin{cases} 1 & \text{if object } r \text{ is ranked ahead of object } s, \\ -1 & \text{if object } r \text{ is ranked behind object } s, \\ 0 & \text{if objects } r \text{ and } s \text{ are tied, or if } r = s. \end{cases} \tag{3.1}$$

The diagonal elements in the score matrix are zero and the lower triangular matrix is the reverse of the upper triangular matrix. Kendall's correlation coefficient $\tau_b$ between two judges $y$ with score matrix $a_{rs}$ and $y^*$ with score matrix $b_{rs}$ is defined as

$$\tau_b(y, y^*) = \frac{\sum_{r=1}^{m} \sum_{s=1}^{m} a_{rs} b_{rs}}{\sqrt{\sum_{r=1}^{m} \sum_{s=1}^{m} a_{rs}^2 \sum_{r=1}^{m} \sum_{s=1}^{m} b_{rs}^2}}, \tag{3.2}$$

which is the sum of the products of the elements of two score matrices, divided by the square root of the product of the sum of squares of the two score matrices. For any two identical rankings the correlation is 1. When two rankings are the reversal of each

other, they are completely dissimilar and $\tau_b$ becomes -1. Emond & Mason (2002, p. 19) pointed out that an all-ties ranking results in a zero filled score matrix and can never be estimated as a solution, because of the zeros in the numerator divided zeros in the denominator results in an unknown number. Kendall's correlation coefficient is a measure of similarity and can be transformed into a dissimilarity or distance measure via the linear transformation $d_{\tau_b} = 1 - \tau_b$, where $d_{\tau_b}$ is Kendall's distance.

### 3.2 Emond & Mason's correlation coefficient $\tau_x$

When dealing with tied rankings Emond & Mason (2002, p. 20) showed that Kendall's distance ($d_{\tau_b}$) violates the triangle inequality. An example with three rankings will illustrate this, where $A$ is the weak ordering <$a$–$b$ $c$>, $B$ the full ordering <$a$ $c$ $b$> and $C$ the full ordering <$a$ $b$ $c$> located in between $A$ and $B$. The corresponding Kendall's distance matrix is given in Table 3. The distance between $A$ and $B$ has to be smaller than the sum of the other two distances but $1.00 < 0.18 + 0.67$, hereby violating the triangle inequality.

Table 3: Triangle inequality.

| $d_{\tau_b}$ | $A$ | $B$ | $C$ |
|---|---|---|---|
| $A$ | 0.00 | 1.00 | 0.18 |
| $B$ | 1.00 | 0.00 | 0.67 |
| $C$ | 0.18 | 0.67 | 0.00 |

To solve this difficulty, Emond & Mason (2000, p. 11–12 and 2002, p. 21) redesigned the elements in Kendall's $\tau_b$ score matrix in equation (3.1) and renamed it to $\tau_x$, where $x$ stands for extension. The elements in the new score matrix $a'_{rs}$ for rank vector $y_i$ are now defined by

$$a'_{rs} = \begin{cases} 1 & \text{if object } r \text{ is ranked ahead or tied with object } s, \\ -1 & \text{if object } r \text{ is ranked behind object } s, \\ 0 & \text{if } r = s. \end{cases} \tag{3.3}$$

Again, the score matrix defines the diagonal elements with zeros but all off-diagonal elements are either -1 or 1 including tied objects. Emond & Mason (2000, p. 12) showed that the value assigned to a tied ranking in equation (3.3) can also be -1. Their correlation coefficient between two complete or tied rankings $y$ and $y^*$ is defined as

$$\tau_x(y, y^*) = \frac{\sum_{r=1}^{m} \sum_{s=1}^{m} a'_{rs} b'_{rs}}{m(m-1)}. \tag{3.4}$$

The denominator is adjusted such that the correlation between a tied ranking and itself is 1. Emond & Mason's distance ($d_{\tau_x}$) is equal to $d_{\tau_x} = 1 - \tau_x$.

### 3.3 Kemeny distance $d_{\mathbf{Kem}}$

Another distance measure that has been developed independently from Kendall's distance is the Kemeny distance, Kemeny (1959, p. 586–590) and Kemeny & Snell (1972, Chapter 2). This distance measure satisfies all properties stated earlier, including the triangular inequality. Like the correlation coefficients, the rank vector $y_i$ is transformed

into a score matrix. It is defined by the same representation of elements given in equation (3.1). The Kemeny distance allows complete and tied rankings and is defined between two rankings $y$ and $y^*$ by

$$d_{\text{Kem}}(y, y^*) = \frac{1}{2} \sum_{r=1}^{m} \sum_{s=1}^{m} \big| a_{rs} - b_{rs} \big|. \tag{3.5}$$

Thus, the Kemeny distance is the sum of the absolute differences of the two score matrices divided by two. The factor a half takes into account that the two triangular matrices that are created by the sum of absolute differences of the score matrices are identical. The Kemeny distance is of city block type and a geodesic distance in the permutation polytope. It takes the shortest path between two rankings.

The associated Kemeny distances with the sample space of three objects are also printed in Figure 1. The Kemeny distance between two complete rankings is always even. From a complete ranking to a tied ranking between two objects has an uneven distance, to three tied rankings again has an even distance. The maximum distance from a complete ranking to its reversal is $m(m-1)$. The reversal of a tied ranking is shorter than for complete rankings. The distance between the weak ordering $<a\ b\text{–}c>$ and its reversal $<c\text{–}b\ a>$ is four by going through the center, as can be seen in Figure 1. In general, the maximum distance of a ranking containing $t$ ties is given by: $m(m-1)-2t$.

Emond & Mason (2002, p. 25–26) proved that the Kemeny distance is equivalent to their correlation coefficient for complete and tied rankings by

$$\tau_x(y, y^*) = 1 - \frac{2 d_{\text{Kem}}(y, y^*)}{m(m-1)}, \tag{3.6}$$

where the denominator is the maximum Kemeny distance with $m$ objects to transform a correlation coefficient into a distance measure and vice versa.

### 3.4 Spearman's distance $d_S$

The Spearman's distance is calculated by taking the square root of the well know Spearman's $\rho$. The Spearman's distance between two rank vectors $y$ and $y^*$ is defined by

$$d_S(y, y^*) = \sqrt{\sum_{r=1}^{m} \big( y(r) - y^*(r) \big)^2}, \tag{3.7}$$

which is the square root of the sum of the squared rank differences. When a ranking contains tied objects, these objects must be given the average of the corresponding rank values. For example, the ordering of person three in the friends example obtains rank values of $(1\frac{1}{2}\ 1\frac{1}{2}\ 3)$.

A problem identified by Emond (1997, p. 4) and Emond & Mason (2000, p. 16) showed that Spearman's $\rho$ suffers from what is known as the sensitivity to irrelevant alternatives. The most simple case wherein three judges order two objects: $<a\ b>$, $<a\ b>$ and $<b\ a>$. The most obvious solution would be $<a\ b>$. If we would add two irrelevant tied objects to judges one and two and two extra objects behind object one to judge three. The new orderings become: $<a\ b\ c\text{–}d>$, $<a\ b\ c\text{–}d>$ and $<b\ c\text{–}d\ a>$. The maximum agreement now is $<b\ a\ c\text{–}d>$, where the first two objects are transposed. The addition of two irrelevant objects according to Emond & Mason (2000, p. 17): "This anomaly appears to occur because Spearman's $\rho$ uses the ranks as if they were

variate values instead of purely order values. Because of this sensitivity to irrelevant alternatives, Spearman's $\rho$ is not suitable as a rank correlation coefficient in the weighted rankings problem."

If we return to the friends stated in section 2, we can create the score matrices of the three persons given by the elements in (3.1) and (3.3). The results are listed in the tables in Table 4. The tied ranking of the third person between objects $b$ and $c$ in the second row is given for both representations because it handles ties differently.

Table 4: Score matrices of the friends example.

| $a_{rs}$ | Person 1 | | |
|---|---|---|---|
| | $a$ | $b$ | $c$ |
| $a$ | 0 | 1 | 1 |
| $b$ | -1 | 0 | 1 |
| $c$ | -1 | -1 | 0 |

| $a_{rs}$ | Person 2 | | |
|---|---|---|---|
| | $a$ | $b$ | $c$ |
| $a$ | 0 | 1 | -1 |
| $b$ | -1 | 0 | -1 |
| $c$ | 1 | 1 | 0 |

| $a_{rs}$ | Person 3 | | |
|---|---|---|---|
| | $a$ | $b$ | $c$ |
| $a$ | 0 | -1 | -1 |
| $b$ | 1 | 0 | 0 |
| $c$ | 1 | 0 | 0 |

| $a'_{rs}$ | Person 3 | | |
|---|---|---|---|
| | $a$ | $b$ | $c$ |
| $a$ | 0 | -1 | -1 |
| $b$ | 1 | 0 | 1 |
| $c$ | 1 | 1 | 0 |

# 4 Summarizing ranking data

When summarizing a sample of judges, we look for that particular ranking which describes the data best. We distinguish between the modal, median and mean ranking. With univariate interval data, it is not hard to identify the median and mean. However, the sample space of ranking data explained in section 2.4 and the associated distance measures discussed in the previous section, making it more complicated.

## 4.1 Modal ranking

The modal ranking is the ranking with highest frequency present in the data

$$\hat{c}_{\text{mode}} = \arg\max_{i} w_i y_i. \tag{4.1}$$

It is the ranking given by most judges in the sample. A problem that can occur is whenever two or more rankings are equally often observed.

## 4.2 Median ranking

The median ranking maximizes the agreement of judges' preferences in the sample. It is defined as the ranking that minimizes the Kemeny distance of all observed rankings to the rankings in the sample space (Kemeny & Snell, 1972, p. 19). So, minimizing the Kemeny distance is equivalent to maximizing Emond & Mason's correlation coefficient. The median ranking is determined by

$$\hat{c}_{\text{median}} = \arg\min_{c \in \Omega_m} \sum_{i=1}^{n} w_i d_{\text{Kem}}(y_i, c), \tag{4.2}$$

where $w_i$ is a non-negative weight vector taking the frequency of ranking $y_i$ into account and $\Omega_m$ the sample space of $m$ objects. The use of weights highly reduces computation time, especially when many judges rank the objects similarly. The median ranking does not have to be an observed ranking, but it is always located in the sample space. The interpretation of the median ranking is the maximization of the judges agreement, since it located closest to everyone's preferences. The median ranking has especially been of interest by researchers in the field of social choice theory, where it is called the consensus ranking (Regenwetter et al., 2006).

A problem of estimating the median ranking is that it may not always be uniquely defined, as was pointed out by Kemeny & Snell (1972, p. 20) and Marden (1996, p. 21). For example, if we take four complete orderings of four objects: $<a\ b\ c\ d>$, $<a\ d\ c\ b>$, $<c\ a\ d\ b>$ and $<c\ a\ b\ d>$. Nine orderings satisfy the minimum Kemeny distance of twelve: $<a\ c\ b\ d>$, $<a\ c\ d\ b>$, $<c\ a\ b\ d>$, $<c\ a\ d\ b>$, $<a$–$c\ b$–$d>$, $<a$–$c\ b\ d>$, $<a$–$c\ d\ b>$, $<a\ c\ b$–$d>$ and $<c\ a\ b$–$d>$. All of them have object $a$ up front or tied with object $c$, but it is far from convenient. When the distance between a few rankings is larger, even more rankings qualify as median ranking.

In addition, finding the median ranking is a known NP-hard problem, meaning that it is not possible to find the median ranking of $m$ objects in polynomial time. Recall the number of possible rankings with $m$ objects in Table 2. Emond & Mason (2000 and 2002) developed a branch and bound algorithm that works with up to twenty objects and speeds up this process. Their method is based on $\tau_x$, but instead of maximizing the correlation coefficient the Kemeny distance could be minimized as well. However, for

the computation in this thesis a brute force approach is implemented by enumerating all possible rankings and finding the median ranking by an exhaustive search.

### 4.3 Mean ranking

Ranking data can also be summarized by the mean ranking (Kemeny & Snell, 1972, p. 19). With the Kemeny distance it is estimated by

$$\hat{c}_{\text{mean}} = \underset{c \in \Omega_m}{\arg\min} \sum_{i=1}^{n} w_i d_{\text{Kem}}(y_i, c)^2. \tag{4.3}$$

It penalizes larger distances harder, since the sum of the squared distances has to be minimized. The mean ranking is also located on the sample space, but tends to prefer tied rankings. It can be highly impractical as the next example shows.

Let us summarize the friends example introduced in section 2. Suppose that they want to decide upon the order of movies to watch. There is no modal ranking, because all three rankings are different. The median ranking coincides with the preferences of the second person, namely $<c\ a\ b>$. This ranking minimizes the sum of Kemeny distances in the next-to-last column in Table 5. The mean ranking on the other hand is highly uninformative. The ranking that minimizes the sum of squared distances given in the last column is the all-ties ranking, where every movie goes. The most informative summary statistic is the median ranking to determine the order of movies to watch.

When calculating Emond & Mason's correlation coefficient between persons two and three, one will obtain zero correlation. This can be explained by the location of person three which is perpendicular to the location of person two and its opposite ranking. The reverse approach is also valid. From the Kemeny distance framework this can be motivated by half the maximum distance.

Table 5: Kemeny distances of the friends example.

| Ordering | Person 1 | Person 2 | Person 3 | $\sum d_{\text{Kem}}$ | $\sum d^2_{\text{Kem}}$ |
|---|---|---|---|---|---|
| $<a\ b\ c>$ | 0 | 4 | 5 | 9 | 41 |
| $<a\ b–c>$ | 1 | 3 | 6 | 10 | 46 |
| $<a\ c\ b>$ | 2 | 2 | 5 | 9 | 33 |
| $<a–c\ b>$ | 3 | 1 | 4 | 8 | 26 |
| $<c\ a\ b>$ | 4 | 0 | 3 | **7** | 25 |
| $<c\ a–b>$ | 5 | 1 | 2 | 8 | 30 |
| $<c\ b\ a>$ | 6 | 2 | 1 | 9 | 41 |
| $<b–c\ a>$ | 5 | 3 | 0 | 8 | 34 |
| $<b\ c\ a>$ | 4 | 4 | 1 | 9 | 33 |
| $<b\ a–c>$ | 3 | 5 | 2 | 10 | 38 |
| $<b\ a\ c>$ | 2 | 6 | 3 | 11 | 49 |
| $<a–b\ c>$ | 1 | 5 | 4 | 10 | 42 |
| $<a–b–c>$ | 3 | 3 | 2 | 8 | **22** |

# 5 Partitioning methods for ranking data

Given the rankings by a heterogeneous sample of $n$ judges, we want to partition them into $K$ homogeneous components or clusters. The following two probabilistic clustering methods assume that within a cluster judges rank objects more similarly than in other clusters. The maximum number of clusters $K$ should be smaller than the total number of unique rankings in the sample. When $K$ is equal to the number of unique rankings, then all centers correspond to these rankings with probability inversely related to their observed frequencies. The center of each cluster is given by the median ranking. So, the ranking defining each cluster is in best agreement with the judges of that cluster.

## 5.1 $K$-Median cluster component analysis

$K$-Median Cluster Component Analysis (CCA) was proposed by Heiser & D'Ambrosio (in press). It is an iterative partitioning algorithm for ranking data. It is a form of soft clustering where each ranking is assigned to all $K$ clusters by a membership probability, a degree of belonging to that cluster. The membership matrix $u_{ik}$ has dimensions $n \times K$ and the probabilities sum to 1 over all clusters. The algorithm is built on the probabilistic clustering framework of Ben-Israel & Iyigun (2008). The working principle of probabilistic clustering states that probabilities and distances are inversely related and their product is constant, so for an individual ranking $y_i$:

$$u_{ik}(y_i)d_{\text{Kem}}(y_i, c_k) = \text{constant, depending on } y_i.$$

Ben-Israel & Iyigun state that a cluster center should not coincide with an observed ranking. This cannot always be true with CCA, where the median ranking may coincide with an observed ranking.

The algorithm works as follows. After initializing the algorithm with $K$ random rankings as centers, the algorithm proceeds by alternating between two steps. The first step estimates the membership probability. The second step updates each cluster's median ranking, depending on the membership probabilities. The membership probabilities for ranking $y_i$ are estimated by

$$\hat{u}_{ik}(y_i) = \frac{\prod_{l \neq k} d_{\text{Kem}}(y_i, c_l)}{\sum_{k'=1}^{K} \prod_{l \neq k'} d_{\text{Kem}}(y_i, c_l)}, k = 1, \ldots, K, \tag{5.1}$$

where the membership probability to cluster $k$ is determined by the product of distances except cluster $k$ divided by the sum of all products, making the denominator a normalizing constant. When a ranking coincides with its center it receives probability one, because for all other membership probabilities multiplying a distance of zero in the numerator will always result in a zero probability.

The second step of the algorithm is to update the cluster centers by its median ranking. The center of each component $k$ is estimated by the weighted median ranking

$$\hat{c}_k = \underset{c_k \in \Omega_m}{\arg\min} \sum_{i=1}^{n} w_i u_{ik} d_{\text{Kem}}(y_i, c_k). \tag{5.2}$$

It is equal to the median ranking defined earlier in equation (4.2), but also takes into account the membership probabilities. In each iteration only the membership probabilities change. In order to force the centers to change and to speed up the algorithm timing, we decided to crisp the membership probabilities. Crisping means that the highest membership probability of a ranking gets probability one and zero probability otherwise.

The probability of each cluster in the population of judges is estimated by

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^{n} w_i u_{ik}, \tag{5.3}$$

which is the weighted average of the membership probabilities or weighted column means of the membership matrix.

A measure of homogeneity of each cluster is estimated by

$$\hat{h}_k = \frac{\sum_{i=1}^{n} w_i u_{ik} \tau_x(y_i, c_k)}{\sum_{i=1}^{n} w_i u_{ik}} \tag{5.4}$$

and is the weighted sum of the product of the membership probabilities and Emond & Mason's $\tau_x$, divided by the weighted sum of the membership probabilities. If the homogeneity is close to 1, then all rankings of that cluster are closely located around the center and is smaller otherwise.

The individual joint distance function (JDF) $D(y_i)$ measures the classificability of a single ranking $y_i$ with respect to all centers

$$D(y_i) = \frac{\prod_{k=1}^{K} d_{\text{Kem}}(y_i, c_k)}{\sum_{k'=1}^{K} \prod_{l \neq k'} d_{\text{Kem}}(y_i, c_l)}, \tag{5.5}$$

where the numerator is the product of Kemeny distances to all centers and the denominator is equal to the denominator of estimating the membership probabilities. It is different from equation (5.1) where the distance to the center $c_k$ in the numerator is not taken into account. When a ranking coincides with the estimated center, the JDF is zero. The JDF of the entire sample of judges is: $D_{\text{total}} = \sum_{i=1}^{n} w_i D(y_i)$.

The output of CCA is the membership matrix and the median rankings but also the homogeneity and mixing probability of each cluster. With each iteration the location of the centers may change. There are two ways for the CCA algorithm to converge. According to Ben-Israel & Iyigun the probabilistic distance clustering algorithm converges if

$$\sum_{k=1}^{K} d_{\text{Kem}}\left(c_k^{(l)}, c_k^{(l-1)}\right) < \epsilon,$$

meaning that the sum of Kemeny distances between the current $(l)$ and previous $(l-1)$ iteration for all centers should be smaller or equal than $\epsilon$. Practically, it converges whenever the centers stop changing, since it is the sum of integers and $\epsilon$ is small. But according to Heiser & D'Ambrosio (in press), it converges if the objective loss function is minimized. The loss function is defined as

$$\text{loss}(u_{ik}, c_k) = \sum_{i=1}^{n} w_i \sum_{k=1}^{K} u_{ik}^2 d_{\text{Kem}}(y_i, c_k), \tag{5.6}$$

the incorrect classification at the $l^{\text{th}}$ iteration. It decreases monotonically until the difference in loss between two consecutive iterations is small to converge if

$$\text{loss}(u_{ik}, c_k)^{(l)} - \text{loss}(u_{ik}, c_k)^{(l-1)} \leq \epsilon, \tag{5.7}$$

where $\epsilon = 10\text{e}^{-6}$. It is possible that in a single run the loss may not have reached its global minimum. If the algorithm estimates two (or more) similar centers and these rankings correspond with an observed ranking, then all membership probabilities in equation (5.1) of that ranking are zero and do not sum up to one. To prevent this from occurring and to minimize the loss over a range of rankings to initialize the algorithm with, it is implemented with 50 starts with $K$ different rankings from the sample space. The starting value that minimizes the overall loss is used as final estimate for $\hat{c}_1, \ldots, \hat{c}_K$, $\hat{p}_1, \ldots, \hat{p}_K$ and $\hat{h}_1, \ldots, \hat{h}_K$. The CCA algorithm can be summarized as given in Table 6.

The algorithm can be demonstrated with $n$ unique rankings (without the weight vector $w_i$) and two clusters. The optimality can be shown, where:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n}\sum_{k=1}^{2} u_{ik}^2(y_i) \cdot d_{\text{Kem}}(y_i, c_k) \\ \text{subject to} \quad & u_{i1}(y_i) + u_{i2}(y_i) = 1 \text{ and} \\ & u_{i1}(y_i),\ u_{i2}(y_i) \geq 0, \text{ for } i = 1, \ldots, n. \end{aligned}$$

The probabilities are squared because it is a smoothed version of the original function, wherein the derivatives of the probabilities are linear. The Lagrangian of this problem is:

$$L(p_1, p_2, \lambda) = u_{i1}^2(y_i) \cdot d_{\text{Kem}}(y_i, c_1) + u_{i2}^2(y_i) \cdot d_{\text{Kem}}(y_i, c_2) - \lambda\big(u_{i1}(y_i) + u_{i2}(y_i) - 1\big)$$

When taking the partial derivatives with respect to $u_{i1}(y_i)$ and $u_{i2}(y_i)$ and set them to zero. The result is

$$\begin{aligned} &2u_{i1}(y_i) \cdot d_{\text{Kem}}(y_i, c_1) + \lambda = 0 \text{ and} \\ &2u_{i2}(y_i) \cdot d_{\text{Kem}}(y_i, c_2) + \lambda = 0. \end{aligned}$$

Under the working principle of probabilistic clustering this is

$$u_{i1}(y_i) \cdot d_{\text{Kem}}(y_i, c_1) = u_{i2}(y_i) \cdot d_{\text{Kem}}(y_i, c_2).$$

To obtain the optimal value for the Lagrangian of all $n$ unique rankings is

$$L(u_{i1}, u_{i2}, \lambda) = \sum_{i=1}^{n} \frac{d_{\text{Kem}}(y_i, c_1) \cdot d_{\text{Kem}}(y_i, c_2)}{d_{\text{Kem}}(y_i, c_1) + d_{\text{Kem}}(y_i, c_2)}.$$

This corresponds with the JDF of the entire sample. So, minimizing the loss is minimizing the JDF. Due to this membership probability matrix it is a form of fuzzy or soft clustering. In fuzzy clustering each ranking is a member of all clusters associated by a membership probability. An advantage of fuzzy clustering is that the membership probability matrix generates more information than deterministic (hard) clustering. In hard clustering rankings are assigned to a single component $u_{ik} = 1$ to some $k$ $\forall i$ (Gordon, 1999). Hard clustering results can still be obtained by assigning the ranking to that cluster with highest membership probability.

Table 6: Summary of the CCA algorithm.

| Step | Procedure |
|---|---|
| 1 | Initialize $K$ different random centers, $(l = 0)$. |
| 2a | Calculate membership probabilities, $\hat{u}_{ik}$. |
| 2b | Update median ranking cluster centers, $\hat{c}_k$. |
| 3 | Repeat steps 2a and 2b until the difference in loss between two iterations $(l)$ and $(l-1) \leq \epsilon$. |
| 4 | Final solution minimizes the loss over multiple starts. |

## 5.2 Mixture of distance-based models

The second method is based on Mallows' seminal paper (1957). Mallows' model assumes that the probability of observing a ranking depends on the distance between the observed rankings and the central ranking. The distance-based model (DBM) also assumes that rankings that have equal distances from the central ranking should have equal probability and the further away from the central ranking the probability decreases. The most common used distance-based model is based on Kendall's distance ($d_{\tau_b}$) and is better known as Mallows' $\phi$-model (Mallows, 1957 and Marden, 1996). It belongs to the family of exponential distributions

$$f(y_i) = \mathrm{P}(y_i|c, \lambda) = \frac{1}{Z(\lambda)} \exp\big(-\lambda d_{\tau_b}(y_i, c)\big), \tag{5.8}$$

where the probability of observing ranking $y_i$ depends on the negative exponential of the spread parameter $\lambda$ and Kendall's distance to center $c$. $Z(\lambda)$ is the normalizing constant making the density integrates to 1.

This model has been extended to a mixture model by Murphy & Martin (2003). An extensive overview about mixture models can be found in McLachlan & Peel (2000). A mixture model assumes that the observed rankings are coming from $K$ probability distributions where each distribution represents a cluster. Each cluster has mixing probability $p_k$, with $0 \leq p_k \leq 1$ of being represented in the population and $\sum_{k=1}^{K} p_k = 1$. Each density $f_k$ has central ranking $c_k$ and spread parameter $\lambda_k$. The complete density of the mixture model is

$$f(y_i) = \sum_{k=1}^{K} p_k f_k(y_i|c_k, \lambda_k). \tag{5.9}$$

Transforming the distance-based model into a mixture of distance-based components means combining equation (5.8) of the population model with the mixture of densities in equation (5.9). The mixture model becomes

$$f(y_i) = \sum_{k=1}^{K} p_k \frac{1}{Z(\lambda_k)} \exp\big(-\lambda_k d_{\tau_b}(y_i, c_k)\big). \tag{5.10}$$

This expression illustrates that each component has its own central ranking $c_k$, spread parameter $\lambda_k$ and mixing probability $p_k$. $Z(\lambda_k)$ is the normalizing constant and depends on the spread parameter and is given in Critchlow (1985, p. 98) by

$$Z(\lambda_k) = \sum_{y_i \in c_k} \exp\big(-\lambda_k d_{\tau_b}(y_i, c_k)\big). \tag{5.11}$$

The likelihood function of the mixture of the weighted distance-based model is

$$L = \prod_{i=1}^{n} w_i \sum_{k=1}^{K} p_k \frac{1}{Z(\lambda_k)} \exp\big( - \lambda_k d_{\tau_b}(y_i, c_k)\big). \tag{5.12}$$

The model is fitted by maximum likelihood using the EM algorithm to obtain maximum likelihood estimates. The EM algorithm is well known for obtaining maximum likelihood estimates with incomplete data (Dempster et al., 1977). Fitting this model with unknown components, a latent Bernoulli variable for allocating rankings to components $u_{ik}$ is introduced to the log-likelihood:
$\ell(c, \lambda, p|y_i) = \sum_{i=1}^{n} w_i \log\big\{ \sum_{k=1}^{K} p_k \frac{1}{Z(\lambda_k)} \exp\big(-\lambda_k d_{\tau_b}(y_i, c_k)\big)\big\}$. This latent variable $u_{ik}$ indicates the probability of ranking $y_i$ belonging to component $k$. The complete-data consists of both $y_i$ and $u_{ik}$. This allocation matrix has dimension $n \times K$, similar to CCA. By including $u_{ik}$ the complete-data log-likelihood becomes

$$\ell_C(p, c, \lambda|y_i, u_{ik}) = \sum_{i=1}^{n} w_i \sum_{k=1}^{K} u_{ik}\Big\{ \log(p_k) - \log\big(Z(\lambda_k)\big) - \lambda_k d_{\tau_b}(y_i, c_k)\Big\}. \tag{5.13}$$

The EM algorithm is applied to the complete-data log-likelihood and iteratively improves maximum likelihood estimates by alternating between two steps, the expectation (E) and maximization (M) step. It heavily depends on starting values, therefore the algorithm is initiated with 50 different allocation matrices sampled from the uniform distribution. Murphy & Martin (2003) and Lee & Yu (2012) fitted mixture models to ranking data and they implemented it with 30 and 50 starts, respectively.

The E-step takes the expectation of the complete-data log-likelihood conditional on the ranking data depending on the current parameter estimates in the $l^{\text{th}}$ iteration of the algorithm

$$\hat{u}_{ik} = \frac{p_k f(y_i|c_k, \lambda_k)}{\sum_{k'=1}^{K} p_{k'} f(y_i|c_{k'}, \lambda_{k'})}. \tag{5.14}$$

This can be interpreted as the posterior probability via Bayes' theorem of belonging to cluster $k$, conditioned on the center and spread parameters.

The M-step maximizes the expected conditional complete log-likelihood given $u_{ik}$ with respect to the central ranking $c_k$ and spread $\lambda_k$. The central ranking of each cluster is the median ranking estimated by

$$\hat{c}_k = \underset{c_k \,\in\, \Omega_m}{\arg\min} \sum_{i=1}^{n} w_i u_{ik} d_{\tau_b}(y_i, c_k). \tag{5.15}$$

The spread parameter can be estimated in two ways. One way is restricting $\lambda$ to be equal for all $K$ clusters $\lambda = \lambda_1, \ldots, \lambda_K$ is

$$\hat{\lambda} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n} w_i u_{ik} d_{\tau_b}(y_i, c_k)}{\sum_{k=1}^{K} \sum_{i=1}^{n} w_i u_{ik}}. \tag{5.16}$$

The other way is estimating unrestricted spread parameters $\lambda_k$ for each cluster $\lambda_1, \ldots, \lambda_K$ via

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{n} w_i u_{ik} d_{\tau_b}(y_i, c_k)}{\sum_{i=1}^{n} w_i u_{ik}}. \tag{5.17}$$

Similar to CCA, we can estimate the mixing probabilities $p_k$ by

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^{n} w_i u_{ik}. \tag{5.18}$$

When estimates of the mixing probabilities, spread parameters and the median rankings are obtained, the complete-data log-likelihood can be calculated with equation (5.13). The EM algorithm converges if the absolute change in complete-data log-likelihood between the current $\ell_C^{(l)}(\Psi_K)$ iteration and the previous $\ell_C^{(l-1)}(\Psi_K)$ iteration is small

$$\left| \frac{\ell_C^{(l)}(\Psi_K) - \ell_C^{(l-1)}(\Psi_K)}{\ell_C^{(l-1)}(\Psi_K)} \right| \leq \epsilon, \tag{5.19}$$

where $\Psi_K$ contains the estimated parameters $(\hat{c}_1, \ldots, \hat{c}_K, \hat{\lambda}_1, \ldots, \hat{\lambda}_K, \hat{p}_1, \ldots, \hat{p}_K)$ and $\epsilon = 10\text{e}^{-6}$. The DBM with unrestricted spread parameters is fitted with $3 \times K$ parameters because for each cluster $c_k$, $\lambda_k$ and $p_k$ have to be estimated. By restricting the spread parameter, the model is fitted with $2 \times K + 1$ parameters. The mixture of distance-based model algorithm can be summarized as given in Table 7. The EM algorithm has a few but noteworthy drawbacks. It may fail to converge to the global optimum and get stuck at a local optimum or even fails to reach maximum likelihood values at all, because of bad starting values. To overcome this issue it is implemented with 50 different starts. Each start has a different matrix $u_{ik}$ and a maximum of 1,000 iterations. The solution that maximizes the complete-data log-likelihood is used as final estimate. Since the iterative nature of the procedure, it can be slow regarding the number of iterations necessary.

Table 7: Summary of the EM algorithm for distance-based models.

| Step | Procedure |
|---|---|
| 1 | Give initial values for $u_{ik}$, $(l = 0)$. |
| | **E step** |
| 2 | Estimate $\hat{u}_{ik}$. |
| | **M step** |
| 3a | Estimate $\hat{c}_k, \hat{p}_k, \hat{\lambda}_k$ and $P(y_i)$. |
| 3b | Calculate the complete-data log-likelihood. |
| 4 | Repeat steps 2 and 3 until the absolute change in complete-data log-likelihood between two iterations $l$ and $l-1$ is smaller than or equal to $\epsilon$. |
| 5 | Final solution maximizes the complete-data log-likelihood over multiple starts. |

# 6 Validation methods

When the methods discussed in the previous section are used in practice, the results need to be analyzed objectively (Gordon, 1999). One issue that needs to be addressed is the number of $K$ clusters hidden in the data. The number of clusters has to be verified quantitatively, because it is generally of major interest to researchers. There is no golden standard, since different algorithms partition the data differently. There are two ways to evaluate the results of a clustering, namely internal and external evaluation. Several criteria are described in the following two subsections.

## 6.1 Internal criteria

Internal criteria only depend on the structure of the observed data. They focus on the compactness, connectedness or isolation of the clusters and aim to give an answer to the issue regarding the selection of the appropriate number of clusters.

### 6.1.1 Partition coefficient

Bezdek (1974) proposed the partition coefficient (PC) as a performance measure based on minimizing the overall information in the membership matrix $u_{ik}$. It is a measure for compactness of the clustering. He defined the index as follows

$$\mathrm{PC}(K) = \frac{1}{n} \sum_{i=1}^{n} w_i \sum_{k=1}^{K} u_{ik}^2, \tag{6.1}$$

where $n$ is the number of judges in the sample, $w_i$ the weight corresponding membership probability $u_{ik}$ and $K$ the number of clusters in the partition. It is the weighted average of the squared membership probabilities. The minimum number of clusters to evaluate is two. It ranges between $1/K$ and 1. When the PC is 1, the result of the clustering is completely crisp and when it is close to $1/K$ this indicates no clustering tendency in the data set or in the clustering algorithm to reveal it. A disadvantage of the PC is that it only relies on the membership matrix. When dealing with the PC we look for the maximum.

### 6.1.2 Partition entropy

Another index by Bezdek (1974) is the partition entropy (PE) defined as

$$\mathrm{PE}(K) = -\frac{1}{n} \sum_{i=1}^{n} w_i \sum_{k=1}^{K} u_{ik} \log(u_{ik}), \tag{6.2}$$

where $n$ is the number of judges in the sample, $w_i$ the weight corresponding membership probability $u_{ik}$ and $K$ the number of clusters. Again, the minimum number of clusters to evaluate is two. The PE ranges between $1/\log(K)$ and 1. When the PE is $1/\log(K)$ the partition is a crisp partition and when it comes close to 1 absence of any clustering structure. The best clustering is obtained when the PE is minimum.

### 6.1.3 Joint distance function

The Joint distance function (JDF) of the entire sample ($D_{\text{total}}$) is used as a convergence criterion of CCA. The JDF decreases monotonically. According to Iyigun (2007, p. 61–63) we look for a 'knee' as a function of the JDF and the number of clusters. To identify

the appropriate number of clusters in the data, a line through the largest number of clusters and the second largest number of clusters. Once the JDF of the entire sample significantly deviates from this line, the number of clusters is found.

### 6.1.4 Bayesian information criterion

To select the most appropriate number of components in a mixture model, the Bayesian information criterion (BIC) by Schwarz (1978) is a simple but well known tool for validation. It is defined as

$$\mathrm{BIC}(K) = -2\ell(\hat{\Psi}_K) + d\log(n), \tag{6.3}$$

where $\ell(\hat{\Psi}_K)$ is the maximized log-likelihood of the $K$-component mixture, $d$ is the number of independent parameters in the model and $n$ is the number of judges in the sample. With the BIC we choose the model that has the smallest value, corresponding to a better fit of the model-based clustering model. The parameter $d$ penalizes the unrestricted DBM harder, because it requires an additional $K-1$ parameters to estimate the spread of the densities.

## 6.2 External criterium

External criteria evaluate a clustering by matching the clustering result to a priori information. In the simulation study performed in section 7.1, we fixed the labels of the centers and sampled rankings by Mallows' model in order to generate a heterogeneous population of judges.

Our external validity index is a recovery measure based on Kendall's $\tau_x$ between centers $(c_1^{\mathrm{S}}, \ldots, c_K^{\mathrm{S}})$ fixed in the sampling procedure and the centers estimated by the partitioning algorithm $(c_1^{\mathrm{A}}, \ldots, c_K^{\mathrm{A}})$. However, the algorithms do not precisely know the order in which the clusters were generated. Therefore, the centers estimated by the algorithm are reordered in such a way that the centers are as similar to the centers specified in the population. When calculating $\tau_x$ between the population centers and the centers estimated by the algorithm we get a squared $K \times K$ correlation matrix. We assume that by maximizing the diagonal elements of this matrix the centers estimated by the algorithm closely approximate the centers defined in the population. The reordering of the elements in this matrix works as follows. If we fix the population centers and re-allocate the centers estimated by the algorithm such that the correlations on the diagonal are maximimal. The recovery index is defined by

$$\mathrm{R}(K) = \frac{1}{K}\sum_{k=1}^{K} \tau_x(c_k^{\mathrm{S}}, c_k^{\mathrm{A}}), \tag{6.4}$$

which is the mean of the reorganized diagonal elements of the correlation matrix. We only use the diagonal elements, because the off-diagonal elements depend on the location of the other centers. The recovery has an upper bound of 1 if and only if the algorithm perfectly recovers all labels of the population median rankings. A lower value indicates that at least one cluster is not perfectly recovered.

# 7 Data analysis

In section 3 the various distance measures for ranking data are described. The violation of the triangle inequality by the Kendall distance suggests that it should not be used as a proper distance measure, especially in a distance based model. The adjustment in equation (3.3) by Emond & Mason lead to Emond & Mason's distance. Mallows' $\phi$–model can easily be fitted with Emond & Mason's distance ($d_{\tau_x}$). This distance measure associates proper distances when tied rankings are involved, equivalent to the Kemeny distance. It can also be used with the Kemeny distance with a small modification of the spread parameter. Kendall's maximum distance is 2, independent of the number of objects. This is similar to the transposition of two adjacent objects with the Kemeny distance. The maximum Kemeny distance depends on the number of objects and is $m(m-1)$. Therefore, the product of the spread parameter and Kemeny distance increase. The negative exponential of this product rapidly decreases, resulting in underestimated probabilities. We know that $\tau_x$ is equal to $d_{\text{Kem}}$ for complete and tied rankings and can be solved with equation (3.6) to ensure that the probabilities are equal again.

We have chosen to fit the mixture of DBM in section 5.2 with the Kemeny distance, because it can be applied to both methods and the external evaluation is similar.

## 7.1 Simulation study

To evaluate the recovery of CCA and DBM for both spread parameters described in section 5, a simulation study has been set up to answer our main research question. Artificial data sets have been generated with factors with different levels to mimic real-life situations of a heterogeneous population of judges based on Mallows' model in equation (5.10). The rankings are generated for each cluster independently and combined.

We designed a full factorial experimental design, containing seven factors with two or more levels. The following factors have been manipulated: (a) the number of objects, (b) type of input rankings, (c) correlation between the centers, (d) the number of centers, (e) the sample size, (f) cluster size and (g) spread parameter. We tried to perform a full factorial design, but if the centers were not identifiable given the number of objects, input rankings and correlation between centers it was left out. A description of each factor is described next. When multiplying the levels of the factors, we obtain a total of: $3 \times 2 \times 3 \times 3 \times 3 \times 2 \times 2 \times 3 = 648$ experiments. However, for only 420 experiments the appropriate centers were identified. Each experiment is replicated ten times, analyzing a total of 4,200 data sets. A summary of the manipulated factors is given in Table 8.

Table 8: Factors and levels in the simulation study.

| Factor | Levels |
|---|---|
| Objects | Four, five and seven objects. |
| Input rankings | Complete rankings, complete and tied rankings. |
| Correlation | Uncorrelated, positive and negative. |
| Clusters | Two, three and four clusters. |
| Sample size | 300 and 1,500 judges. |
| Group size | Equal and unequal. |
| Spread | Low, high and varying. |

### 7.1.1 Description of the factors

**Number of objects** As we have seen, the number of objects is crucial. It defines the sample space and the number of rankings rapidly increases. The simulation studies in Lee & Yu (2012) and Murphy & Martin (2003) applied their methods to four and five objects, respectively. We also look at seven objects, so this factor has three levels.

**Input rankings** Not only is the number of objects important, but distinguishing between complete rankings and tied rankings may also be of interest. The number of tied rankings given any number of objects exceeds the number of complete rankings. As we have seen, tied rankings are located at the intersection of complete rankings. Therefore, tied rankings are located closer when the center is a complete ranking. Here, we distinguish between input space. With a tied ranking as center, we can never recover it by sampling complete rankings as input rankings. When dealing with complete rankings the centers are complete rankings and with tied rankings it can be a combination of both.

**Number of clusters** The number of clusters that can be discovered in a population of judges is simulated by this factor. Here, we distinguish a population of judges that can be decomposed into two, three and four clusters. If the number of clusters increase, we expect that it becomes more difficult to correctly recover them.

**Correlation between centers** We distinguish three different locations between the cluster centers namely negatively, uncorrelated and positively correlated centers. The first center in any simulation study is always the full ordering $<1 \cdots m>$ with $m$ objects is the reference ranking. Uncorrelated centers are located halfway the reference ranking and its reversal, simply dividing the maximum distance by two. Positively correlated cluster centers are located closer, whereas negatively correlated cluster centers are located further away from the reference ranking. When dealing with tied rankings, if possible, the all-ties ranking is not chosen as center.

In this simulation study all centers are equidistant, meaning that all off-diagonal elements in the distance matrix are similar. With three clusters, we can identify an equilateral triangle which satisfies the property that the distance between the centers is equal. Let $A$, $B$ and $C$ denote these centers, so: $d(A,C) = d(B,C) = d(A,B)$ in Figure 3. When moving from three to four centers, this is done similarly. In Figure 4 a rhombus with four centers $A$, $B$, $C$ and $D$ is given. This rhombus is a combination of four equilateral triangles, where:

$$\begin{aligned}
&\text{Triangle } \alpha \ (ABD) : d(A,B) = d(B,D) = d(A,D),\\
&\text{Triangle } \beta \ (DBC) : d(D,B) = d(B,C) = d(D,C),\\
&\text{Triangle } \gamma \ (ACD) : d(A,C) = d(C,D) = d(A,D) \text{ and}\\
&\text{Triangle } \delta \ (ACB) : d(A,C) = d(C,B) = d(A,B).
\end{aligned}$$

Adding triangle $\beta$ to triangle $\alpha$ we get: $d(A,B) = d(B,D) = d(A,D) = d(B,C) = d(D,C)$. Then adding triangle $\gamma$ to $\alpha$ and $\beta$ we get: $d(A,B) = d(B,D) = d(A,D) = d(B,C) = d(D,C) = d(A,C)$. Finally, adding triangle $\delta$ to the previous three, to conclude that: $d(A,B) = d(B,D) = d(A,D) = d(B,C) = d(D,C) = d(A,C) = d(A,C)$. The distance from one center to any other is equal. By combining multiple equilateral triangles it is possible to identify situations with more than four equidistant centers.
To form equilateral triangles on the polytope between complete rankings we can only use even distances. The correlations between the centers and the reference ranking are chosen such that $\tau_x$ comes close to -0.50 and 0.50. Given the distances associated

Table 9: Distance $d_{\text{Kem}}$ and correlation ($\tau_x$) between centers.

| | Objects | | |
|---|---|---|---|
| Correlation | 4 | 5 | 7 |
| Negative | 8 (-0.33) | 14 (-0.40) | 30 (-0.43) |
| Uncorrelated | 6 (0.00) | 10 (0.00) | 20 (0.05) |
| Positive | 4 (0.33) | 6 (0.40) | 12 (0.43) |

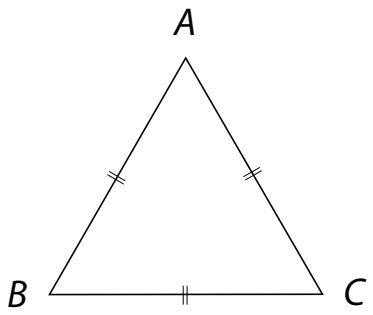



Figure 3: Equilateral triangle between three centers.

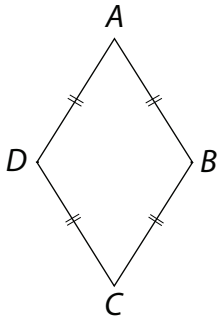



Figure 4: Rhombus between four centers.

with the number of objects it is not possible to identify centers with these specific correlations. The correlation between clusters is given in Table 9. The identified cluster centers $(c_1^{\text{S}}, \ldots, c_K^{\text{S}})$ associated rank vectors are listed in Table 10. The correlation between centers can be seen as the location parameter between clusters, wherein judges rank objects more similar. The location means that the difference between the judges that agree to the reference ranking rank objects more similar compared to the negatively correlated centers. We expect a better recovery when the centers are more apart from each other.

**Sample size** We consider two sample sizes for the entire population of judges, a small sample containing 300 judges and a large sample containing in total 1,500 judges. We expect that the larger sample size leads to a better recovery.

**Cluster size** This factor determines the size and probability of being represented in the population of judges. We distinguish between balanced and unbalanced cluster sizes. With equal cluster size, it is determined by the entire sample size (see previous factor) divided by the number of clusters: $n_1, \ldots, n_K = \frac{n}{K}$. If we deal with unbalanced cluster sizes, the size of the first cluster is taken to be twice as large as the remaining clusters. The size of the first cluster is given by: $n_1 = \frac{2n}{K+1}$ and the remaining $K-1$ clusters: $n_2, \ldots, n_K = \frac{n}{K+1}$.

**Spread parameter** Lastly, the spread parameter controls the peakedness of the density and can be seen as the within group parameter. It is controlled by $\lambda$ in Mallows' model. If the spread parameter is zero, then all rankings are uniformly distributed over the sample space. When the spread parameter increases the probabilities around the center increase. With this factor we distinguish three levels for $\lambda$. A low level for $\lambda$ is fixed at 0.3, so $\lambda_1, \ldots, \lambda_K = 0.3$ for all centers. The high level is fixed at 0.5, so $\lambda_1, \ldots, \lambda_K = 0.5$. To vary the levels, values for $\lambda$ alternate between 0.5 and 0.3 for up to $K$ clusters. With three centers the vector for $\lambda$ is: $\lambda_1 = 0.5, \lambda_2 = 0.3$ and $\lambda_3 = 0.5$ and with four centers the vector for $\lambda$ is: $\lambda_1 = 0.5, \lambda_2 = 0.3, \lambda_3 = 0.5$ and $\lambda_4 = 0.3$.

Table 10: Median rankings defined in the experimental design.

| | | 4 objects | | 5 objects | | 7 objects | |
|---|---|---|---|---|---|---|---|
| | | Complete | Tied | Complete | Tied | Complete | Tied |
| Cor [1] | $K$ | $a\ b\ c\ d$ | $a\ b\ c\ d$ | $a\ b\ c\ d\ e$ | $a\ b\ c\ d\ e$ | $a\ b\ c\ d\ e\ f\ g$ | $a\ b\ c\ d\ e\ f\ g$ |
| Neg | 2 | (1 2 3 4)<br>(3 4 1 2) | (1 2 3 4)<br>(2 1 2 1) | (1 2 3 4 5)<br>(3 4 5 2 1) | (1 2 3 4 5)<br>(3 3 1 2 2) | (1 2 3 4 5 6 7)<br>(7 6 5 1 2 3 4) | (1 2 3 4 5 6 7)<br>(5 4 2 4 1 3 3) |
| | 3 | (1 2 3 4)<br>(3 4 1 2)<br>(3 2 4 1) | (1 2 3 4)<br>(3 4 1 2)<br>(3 2 4 1) | | | | |
| Unc | 2 | (1 2 3 4)<br>(2 4 1 3) | (1 2 3 4)<br>(1 2 2 1) | (1 2 3 4 5)<br>(2 5 3 1 4) | (1 2 3 4 5)<br>(2 3 1 3 2) | (1 2 3 4 5 6 7)<br>(4 3 6 5 1 2 7) | (1 2 3 4 5 6 7)<br>(4 2 3 2 4 1 5) |
| | 3 | | (1 2 3 4)<br>(1 2 2 1)<br>(2 4 1 3) | | (1 2 3 4 5)<br>(2 3 1 3 2)<br>(2 5 3 1 4) | (1 2 3 4 5 6 7)<br>(4 3 6 5 1 2 7)<br>(7 4 1 3 2 5 6) | (1 2 3 4 5 6 7)<br>(4 2 3 2 4 1 5)<br>(7 4 1 3 2 5 6) |
| | 4 | | | | (1 2 3 4 5)<br>(2 3 1 3 2)<br>(2 5 3 1 4)<br>(2 1 1 1 2) | (1 2 3 4 5 6 7)<br>(4 3 6 5 1 2 7)<br>(7 4 1 3 2 5 6)<br>(6 2 3 4 7 1 5) | (1 2 3 4 5 6 7)<br>(4 2 3 2 4 1 5)<br>(7 4 1 3 2 5 6)<br>(5 2 7 1 3 6 4) |
| Pos | 2 | (1 2 3 4)<br>(2 3 1 4) | (1 2 3 4)<br>(1 2 1 2) | (1 2 3 4 5)<br>(2 4 1 3 5) | (1 2 3 4 5)<br>(2 1 1 3 3) | (1 2 3 4 5 6 7)<br>(3 2 1 6 4 7 5) | (1 2 3 4 5 6 7)<br>(2 4 3 4 1 5 5) |
| | 3 | (1 2 3 4)<br>(2 3 1 4)<br>(1 4 2 3) | (1 2 3 4)<br>(2 3 1 4)<br>(1 4 2 3) | | (1 2 3 4 5)<br>(2 1 1 3 3)<br>(2 4 1 3 5) | (1 2 3 4 5 6 7)<br>(3 2 1 6 4 7 5)<br>(3 1 5 2 7 4 6) | (1 2 3 4 5 6 7)<br>(2 4 3 4 1 5 5)<br>(5 2 1 4 3 6 7) |
| | 4 | | | | (1 2 3 4 5)<br>(2 1 1 3 3)<br>(2 4 1 3 5)<br>(1 1 1 1 2) | (1 2 3 4 5 6 7)<br>(3 2 1 6 4 7 5)<br>(3 1 5 2 7 4 6)<br>(3 2 1 4 7 6 5) | (1 2 3 4 5 6 7)<br>(2 4 3 4 1 5 5)<br>(5 2 1 4 3 6 7)<br>(3 5 1 2 4 7 6) |

[1] Cor is the abbreviation for correlation between centers, where 'Neg' stands for negatively correlated centers, 'Unc' stands for uncorrelated centers and 'Pos' stands for positively correlated centers.

Figure 5 shows the density of Mallows' model with four objects and the reference ranking as its center. The upper figure shows the density considering only complete rankings, whereas the lower figure shows both complete and tied rankings. The two lines represent the high and low value of the spread parameter. If we increase the value of the spread parameter, the probability of observing the cluster center also increases. Therefore, we expect a better recovery when the value of the spread parameter is high.

**Sampling ranking data** In Table 11 a description is given about the sampling procedure that was applied for generating the heterogeneous population of judges.

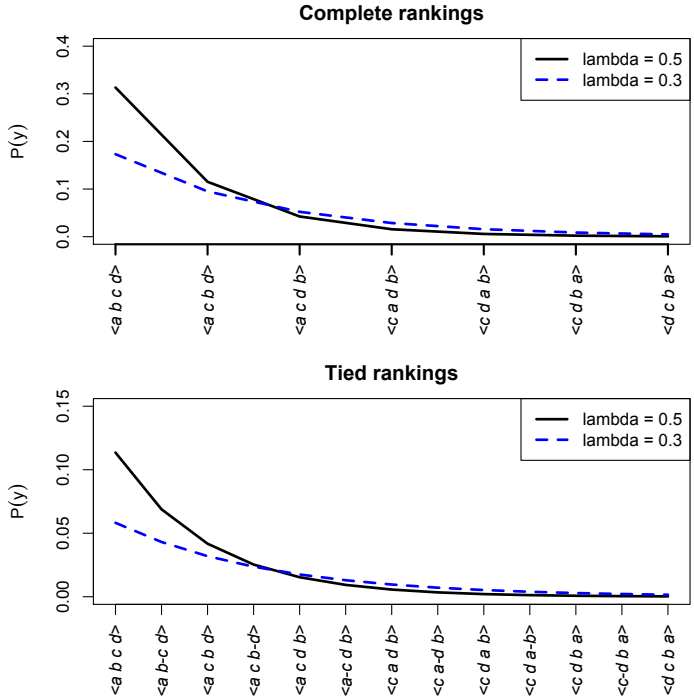



Figure 5: Observed probabilities under Mallows' model.

Table 11: Sampling ranking data.

| Step | Procedure |
|---|---|
| 1 | With $m$ objects generate the input rankings. |
| 2 | Given the input rankings and centers, calculate the distance matrix. |
| 3 | Estimate the probabilities with $\lambda$ of these rankings. |
| 4 | Calculate the size for each cluster $n_k$. |
| 5 | Sample rankings for each cluster by the probability of that ranking. |
| 6 | Combine all sampled rankings to obtain the data of the population. |

### 7.1.2 Effect of the factors

It is not possible to identify more than two negatively correlated centers, except with four objects. Similarly, with five objects it is not possible to identify uncorrelated and positively correlated cluster centers for complete rankings. The 35 identified centers in Table 10 are evaluated with the manipulations for sample and cluster size and the spread parameter. That is why 420 out of 648 experiments have been performed. The most global results are given in Appendix A, where each cell is the mean recovery of ten replications.

Analyzing the results based with a complete ANOVA would not be appropriate because the underlying assumptions, where $e_i \overset{iid}{\sim} N(0, \sigma_e^2)$ would be violated. There is no constant variance and many recoveries are close to one. After identifying the centers, the design of this experiment became highly unbalanced. Therefore we analyze the mean recovery and standard deviation per factor level averaged out over the other factors. The results are given in Table 12. Lets look at the recovery of each factor in turn.

Table 12: Marginal mean and (standard error) recovery.

| Factor | Levels | Method | | | | | |
|---|---|---|---|---|---|---|---|
| | | CCA | | Restricted DBM | | Unrestricted DBM | |
| Objects | 4 | 0.95 | (0.08) | 0.79 | (0.22) | 0.69 | (0.23) |
| | 5 | 0.95 | (0.08) | **0.83** | **(0.19)** | 0.73 | (0.21) |
| | 7 | **0.98** | **(0.05)** | 0.80 | (0.20) | **0.74** | **(0.19)** |
| Input rankings | Complete rankings | **0.99** | **(0.03)** | **0.90** | **(0.16)** | **0.78** | **(0.21)** |
| | Tied rankings | 0.94 | (0.09) | 0.73 | (0.20) | 0.68 | (0.20) |
| Correlation | Negative | **0.98** | **(0.06)** | 0.79 | (0.28) | 0.68 | (0.29) |
| | Uncorrelated | 0.97 | (0.07) | 0.77 | (0.21) | 0.68 | (0.21) |
| | Positive | 0.94 | (0.08) | **0.84** | **(0.12)** | **0.78** | **(0.12)** |
| Clusters | 2 | **0.98** | **(0.04)** | **0.87** | **(0.20)** | **0.78** | **(0.22)** |
| | 3 | 0.94 | (0.09) | 0.75 | (0.19) | 0.67 | (0.18) |
| | 4 | 0.92 | (0.09) | 0.71 | (0.15) | 0.64 | (0.16) |
| Sample size | Small | 0.95 | (0.08) | **0.81** | **(0.20)** | 0.71 | (0.20) |
| | Large | **0.97** | **(0.06)** | 0.80 | (0.21) | **0.73** | **(0.21)** |
| Cluster size | Equal | **0.97** | **(0.06)** | **0.83** | **(0.21)** | **0.74** | **(0.22)** |
| | Unequal | 0.95 | (0.08) | 0.78 | (0.19) | 0.70 | (0.19) |
| Spread | Low | 0.93 | (0.09) | 0.74 | (0.21) | 0.61 | (0.17) |
| | Varying | 0.96 | (0.08) | 0.80 | (0.19) | 0.72 | (0.20) |
| | High | **0.99** | **(0.04)** | **0.86** | **(0.18)** | **0.83** | **(0.20)** |

**Objects** The best recovery is obtained with seven objects for CCA, followed by the unrestricted DBM. The best recovery with the restricted DBM is with five objects. The highest recovery is associated with the smallest standard error. The worst recovery for all models is with four objects. An explanation could be that the sample size is limited and the recovery quickly decreases by a small deviation from the population centers. For example, if the estimated median ranking is only a transposition of two adjacent objects wrong with four objects this already leads to a decrease of 0.33 with Emond & Mason's $\tau_x$.

**Input rankings** The average recovery in Table 12 clearly indicates that all methods recover the centers better with complete rankings. It shows the highest recovery for CCA and restricted DBM for complete rankings. Given the design of our simulation study, we only identified centers listed in Table 10. It shows that there are more situations possible with tied rankings than with complete rankings. Recall Figure 5, where the probability of observing the reference ranking with tied rankings is much smaller than with complete rankings.

The boxplots in Figure 6 show the variation between the number of objects and input rankings. For CCA there is no variation with complete rankings, but the recovery increases and variation decreases when the number of objects increase with tied rankings. When looking at the DBM models and complete rankings, the picture is much more diverse; the best recovery is with five objects and four and seven objects it is less.

**Correlation** This factor reveals an unexpected trend. We expected to see better recovery with negatively correlated centers, because they are further away from the reference ranking. Therefore, it is easier for the algorithms to recover them. This is indeed what happened with CCA. With DBM the reverse seemed to happen, where the best recovery is obtained by positively correlated centers. An explanation of this could be found in the range of sampled rankings. The range of sampled rankings is much wider when considering negatively correlated centers.

The thick black line in Figure 7 is the same in Figure 5, where the density of Mallows' model is printed with four objects, complete rankings and a high value for $\lambda$. In this figure, the density of a second cluster with negatively correlated center at $<c\ d\ a\ b>$ is given by the dashed line. The dotted line is the average of these two densities, to show that the sampling procedure in Table 11 does not lead to unintended situations. However, the sampling of rankings in step 5 may result in sampling the ordering $<a\ c\ d\ b>$ and rankings that overlap for different clusters. Another explanation could be that the rankings in between cluster centers are sampled more often, resulting in a local maximum for the DBM.

The variation between the correlation between centers and input rankings in Figure 8 shows that with CCA for complete rankings there is no variation but almost perfect recovery. The variation in recovery with tied rankings increase when centers are located closer together and the median recovery with positive correlation is not 1 anymore. With the DBM models the median recovery with negative correlation and complete rankings is 1, but with tied rankings it comes close to 0.6. With complete rankings and uncorrelated centers the median recovery of the restricted DBM is one, whereas for the unrestricted DBM it is much lower, around 0.7. With tied rankings the recovery of the DBM for uncorrelated and positively correlated centers is somewhat smaller.

**Clusters** We expected to see a decrease in recovery when the number of clusters increase. This is indeed what has happened. With two centers it returns the second best recovery of all averaged medians in Table 12, with only the high value of lambda and the input of only complete rankings in advance. Even though the recovery may be better, the standard errors of the DBM's indicated that there is more variation than with more clusters.

**Sample size** This factor is not of much influence. Our expectation that recovery increases with sample size does not show noticeable differences on the recovery of the methods. If we look at the average recovery over ten replications in Appendix A, there are small improvements with CCA if the sample size is increased. The DBM on the other hand shows more variation between sample sizes.

**Cluster size** All methods prefer balanced clusters. When fixing the first cluster size to be twice as large as the remaining clusters it leads to a small decrease in recovery, but the standard error is smaller with unequal cluster size for the DBM's.

**Spread** The high value for lambda in our data generation model leads to the best recovery for CCA. The unrestricted DBM and with low values for the spread parameter lambda lead to the second worst and worst recovery, respectively. We expected that the unrestricted spread parameter in equation (5.17) would show better recovery compared to its restricted counterpart. The mean recovery however suggests otherwise. With 0.72 the recovery of the unrestricted DBM is worse than the restricted DBM with 0.80.

When looking at the boxplots in Figure 9 where the variation between spread and input rankings are displayed, CCA shows a perfect median recovery of the centers with complete rankings, irrespective of the value of the spread parameter. The recovery increases and the variation decreases when the spread parameter increases with tied rankings. In addition, the restricted DBM has a perfect median recovery with complete rankings. However, there is a big difference with tied rankings. Even with high values for lambda, the median recovery is around 0.80 and gets lower if there is more variation within a cluster. The unrestricted DBM only attains perfect recovery with complete rankings and a high value of the spread parameter. With lower values of lambda and complete rankings, the recovery decreases more rapidly than with tied rankings.

### 7.1.3 Discussion of the simulation results

Given the results of our simulation experiment, we conclude the following. CCA recovers the median rankings best for each factor for all levels. In addition, CCA has the smallest standard error indicating more stable clustering results. All models tend to prefer more objects, complete rankings, two clusters, a large and balanced sample size and a high value for lambda. The worst model in terms of recovering the population centers is the unrestricted DBM, followed by the restricted distance based model. In general, the better the recovery, the smaller the standard error.

A counterintuitive observation is that the DBM's have more difficulties in recovering cluster centers that are located further away from the reference ranking, instead of centers that are located closer to each other. In addition, the unrestricted DBM is not an improvement over the restricted DBM when it comes to the recovery with unequal values for lambda.

The differences between CCA and the DBM's cannot be easily explained. Both methods estimate the median rankings as cluster centers, spread parameter and mixing probabilities similarly. The difference is in the estimating the membership probabilities. In each iteration the membership probabilities with CCA only depend on the distances of the ranking to the centers, whereas with DBM the densities have to be evaluated. The densities of Mallow's model depend on the centers, as well as the spread parameters and mixing probabilities (Iyigun, 2007, Chapter 6).

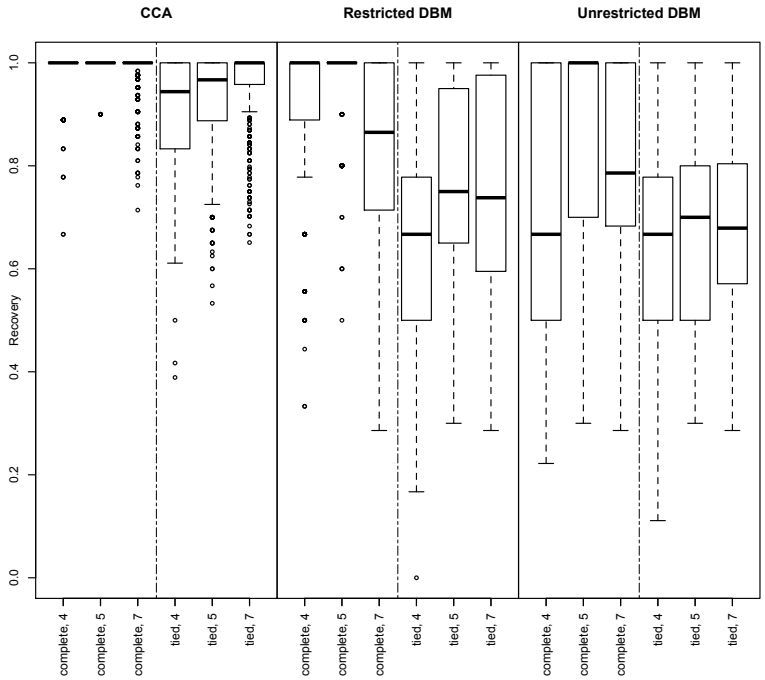



Figure 6: Variation between number of objects and input rankings.

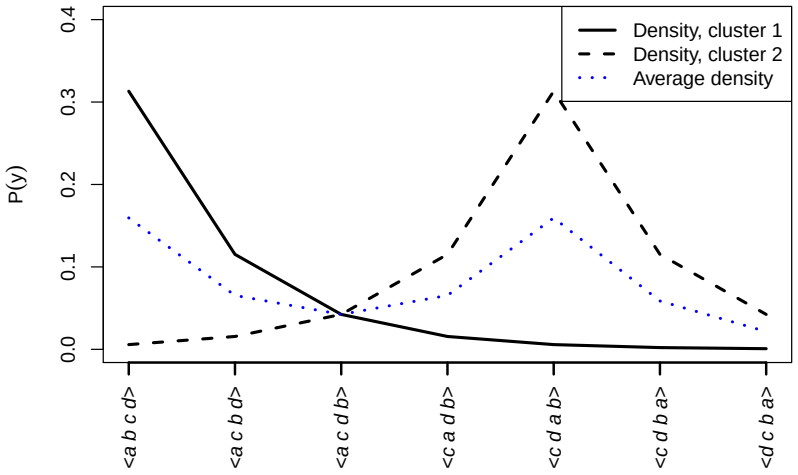



Figure 7: Densities of two negatively correlated centers.

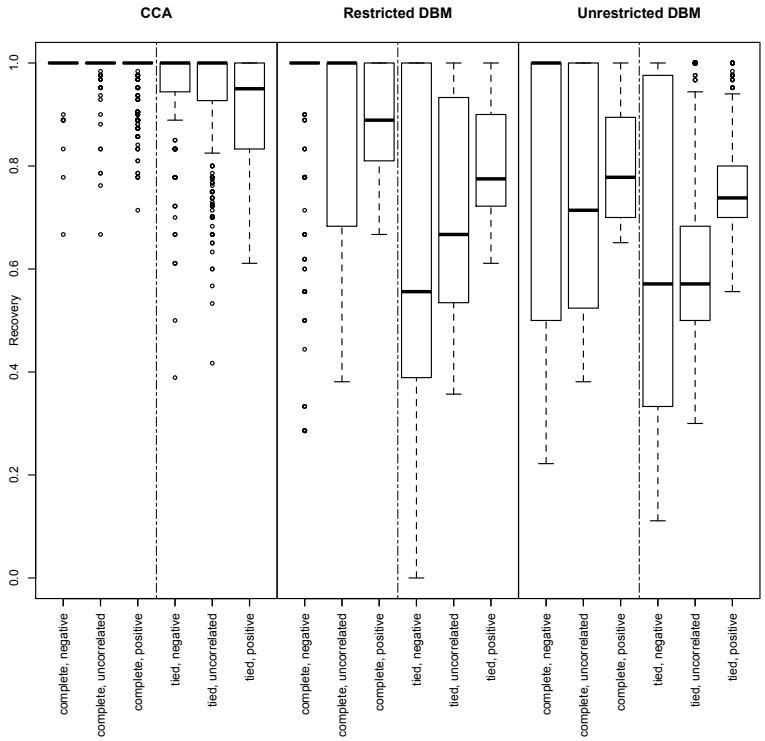



Figure 8: Variation between the correlation between centers and input rankings.

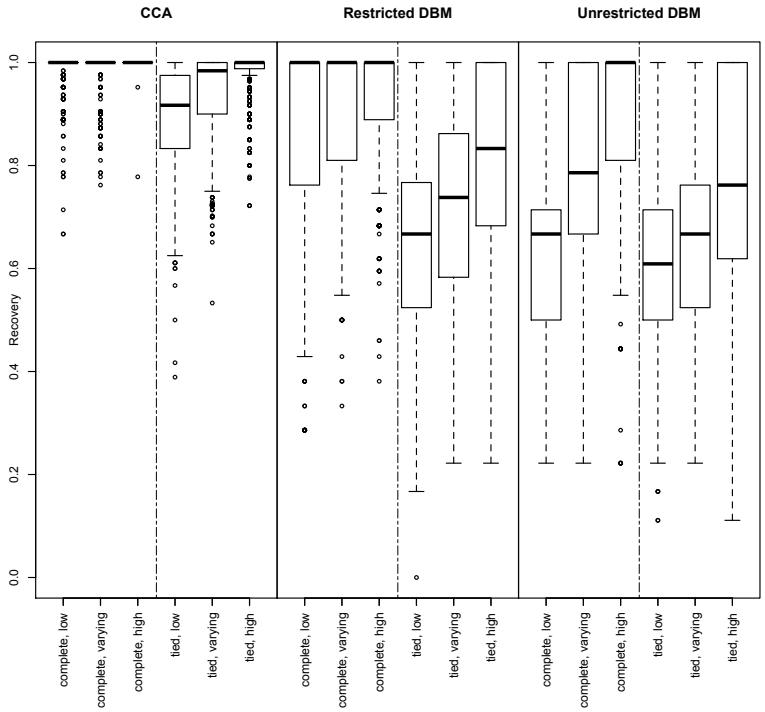



Figure 9: Variation between the spread and input rankings.

## 7.2 Real data applications

The aim of this subsection is to identify which internal measure seems to be able to identify the number of clusters in the data. Eight data sets have been published and analyzed by different authors. They will be analyzed again by the partitioning methods. These data sets are: 'Voting', 'Living places', 'Political goals', 'Song', 'Idea', 'Rice subset', 'APA subset' and 'Sports'. The observed rankings and observed frequencies are given in Appendix B. A description of these data sets is given next.

### 7.2.1 Description of the data sets

**Voting** This data set, described in Plackett (1975, p. 197) and Marden (1996, p. 30), is about the order of candidates $a$, $b$ and $c$ that appeared on the ballot. These ballots have been disseminated in six areas. The data are the aggregated 908 votes of these areas.

**Living places** This data set is described in Diaconis (1988, p. 92). It is a question of a larger questionnaire from a survey of the National Opinion Research Center about where people want to live. The choicer were: $a$ = in a big city, $b$ = near a big city (suburbs; $\leq 50$ miles) and $c$ = far from a big city ($\geq 50$ miles). In total 1,493 people answered this question.

**Political goals** This data set comes from Croon (1989, p. 111) and has been analyzed in Lee & Yu (2012) and describes "Changing mass publics". A subset of $N = 2,262$ German respondents were asked to rank the following four political goals according to their desirability: $a$ = Maintain order in the nation, $b$ = Give people more say in the decision of the government, $c$ = Fight rising prices and $d$ = Protect freedom of speech. Of the four alternatives there is a distinction between materialistic and post-materialistic value orientation. The first entails social and economic stability and security and goals like $a$ and $c$, whereas $b$ and $d$ agree more to humane and spiritual aspects of life.

**Song** This data set comes from Critchlow et al. (1991, p. 313) and is about the ranking of five words according to strength of association with a target word "song". The five choices to rank were: $a$ = score, $b$ = instrument, $c$ = solo, $d$ = benediction and $e$ = suit. In this study 83 university students participated.

**Idea** This data set from Fligner & Verducci (1986, p. 364) has a similar structure as the 'Song' data set. It is about the ranking of five words according to strength of association with a target word "idea". The five choices were: $a$ = though, $b$ = play, $c$ = theory, $d$ = dream, and $e$ = attention. In total 98 university students participated.

**Rice subset** This data set comes from Baggerly (1995, p. 105–106) and the rankings comprise the ballots in a preferential election to choose a faculty member to serve on the Rice Presidential Search Committee. In total 300 people casted their vote. However, only a total of 210 people completely listed their preferences of five persons: $a, b, d$ and $e$ with the remaining 90 rankings incomplete.

**APA subset** This data set comes from Diaconis (1988, p. 96) and is a subset where 15,449 psychologists were asked to rank five candidates for the 1980 American Psychological Association (APA). It has been analyzed in Diaconis (1988), Marden (1996, p. 37), Murphy & Martin (2003) and Busse et al. (2007). The candidates are $a$ = William Bevan, $b$ = Ira Iscoe, $c$ = Charles Kiesler, $d$ = Max Siegle and $e$ = Logan Wriths. Furthermore, candidates $a$ and $c$ are research psychologists, $d$ and $e$ are clinical psychologists and $b$ is a community psychologist. From all rankings, we only analyzed the complete rankings, which is a subset of 5,738 rankings (only 37%).

**Sports** The 'Sports' data set comes from Louis Roussos and is described in Marden (1996). He asked 130 students at the University of Illinois to rank seven sports according to their preference of participating in. The sports to choose from were: $a$ = baseball, $b$ = football, $c$ = basketball, $d$ = tennis, $e$ = cycling, $f$ = swimming and $g$ = jogging.

### 7.2.2 Clustering outcomes

When summarizing these data sets, the following summary statistics of the entire sample with the Kemeny distance are listed in Table 13. For some data sets the median ordering coincides with the modal ordering. The mean ranking on the other hand, tends to prefer tied rankings. Lets take a closer look at the outcomes of each data set separately. The outcomes of the internal validity measures are given in Table 14. Figure 10 shows the JDF of the entire sample of judges of the data sets. In Tables 15 to 17 the outcomes of the best fitting models are given, ordered by the probability of belonging to the population.

Table 13: Summary statistics of the data sets.

| Data set | Modal ordering | Median ordering | Mean ordering |
|---|---|---|---|
| Voting | $<a\ b\ c>$ | $<a\ b\ c>$ | $<a–b–c>$ |
| Living places | $<c\ a\ b>$ | $<c\ a\ b>$ | $<c–a\ b>$ |
| Political goals | $<b\ c\ a\ d>$ | $<a\ b\ c\ d>$ | $<a–b\ c–d>$ |
| Song | $<c\ b\ a\ d\ e>$ | $<c\ b\ a\ d\ e>$ | $<c\ b\ a\ d\ e>$ |
| Idea | $<b\ e\ d\ c\ a>$ | $<b\ e\ d\ c\ a>$ | $<b\ e\ d–c\ a>$ |
| Rice subset | $<a\ c\ b\ e\ d>$ | $<a\ b\ c\ e\ d>$ | $<a–b\ c\ e\ d>$ |
| APA subset | $<c\ a\ b\ e\ d>$ | $<a\ c\ e\ d\ b>$ | $<a–b–c–d–e>$ |
| Sports | $<g\ e\ f\ d\ c\ a\ b>$ | $<e\ f\ c\ a\ d\ b\ g>$ | $<e\ c\ a\ f\ d\ b\ g>$ |

**Voting** The modal and median ranking are the same, listing the candidates $<a\ b\ c>$. The mean ranking is of little use, it the all ties ranking of the three candidates. When partitioning this data, the internal validity measures are give in Table 14. With CCA applied, the best result is with six clusters. This is the complete crisp partition of the membership matrix, indicated by a JDF of zero. The mixing probability is about the inverse of its frequency. The BIC of both DBM's suggest a two component mixture where they both estimate the largest cluster as the modal ranking of the entire sample. The second cluster is the ordering $<a\ c\ b>$ with a small probability of being represented for the restricted DBM, whereas the second cluster of the unrestricted DBM has zero probability of being present in the data.

**Living places** The modal and median ranking are the same namely $<c\ a\ b>$, the mean ranking is a tied ranking of the first two objects $c$ and $a$ of the modal ranking. When partitioning this data set we obtain internal validity measures similar to the partition of the Voting data set. The PC and PE indicate a six component mixture with CCA. If we look for a knee with the JDF, it is located at four centers in Figure 10. When comparing the outcomes of the four and six cluster results, then the four clusters are estimated with similar probabilities and spread parameter and the estimated centers coincide with the observed frequencies. The additional two clusters do not account for more than 3% of the entire sample. The BIC indicates a two component mixture for both DBM's. They estimate the same centers with the first center equal to $<c\ a\ b>$ and the second center as $<a\ b\ c>$. These centers correspond with the modal ranking and the

third most observed ranking. The unrestricted DBM has the smallest BIC indicating a better fit to the data.

**Political goals** The modal ordering is $<b\ c\ a\ d>$, the median ordering is different, namely $<a\ b\ c\ d>$. The mean ranking places the first and last two objects as tied. When we partition the data the PC indicates that six clusters fits the data best and the PE with two clusters.

This data set has been analyzed in Lee & Yu (2012). They suggest that the best fitting model contains a three clusters with orderings: $<a\ c\ b\ d>$ with 0.441 probability, $<c\ a\ b\ d>$ with 0.352 probability and $<b\ d\ c\ a>$ with 0.208 probability. In the obtained clustering there is a distinction between the value orientations. Croon (1989) also fitted a different kind of model to the same data and also identified that a three component mixture would fit this data best. However, none of the methods identify the same three cluster outcome.

The PE for CCA identifies a two cluster outcome, where the ordering of the centers are $<c\ a\ b\ d>$ and $<a\ d\ c\ b>$. The largest estimated center corresponds with the second largest center. The six cluster outcome has the first center estimated as in Lee & Yu.

The BIC indicates that a four components mixture should be fitted with the unrestricted DBM and a two component mixture with the unrestricted DBM. There is a large difference between the BIC's in favour of the unrestricted DBM. With this model, the two estimated centers coincide with the median ranking with probabilities of one and zero. The restricted DBM estimates the largest center $<a\ c\ b\ d>$ with an inverse of values $d$ and $b$ compared to the four cluster CCA outcome. The second centers are similar and the remaining two centers are reversed.

**Song** For this particular data set the modal, median and mean ranking are the same and the ordering is $<c\ b\ a\ d\ e>$. Clearly, the word 'solo' is associated by the word song. When partitioning this data all indices identify that a two cluster outcome fits the data best. The BIC is slightly smaller in favour of the unrestricted DBM. Both DBM's estimate the same cluster centers with very similar spread parameters and mixing probabilities. The estimated centers correspond with the two most observed rankings. CCA's validity measures also indicate a two cluster outcome. The largest estimated cluster center is similar to the median ranking. The estimated center of the second cluster is the ordering $<b\ a\ c\ d\ e>$ with probability 0.380. This center has the first three objects reversed, compared to the center of the DBM model.

**Idea** The modal and median ranking are equal and the ordering is $<b\ e\ d\ c\ a>$. The mean ranking ranks the words $d$ and $c$ of the modal ranking as tied. This data set has not been partitioned before. With CCA the PE identifies the two cluster solution as best fitting model, whereas the PC suggests six clusters. The largest cluster of these models is the median ranking of all students and the second largest cluster is the inverse of the words $a$ and $c$. From the remaining four centers, the first is the inverse of words $b$ and $e$, the second center is the inverse of words $d$ and $e$, the third cluster the inverse of words $a$ and $c$ of the median ranking and the final ranking is with only 2% mixing probability entirely different. Of these four centers, three are small deviations from the median ranking. The BIC for the DBM's identifies a two component mixture with centers equal to the two cluster outcome with PE.

**Rice subset** The modal ranking is $<a\ c\ b\ e\ d>$. The median ranking is the inverse of candidates $c$ and $b$ of the modal ranking. The mean ranking places the candidates $a$ and $b$ of the median ranking as tied. When partitioning this data, we obtain for all methods a best fitting solution with two clusters. The largest cluster has the ordering $<b\ a\ c\ e\ d>$ and is the second most observed ranking in the data. The second largest

estimated cluster center is $<c\ a\ b\ e\ d>$, where the first three candidates ($b$, $a$ and $c$) are reversed and this ranking is the third most observed ranking. The mixing probabilities of CCA and the restricted DBM are about the same. The unrestricted DBM has a much lower BIC and estimates the mixing probability of the first clusters much higher than the other two. The BIC indicates that the unrestricted DBM fits the data better.

**APA subset** The three summary rankings are completely different. The modal ordering is $<c\ a\ b\ e\ d>$, the median ordering is $<a\ c\ e\ d\ b>$ and the mean ranking is the all-ties ranking of the psychologists. The partitioning of all models suggest that a two cluster solution fits the data best. Marden (1996, p. 36-37) has analyzed the same data with an adjusted K-Means clustering algorithm to ranking data based on Kendall's distance. Even though we showed that Kendall's distance violates the triangle inequality, the estimated centers and the mixing probabilities are exactly the same with respect to the CCA algorithm, namely the orderings: $<c\ a\ b\ e\ d>$ and $<d\ e\ b\ a\ c>$ with mixing probabilities 52% and 48%, respectively. The result is a distinction between the research psychologists $a$ and $c$ and the clinical psychologists $d$ and $e$. The first estimated center by CCA and the restricted DBM is the modal ranking of the entire data set. The second estimated center is the reversal of the modal ranking. The unrestricted DBM on the other hand estimates the median ranking of the entire sample twice with mixing probabilities of one and zero, respectively. Murphy & Martin (2003) also analyzed the subset of the APA data and obtained with another distance measure, the Cayley distance, a five cluster solution with orderings: $<d\ b\ e\ c\ a>$, $<c\ d\ e\ a\ b>$, $<b\ c\ a\ d\ e>$, $<b\ c\ a\ e\ d>$ and $<b\ d\ a\ e\ c>$. This outcome does not reveal the distinction between the specialization of psychologists that well.

**Sports** It is interesting to see that the modal ranking is given by just three students that rank jogging as their most preferred sport of participating in, whereas it is least preferred in the median and mean ranking. The classification in Marden (1996, p. 37) identifies the following centers of the clusters $<e\ f\ c\ a\ d\ b\ g>$ with 53.46 mixing probability and $<a\ b\ c\ d\ e\ f\ g>$ with 46.5% of being represented in the data. The CCA algorithm identifies the second cluster correctly, but estimates a different ranking for the first cluster. The unrestricted DBM identifies the first cluster similar to the student population median, but with mixing probability of zero for the second cluster indicating a homogeneous population of students.

### 7.2.3 Discussion of the real data sets

The internal validity measures with CCA show counterintuitive results for the real data sets: Voting, Living places, Political Goals and Idea. Of the data sets with three objects it tends to look after the crisp partition, overestimating the number of clusters. That is obtained by the maximum possible number of clusters that coincides with the number of unique rankings present in the data set. The number of clusters identified by JDF of the entire sample as validity measure in Figure 10 ranges in between the results of PC and PE. In the case of the data examples with three objects, the JDF identifies less clusters than the crisp partition. In the other cases it identifies more clusters than the PE. For the Political goals and the Idea data set, the PC identifies more clusters than the PE. Therefore, we suggest to use the PE as internal validity criterium. The number of clusters in the real data examples may be small but the outcomes between CCA and the unrestricted DBM are largely the same namely: Idea, Rice subset and APA subset. The unrestricted DBM identifies in half of the data sets the first cluster to have a mixing probability of one and the other cluster to have zero mixing probability.

Table 14: Internal validity measures of real data sets.

| | | CCA | | | DBM | |
|---|---|---|---|---|---|---|
| | | | | | Restricted | Unrestricted |
| Data set | # clusters | PC | PE | JDF | BIC | BIC |
| Voting | 2 | 0.730 | 0.386 | 766.667 | **3,982.80** | **3,410.97** |
| | 3 | 0.745 | 0.424 | 376.618 | 4,181.30 | 4,385.64 |
| | 4 | 0.810 | 0.351 | **166.667** | 4,895.99 | 4,426.85 |
| | 5 | 0.908 | 0.183 | 68.400 | 5,381.67 | 4,708.92 |
| | 6 | **1.000** | **0.000** | 0.000 | 5,403.15 | 4,772.33 |
| Living Places | 2 | 0.848 | 0.226 | 845.500 | **5,273.05** | **5,300.17** |
| | 3 | 0.912 | 0.147 | 219.855 | 5,865.69 | 5,831.90 |
| | 4 | 0.980 | 0.037 | **26.667** | 6,763.15 | 5,889.04 |
| | 5 | 0.994 | 0.013 | 7.200 | 6,783.11 | 6,920.98 |
| | 6 | **1.000** | **0.000** | 0.000 | 7,867.74 | 7,400.78 |
| Political goals | 2 | 0.654 | **0.499** | 4,611.429 | 16,253.75 | **14,124.60** |
| | 3 | 0.599 | 0.672 | 2,440.613 | 16,482.14 | 14,147.77 |
| | 4 | 0.675 | 0.599 | **1,294.702** | **15,483.68** | 14,171.12 |
| | 5 | 0.674 | 0.648 | 907.782 | 15,499.16 | 14,194.29 |
| | 6 | **0.695** | 0.646 | 692.385 | 15,514.62 | 15,177.74 |
| Song | 2 | **0.740** | **0.421** | 96.000 | **529.14** | **533.83** |
| | 3 | 0.494 | 0.856 | 54.590 | 610.28 | 620.43 |
| | 4 | 0.382 | 1.095 | 34.879 | 658.80 | 645.66 |
| | 5 | 0.361 | 1.253 | **24.676** | 675.10 | 695.08 |
| | 6 | 0.308 | 1.491 | 17.874 | 742.14 | 762.09 |
| Idea | 2 | 0.782 | **0.310** | 86.611 | **553.21** | **558.19** |
| | 3 | 0.774 | 0.374 | 47.862 | 606.67 | 608.40 |
| | 4 | 0.790 | 0.388 | **32.714** | 678.59 | 687.71 |
| | 5 | 0.780 | 0.434 | 27.493 | 728.37 | 769.92 |
| | 6 | **0.808** | 0.405 | 20.339 | 767.64 | 779.10 |
| Rice subset | 2 | **0.628** | **0.530** | 525.477 | **1,912.25** | **1,842.04** |
| | 3 | 0.519 | 0.798 | 333.677 | 1,935.12 | 1,858.08 |
| | 4 | 0.506 | 0.913 | **234.310** | 1,945.82 | 1,874.15 |
| | 5 | 0.440 | 1.119 | 190.610 | 2,102.95 | 1,890.19 |
| | 6 | 0.472 | 1.118 | 149.847 | 2,082.50 | 1,948.54 |
| APA subset | 2 | **0.632** | **0.542** | 21,109.000 | **58,319.48** | **54,923.77** |
| | 3 | 0.471 | 0.891 | **13,951.153** | 60,490.18 | 57,014.59 |
| | 4 | 0.381 | 1.147 | 10,328.322 | 61,749.40 | 57,609.80 |
| | 5 | 0.341 | 1.314 | 8,022.973 | 62,591.26 | 57,650.10 |
| | 6 | 0.306 | 1.469 | 6,576.992 | 63,601.73 | 57,504.06 |
| Sports | 2 | **0.616** | **0.566** | 972.019 | **1,541.79** | **1,410.30** |
| | 3 | 0.430 | 0.956 | **655.376** | 1,679.89 | 1,424.91 |
| | 4 | 0.338 | 1.224 | 490.412 | 1,781.34 | 1,439.53 |
| | 5 | 0.290 | 1.415 | 388.986 | 1,813.05 | 1,463.07 |
| | 6 | 0.268 | 1.554 | 318.330 | 1,721.49 | 1,426.59 |

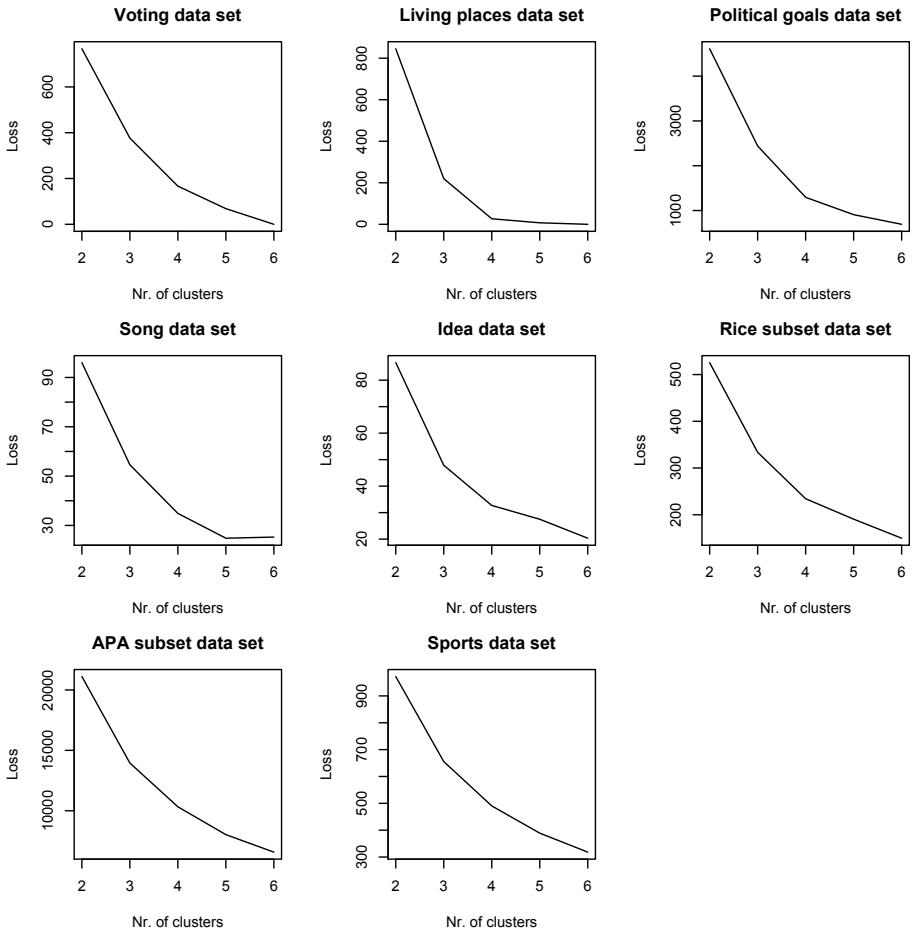



Figure 10: Internal validation measures of the JDF of the entire data set.

Table 15: Best clustering estimates of data sets, 1/3.

| Data set | Method | Index | Center | Spread[1] | $p_k$ |
|---|---|---|---|---|---|
| Voting | CCA | JDF | $<a\ b\ c>$ | 0.813 | 0.313 |
| | | | $<b\ a\ c>$ | 0.764 | 0.248 |
| | | | $<c\ a\ b>$ | 0.744 | 0.227 |
| | | | $<c\ b\ a>$ | 0.724 | 0.213 |
| | | PC & PE | $<a\ b\ c>$ | 1.000 | 0.245 |
| | | | $<b\ a\ c>$ | 1.000 | 0.184 |
| | | | $<c\ a\ b>$ | 1.000 | 0.159 |
| | | | $<c\ b\ a>$ | 1.000 | 0.149 |
| | | | $<a\ c\ b>$ | 1.000 | 0.143 |
| | | | $<b\ c\ a>$ | 1.000 | 0.120 |
| | Restricted DBM | BIC | $<a\ b\ c>$ | 0.119 | 0.908 |
| | | | $<a\ c\ b>$ | 0.119 | 0.092 |
| | Unrestricted DBM | BIC | $<a\ b\ c>$ | 0.112 | 1.000 |
| | | | $<b\ a\ c>$ | 0.652 | 0.000 |

[1] The spread parameters are estimated different, for CCA equation (5.4), for DBM equations (5.16) and (5.17), respectively.

Table 16: Best clustering estimates of data sets, 2/3.

| Data set | Method | Index | Center | Spread[1] | $p_k$ |
|---|---|---|---|---|---|
| Living places | CCA | JDF | <c a b> | 0.986 | 0.444 |
| | | | <c b a> | 0.975 | 0.256 |
| | | | <a b c> | 0.966 | 0.176 |
| | | | <b a c> | 0.950 | 0.124 |
| | | PC & PE | <c a b> | 0.999 | 0.436 |
| | | | <c b a> | 0.998 | 0.249 |
| | | | <a b c> | 0.999 | 0.168 |
| | | | <b a c> | 0.999 | 0.118 |
| | | | <a c b> | 0.976 | 0.019 |
| | | | <b c a> | 0.959 | 0.008 |
| | Restricted DBM | BIC | <c a b> | 0.637 | 0.740 |
| | | | <a b c> | 0.637 | 0.260 |
| | Unrestricted DBM | BIC | <c a b> | 0.651 | 0.733 |
| | | | <a b c> | 0.597 | 0.267 |
| Political goals | CCA | PE | <c a b d> | 0.352 | 0.525 |
| | | | <a d c b> | 0.284 | 0.475 |
| | | JDF | <a c b d> | 0.644 | 0.267 |
| | | | <b c a d> | 0.635 | 0.263 |
| | | | <b c d a> | 0.594 | 0.236 |
| | | | <a d b c> | 0.594 | 0.234 |
| | | PC | <a c b d> | 0.756 | 0.210 |
| | | | <b c a d> | 0.748 | 0.202 |
| | | | <b d a c> | 0.740 | 0.196 |
| | | | <a d b c> | 0.730 | 0.190 |
| | | | <c b a d> | 0.518 | 0.106 |
| | | | <c d a b> | 0.463 | 0.095 |
| | Restricted DBM | BIC | <a c d b> | 0.700 | 0.295 |
| | | | <b c a d> | 0.700 | 0.260 |
| | | | <a d b c> | 0.700 | 0.230 |
| | | | <b d a c> | 0.700 | 0.215 |
| | Unrestricted DBM | BIC | <a b c d> | 0.195 | 1.000 |
| | | | <a b c d> | 0.414 | 0.000 |
| Song | CCA | PC & PE | <c b a d e> | 0.814 | 0.620 |
| | | | <b a c d e> | 0.696 | 0.380 |
| | | JDF | <c b a d e> | 0.904 | 0.314 |
| | | | <c a b d e> | 0.855 | 0.206 |
| | | | <c b d a e> | 0.835 | 0.182 |
| | | | <a b c d e> | 0.813 | 0.158 |
| | | | <b c a d e> | 0.784 | 0.140 |
| | Restricted DBM | BIC | <c b a d e> | 0.779 | 0.685 |
| | | | <c a b d e> | 0.779 | 0.315 |
| | Unrestricted DBM | BIC | <c b a d e> | 0.781 | 0.682 |
| | | | <c a b d e> | 0.775 | 0.318 |

[1] The spread parameters are estimated different, for CCA equation (5.4), for DBM equations (5.16) and (5.17), respectively.

Table 17: Best clustering estimates of data sets, 3/3.

| Data set | Method | Index | Center | Spread[1] | $p_k$ |
|---|---|---|---|---|---|
| Idea | CCA | PE | $<b\ e\ d\ c\ a>$ | 0.852 | 0.595 |
| | | | $<b\ e\ c\ d\ a>$ | 0.783 | 0.405 |
| | | JDF | $<b\ e\ d\ c\ a>$ | 0.918 | 0.410 |
| | | | $<b\ e\ c\ d\ a>$ | 0.877 | 0.273 |
| | | | $<e\ b\ d\ c\ a>$ | 0.834 | 0.199 |
| | | | $<b\ d\ e\ c\ a>$ | 0.716 | 0.119 |
| | | PC | $<b\ e\ d\ c\ a>$ | 0.944 | 0.377 |
| | | | $<b\ e\ c\ d\ a>$ | 0.916 | 0.246 |
| | | | $<e\ b\ d\ c\ a>$ | 0.879 | 0.174 |
| | | | $<b\ d\ e\ c\ a>$ | 0.791 | 0.097 |
| | | | $<b\ e\ d\ a\ c>$ | 0.752 | 0.084 |
| | | | $<a\ b\ c\ d\ e>$ | 0.045 | 0.022 |
| | Restricted DBM | BIC | $<b\ e\ d\ c\ a>$ | 0.826 | 0.670 |
| | | | $<b\ e\ c\ d\ a>$ | 0.826 | 0.330 |
| | Unrestricted DBM | BIC | $<b\ e\ d\ c\ a>$ | 0.852 | 0.643 |
| | | | $<b\ e\ c\ d\ a>$ | 0.779 | 0.357 |
| Rice subset | CCA | PC & PE | $<b\ a\ c\ e\ d>$ | 0.507 | 0.507 |
| | | | $<c\ a\ b\ e\ d>$ | 0.493 | 0.493 |
| | | JDF | $<b\ a\ c\ e\ d>$ | 0.604 | 0.281 |
| | | | $<a\ c\ b\ e\ d>$ | 0.590 | 0.271 |
| | | | $<c\ a\ b\ e\ d>$ | 0.576 | 0.262 |
| | | | $<b\ a\ d\ e\ c>$ | 0.400 | 0.185 |
| | Restricted DBM | BIC | $<b\ a\ c\ e\ d>$ | 0.547 | 0.549 |
| | | | $<c\ a\ b\ e\ d>$ | 0.547 | 0.451 |
| | Unrestricted DBM | BIC | $<b\ a\ c\ e\ d>$ | 0.441 | 0.759 |
| | | | $<c\ a\ b\ e\ d>$ | 0.771 | 0.241 |
| APA subset | CCA | PC & PE | $<c\ a\ b\ e\ d>$ | 0.296 | 0.523 |
| | | | $<d\ e\ b\ a\ c>$ | 0.229 | 0.477 |
| | | JDF | $<c\ a\ b\ e\ d>$ | 0.355 | 0.377 |
| | | | $<e\ d\ a\ b\ c>$ | 0.268 | 0.332 |
| | | | $<d\ b\ a\ c\ e>$ | 0.165 | 0.291 |
| | Restricted DBM | BIC | $<c\ a\ b\ e\ d>$ | 0.408 | 0.535 |
| | | | $<d\ e\ b\ a\ c>$ | 0.408 | 0.465 |
| | Unrestricted DBM | BIC | $<a\ c\ e\ d\ b>$ | 0.060 | 1.000 |
| | | | $<a\ c\ e\ d\ b>$ | 0.817 | 0.000 |
| Sports | CCA | PC & PE | $<f\ e\ d\ c\ g\ a\ b>$ | 0.295 | 0.281 |
| | | | $<a\ b\ c\ d\ e\ f\ g>$ | 0.139 | 0.505 |
| | | JDF | $<a\ b\ c\ d\ e\ f\ g>$ | 0.332 | 0.359 |
| | | | $<e\ f\ d\ g\ c\ a\ b>$ | 0.273 | 0.330 |
| | | | $<f\ e\ a\ c\ d\ g\ b>$ | 0.228 | 0.311 |
| | Restricted DBM | BIC | $<d\ a\ e\ f\ c\ b\ g>$ | 0.234 | 0.501 |
| | | | $<e\ c\ f\ a\ b\ d\ g>$ | 0.234 | 0.499 |
| | Unrestricted DBM | BIC | $<e\ f\ c\ a\ d\ b\ g>$ | 0.144 | 1.000 |
| | | | $<e\ f\ d\ c\ g\ a\ b>$ | 0.953 | 0.000 |

[1] The spread parameters are estimated different, for CCA equation (5.4), for DBM equations (5.16) and (5.17), respectively.

# 8 Discussion

In this thesis, we have studied partitioning methods for ranking data. Let us recall our research questions: which method is the most suitable for recovering the population median ranking in a simulation study and which internal validity index adequately reveals the number of clusters. Based on the extensive simulation experiment in section 7.1 we conclude that the $K$-Median Cluster Component Analysis (CCA) algorithm performs best under all possible factor levels, with smallest standard error: indicating stable and good recovery. Of the two mixtures of distance-based models (DBM), the configuration with restricted spread parameter recovers the estimated population rankings better than its unrestricted counterpart. In addition, the unrestricted DBM shares more similarities when looking at real data sets. Based on the partitioning of the real data sets in section 7.2 we suggest to use CCA and to verify the number of components in the data with the partition entropy (PE).

Four distance measures that measure dissimilarity between pairs of judges have been distinguished. The traditional Kendall distance ($d_{\tau_b}$) violates the triangle inequality between complete and tied rankings. This has been corrected by Emond & Mason (2000, 2002) to $d_{\tau_x}$ that properly deals with tied rankings and is equivalent to the Kemeny distance. Spearman's $\rho$ suffers from the sensitivity of irrelevant alternatives. The appropriate distance measures are Emond & Mason's distance and the Kemeny distance, which are equivalent.

A homogeneous population of $n$ judges of $m$ objects can be summarized by the modal, median or mean ranking. The modal ranking is defined as the most observed ranking in the sample. It can be troublesome if two (or more) rankings are observed equally often. The median ranking is defined as the ranking that minimizes the distance to a single ranking on the space of rankings. However, this ranking may not be uniquely defined when many rankings satisfy the minimum distance. With many observed rankings and high frequencies this is not a problem as the data examples show. The mean ranking based on the Kemeny distance does not always generate much insight into the data. It prefers tied rankings and can return the uninformative all-ties ranking as the estimate.

If a population may be composed of multiple groups, it can be partitioned into a finite number of $K$ clusters. The two partitioning methods are CCA, and the DBM. They estimate the central ranking, mixing probabilities and spread parameter similarly. However, they estimate the membership probabilities differently. When updating the membership probabilities CCA only depends on the Kemeny distances to the cluster centers of the previous iteration, whereas the DBM also depends on the estimated spread and mixing probability parameters. Both algorithms were fitted with 50 different starting values to minimize the loss function (CCA) and to maximize the complete-data log-likelihood (DBM).

An interesting proposition for future research would be to extend these models to be fitted with partial and incomplete rankings. If the missing objects in the case of partial rankings are treated as tied on the last position, they are located on the sample space. Then the models can be fitted, without any further adjustment.

Another suggestion for further research is to model the ranking process instead of the population of judges like we did here, Marden (1996, p. 111). This can be done by the Plackett-Luce model proposed by Plackett (1975) and Luce (1959). The ranking process is modeled by decomposing the most preferred of $m$ objects into $m-1$ stages. This model has recently been extended to a heterogeneous population of judges by Gormley & Murphy (2006) and Csiszár (2012) to multiple groups with the MM and EM algorithm,

respectively. The difference between these algorithms can be found in Hunter (2004). A third model to sort rankings has very recently been proposed by Biernacki & Jacques (2013) based on the insertion sort algorithm.

A final suggestion that could be interesting to follow up is the procedure given in Iyigun (2007, Chapter 4), where the probabilistic $d$-clustering model is extended to probabilistic $dq$-clustering. The membership probabilities are adjusted for cluster size. The numerator and denominator in equation (5.1) are adjusted by dividing the distances with the cluster probability in equation (5.3). This extension could be an improvement of the CCA algorithm. In our simulation study CCA showed a lower recovery with unequal cluster sizes.

## References


Baggerly, K. A. (1995), Visual estimation of structure in ranked data, PhD thesis, Rice University.

Ben-Israel, A. & Iyigun, C. (2008), 'Probabilistic d-clustering', *Journal of Classification* **25**(1), 5–26.

Bezdek, J. C. (1974), 'Cluster validity with fuzzy sets', *Journal of Cybernetics* **3**(3), 58–73.

Biernacki, C. & Jacques, J. (2013), 'A generative model for rank data based on insertion sort algorithm', *Journal of Computational Statistics & Data Analysis* **58**(1), 162–176.

Busse, L. W., Orbanz, P. & Buhmann, J. M. (2007), Cluster analysis of heterogeneous rank data, *in* 'Proceedings of the 24th International Conference on Machine Learning (ICML)', pp. 113–120.

Critchlow, D. E. (1985), *Metric methods for analyzing partially ranked data*, Vol. 34 of *Lecture Notes in Statistics*, Springer-Verlag: Berlin.

Critchlow, D. E., Fligner, M. E. & Verducci, J. E. (1991), 'Probability models on rankings', *Journal of Mathematical Psychology* **35**(3), 294–318.

Croon, M. A. (1989), Latent class models for the analysis of rankings, *in* G. de Soete, H. Feger & K. C. Klauer, eds, 'New developments in psychological choice modeling', Elsevier: Amsterdam, pp. 99–121.

Csiszár, V. (2012), 'EM algorithms for generalized Bradley-Terry models', *Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös Nominatae (Sectio Computatorica)* **36**(1), 143–157.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM Algorithm', *Journal of the Royal Statistical Society, Series B (Methodological)* **39**(1), 1–38.

Diaconis, P. (1988), *Group representations in probability and statistics*, Vol. 11 of *Lecture Notes–Monograph Series*, Institute of Mathematical Statistics: Hayward, California.

Emond, E. J. (1997), Maximum rank correlation as a solution concept in the $m$ rankings problem with application to multi criteria decision analysis, Technical Report DOR (CAM) Research Note 9705, Department of National Defence, Canada.

Emond, E. J. & Mason, D. W. (2000), A new technique for high level decision support, Technical Report DOR (CAM) Project Report 2000/13, Department of National Defence, Canada.

Emond, E. J. & Mason, D. W. (2002), 'A new rank correlation coefficient with application to the consensus ranking problem', *Journal of Multi-Criteria Decision Analysis* **11**(1), 17–28.

Fligner, M. A. & Verducci, J. S. (1986), 'Distance based ranking models', *Journal of the Royal Statistical Society, Series B (Methodological)* **48**(3), 359–369.

Gordon, A. D. (1999), *Classification*, Vol. 82 of *Monographs on Statistics and Applied Probability*, 2nd edn, Chapman & Hall: Boca Raton, Florida.

Gormley, I. C. & Murphy, T. B. (2006), 'Analysis of Irish third-level college applications data', *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **169**(2), 361–379.

Gregory, E. (2012), *RMallow: fit multi-modal Mallows' models to ranking data*. `R` package version 1.0.

Gross, O. A. (1962), 'Preferential arrangements', *American Mathematical Monthly* **69**(1), 4–8.

Heiser, W. J. (2004), 'Geometric representation of association between categories', *Psychometrika* **69**(4), 513–545.

Heiser, W. J. & D'Ambrosio, A. (in press), Clustering and prediction of rankings within a Kemeny distance framework, *in* B. Lausen, D. van den Poel & A. Ultsch, eds, 'Algorithms from and for Nature and Life', Springer-Verlag: Berlin.

Hunter, D. R. (2004), 'MM algorithms for generalized Bradley-Terry models', *The Annals of Statistics* **32**(1), 384–406.

Iyigun, C. (2007), Probabilistic distance clustering, PhD thesis, Rutgers University.

Kemeny, J. G. (1959), 'Mathematics without numbers', *Daedalus* **88**(4), 577–591.

Kemeny, J. G. & Snell, J. L. (1972), *Mathematical models in the social sciences*, MIT Press: Cambridge, Massachusetts.

Kendall, M. G. (1948), *Rank correlation methods*, 4 edn, Griffin: London.

Lee, P. H. & Yu, P. L. H. (2011), *pmr: probability models for ranking data*. `R` package version 1.1.1.

Lee, P. H. & Yu, P. L. H. (2012), 'Mixtures of weighted distance-based models for ranking data with applications in political studies', *Journal of Computational Statistics & Data Analysis* **56**(8), 2486–2500.

Luce, D. R. (1959), *Individual choice behavior*, Wiley & Sons: New York, New York.

Mallows, C. L. (1957), 'Non-null ranking models. I', *Biometrika* **44**(1), 114–130.

Marden, J. I. (1996), *Analyzing and modeling rank data*, Vol. 64 of *Monographs on Statistics and Applied Probability*, Chapman & Hall: Boca Raton, Florida.

McLachlan, G. & Peel, D. (2000), *Finite mixture models*, Wiley: New York, New York.

Murphy, T. B. & Martin, D. (2003), 'Mixtures of distance-based models for ranking data', *Journal of Computational Statistics & Data Analysis* **41**(3), 645–655.

Plackett, R. L. (1975), 'The analysis of permutations', *Applied Statistics* **24**(2), 193–202.

R Development Core Team (2012), *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna.

Regenwetter, M., Grofman, B., Marley, A. A. J. & Tsetlin, I. (2006), *Behavioral social choice: probabilistic models, statistical inference, and applications*, Cambridge University Press: Cambridge.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**(2), 461–464.

Spearman, C. (1904), 'The proof and measurement of association between two things', *The American Journal of Psychology* **15**(1), 72–101.

Thompson, G. L. (1993), 'Generalized permutation polytopes and exploratory graphical models for ranked data', *The Annals of Statistics* **21**(3), 1401–1430.

# A Results of the recovery simulation

The following tables present the results of the simulation study described in section 7.1. Each cell represents the mean recovery of ten replications. Table 18 gives the results with four objects. Table 19 gives the results with five objects. Table 20 gives the results of the first part with seven objects and Table 21 the remaining part with seven objects.

Table 18: Mean recovery of the clustering with four objects with CCA and DBM with restricted and unrestricted spread parameter in between brackets.

| | | | | 2 centers Complete rankings | | 2 centers Tied rankings | | 3 centers Complete rankings | | 3 centers Tied rankings | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Algorithm | Correlation | Spread | Cluster / Size | Small | Large | Small | Large | Small | Large | Small | Large |
| CCA | Negative | Low | Equal | 1.00 | 1.00 | 0.93 | 0.94 | 0.99 | 1.00 | 0.76 | 0.84 |
| | | | Unequal | 0.98 | 1.00 | 0.95 | 0.96 | 0.92 | 0.99 | 0.75 | 0.82 |
| | | Varying | Equal | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.91 | 0.97 |
| | | | Unequal | 1.00 | 1.00 | 0.99 | 1.00 | 0.98 | 1.00 | 0.87 | 0.89 |
| | | High | Equal | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.96 | 1.00 |
| | | | Unequal | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.95 | 0.99 |
| | Uncorrelated | Low | Equal | 1.00 | 1.00 | 0.91 | 0.92 | | | 0.82 | 0.89 |
| | | | Unequal | 0.97 | 1.00 | 0.87 | 0.92 | | | 0.80 | 0.84 |
| | | Varying | Equal | 1.00 | 1.00 | 0.97 | 1.00 | | | 0.93 | 0.97 |
| | | | Unequal | 1.00 | 0.97 | 0.97 | 1.00 | | | 0.87 | 0.88 |
| | | High | Equal | 1.00 | 1.00 | 0.99 | 1.00 | | | 0.96 | 0.99 |
| | | | Unequal | 1.00 | 1.00 | 0.99 | 1.00 | | | 0.96 | 0.95 |
| | Positive | Low | Equal | 1.00 | 1.00 | 0.86 | 0.83 | 0.96 | 0.96 | 0.76 | 0.77 |
| | | | Unequal | 1.00 | 1.00 | 0.84 | 0.83 | 0.92 | 0.92 | 0.72 | 0.76 |
| | | Varying | Equal | 1.00 | 1.00 | 0.93 | 0.93 | 1.00 | 1.00 | 0.84 | 0.89 |
| | | | Unequal | 1.00 | 1.00 | 0.93 | 0.94 | 0.96 | 1.00 | 0.77 | 0.76 |
| | | High | Equal | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.93 | 0.98 |
| | | | Unequal | 1.00 | 1.00 | 0.96 | 1.00 | 0.96 | 1.00 | 0.88 | 0.93 |
| DBM ($d_{\text{Kem}}$) | Negative | Low | Equal | 1.00 (0.38) | 1.00 (0.45) | 0.67 (0.50) | 0.80 (0.61) | 0.74 (0.39) | 0.91 (0.39) | 0.33 (0.26) | 0.22 (0.26) |
| | | | Unequal | 0.95 (0.53) | 1.00 (0.52) | 0.54 (0.37) | 0.47 (0.35) | 0.64 (0.49) | 0.63 (0.51) | 0.31 (0.26) | 0.23 (0.26) |
| | | Varying | Equal | 1.00 (0.87) | 1.00 (0.40) | 0.84 (0.43) | 0.75 (0.33) | 1.00 (0.41) | 1.00 (0.51) | 0.46 (0.45) | 0.44 (0.48) |
| | | | Unequal | 0.77 (0.90) | 0.87 (0.93) | 0.48 (0.45) | 0.47 (0.42) | 0.87 (0.82) | 0.76 (0.66) | 0.51 (0.42) | 0.47 (0.55) |
| | | High | Equal | 1.00 (1.00) | 1.00 (1.00) | 1.00 (0.93) | 1.00 (1.00) | 1.00 (0.47) | 1.00 (0.37) | 0.65 (0.52) | 0.46 (0.71) |
| | | | Unequal | 1.00 (1.00) | 1.00 (1.00) | 0.68 (0.88) | 0.63 (0.71) | 1.00 (0.98) | 1.00 (1.00) | 0.58 (0.46) | 0.59 (0.56) |
| | Uncorrelated | Low | Equal | 0.92 (0.52) | 1.00 (0.50) | 0.73 (0.67) | 0.81 (0.72) | | | 0.50 (0.58) | 0.57 (0.57) |
| | | | Unequal | 0.88 (0.58) | 0.98 (0.63) | 0.58 (0.63) | 0.59 (0.58) | | | 0.56 (0.57) | 0.48 (0.59) |
| | | Varying | Equal | 1.00 (0.60) | 1.00 (0.50) | 0.69 (0.71) | 0.82 (0.93) | | | 0.66 (0.57) | 0.60 (0.65) |
| | | | Unequal | 0.80 (0.72) | 0.87 (0.77) | 0.62 (0.56) | 0.71 (0.52) | | | 0.52 (0.58) | 0.56 (0.53) |
| | | High | Equal | 1.00 (1.00) | 1.00 (1.00) | 0.99 (0.97) | 1.00 (1.00) | | | 0.63 (0.64) | 0.62 (0.66) |
| | | | Unequal | 0.95 (0.98) | 1.00 (1.00) | 0.78 (0.82) | 0.85 (0.97) | | | 0.60 (0.61) | 0.64 (0.59) |
| | Positive | Low | Equal | 0.88 (0.73) | 0.87 (0.67) | 0.78 (0.82) | 0.79 (0.90) | 0.84 (0.69) | 0.90 (0.67) | 0.72 (0.67) | 0.71 (0.66) |
| | | | Unequal | 0.95 (0.78) | 1.00 (0.75) | 0.68 (0.75) | 0.69 (0.78) | 0.80 (0.67) | 0.83 (0.69) | 0.72 (0.72) | 0.73 (0.72) |
| | | Varying | Equal | 1.00 (0.77) | 1.00 (0.88) | 0.84 (0.81) | 0.89 (0.84) | 0.90 (0.78) | 0.91 (0.80) | 0.74 (0.73) | 0.75 (0.73) |
| | | | Unequal | 0.85 (0.67) | 0.87 (0.67) | 0.69 (0.73) | 0.71 (0.69) | 0.88 (0.78) | 0.88 (0.78) | 0.77 (0.73) | 0.76 (0.75) |
| | | High | Equal | 1.00 (1.00) | 1.00 (1.00) | 0.95 (0.90) | 1.00 (1.00) | 0.92 (0.84) | 0.92 (0.89) | 0.73 (0.72) | 0.72 (0.73) |
| | | | Unequal | 0.97 (0.90) | 0.97 (0.70) | 0.95 (0.84) | 0.93 (0.87) | 0.87 (0.79) | 0.89 (0.80) | 0.77 (0.76) | 0.74 (0.78) |

Table 19: Mean recovery of the clustering with five objects with CCA and DBM with restricted and unrestricted spread parameter in between brackets.

| | | | | 2 centers Complete rankings | | 2 centers Tied rankings | | 3 centers Tied rankings | | 4 centers Tied rankings | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Algorithm | Correlation | Spread | Cluster / Size | Small | Large | Small | Large | Small | Large | Small | Large |
| CCA | Negative | Low | Equal | 1.00 | 1.00 | 0.95 | 0.98 | | | | |
| | | | Unequal | 0.99 | 1.00 | 0.92 | 1.00 | | | | |
| | | Varying | Equal | 1.00 | 1.00 | 1.00 | 1.00 | | | | |
| | | | Unequal | 1.00 | 1.00 | 0.96 | 1.00 | | | | |
| | | High | Equal | 1.00 | 1.00 | 1.00 | 1.00 | | | | |
| | | | Unequal | 1.00 | 1.00 | 1.00 | 1.00 | | | | |
| | Uncorrelated | Low | Equal | 1.00 | 1.00 | 0.95 | 0.98 | 0.92 | 0.93 | 0.84 | 0.94 |
| | | | Unequal | 0.99 | 1.00 | 0.88 | 0.94 | 0.80 | 0.91 | 0.76 | 0.92 |
| | | Varying | Equal | 1.00 | 1.00 | 1.00 | 1.00 | 0.95 | 0.99 | 0.93 | 0.98 |
| | | | Unequal | 1.00 | 1.00 | 0.98 | 0.96 | 0.84 | 0.95 | 0.86 | 0.96 |
| | | High | Equal | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 0.98 | 1.00 |
| | | | Unequal | 1.00 | 1.00 | 0.98 | 1.00 | 0.97 | 0.99 | 0.95 | 0.99 |
| | Positive | Low | Equal | 1.00 | 1.00 | 0.92 | 0.96 | 0.82 | 0.88 | 0.71 | 0.75 |
| | | | Unequal | 1.00 | 1.00 | 0.90 | 0.96 | 0.85 | 0.86 | 0.72 | 0.73 |
| | | Varying | Equal | 1.00 | 1.00 | 0.96 | 1.00 | 0.96 | 0.99 | 0.78 | 0.80 |
| | | | Unequal | 0.94 | 1.00 | 0.83 | 0.99 | 0.89 | 0.96 | 0.80 | 0.80 |
| | | High | Equal | 1.00 | 1.00 | 0.99 | 1.00 | 0.97 | 1.00 | 0.86 | 0.83 |
| | | | Unequal | 1.00 | 1.00 | 1.00 | 1.00 | 0.95 | 0.98 | 0.85 | 0.87 |
| DBM ($d_{\text{Kem}}$) | Negative | Low | Equal | 1.00 (0.66) | 1.00 (0.56) | 0.94 (0.78) | 0.93 (0.89) | | | | |
| | | | Unequal | 0.97 (0.48) | 1.00 (0.72) | 0.60 (0.45) | 0.52 (0.48) | | | | |
| | | Varying | Equal | 1.00 (1.00) | 1.00 (1.00) | 0.94 (0.69) | 1.00 (1.00) | | | | |
| | | | Unequal | 0.96 (0.93) | 0.96 (1.00) | 0.69 (0.62) | 0.64 (0.62) | | | | |
| | | High | Equal | 1.00 (1.00) | 1.00 (1.00) | 1.00 (0.86) | 1.00 (1.00) | | | | |
| | | | Unequal | 1.00 (1.00) | 1.00 (1.00) | 1.00 (0.90) | 0.94 (1.00) | | | | |
| | Uncorrelated | Low | Equal | 1.00 (0.51) | 1.00 (0.50) | 0.90 (0.55) | 0.91 (0.52) | 0.62 (0.49) | 0.52 (0.45) | 0.52 (0.45) | 0.55 (0.41) |
| | | | Unequal | 0.96 (0.62) | 1.00 (0.65) | 0.64 (0.55) | 0.62 (0.52) | 0.60 (0.51) | 0.60 (0.54) | 0.54 (0.40) | 0.48 (0.38) |
| | | Varying | Equal | 1.00 (0.80) | 1.00 (1.00) | 0.96 (0.74) | 1.00 (0.95) | 0.66 (0.56) | 0.55 (0.43) | 0.64 (0.46) | 0.44 (0.36) |
| | | | Unequal | 0.96 (0.80) | 0.95 (0.85) | 0.60 (0.56) | 0.61 (0.60) | 0.77 (0.60) | 0.84 (0.68) | 0.55 (0.45) | 0.56 (0.44) |
| | | High | Equal | 1.00 (1.00) | 1.00 (1.00) | 1.00 (1.00) | 1.00 (1.00) | 0.97 (0.68) | 1.00 (0.90) | 0.78 (0.56) | 0.65 (0.52) |
| | | | Unequal | 1.00 (1.00) | 1.00 (1.00) | 0.98 (0.88) | 0.85 (0.95) | 0.85 (0.66) | 0.83 (0.93) | 0.48 (0.45) | 0.50 (0.49) |
| | Positive | Low | Equal | 0.96 (0.70) | 0.97 (0.70) | 0.82 (0.72) | 0.76 (0.76) | 0.74 (0.66) | 0.71 (0.69) | 0.73 (0.71) | 0.70 (0.68) |
| | | | Unequal | 0.92 (0.76) | 0.84 (0.72) | 0.80 (0.76) | 0.79 (0.74) | 0.76 (0.70) | 0.70 (0.70) | 0.74 (0.70) | 0.70 (0.68) |
| | | Varying | Equal | 1.00 (0.76) | 1.00 (0.70) | 0.90 (0.76) | 0.80 (0.74) | 0.81 (0.76) | 0.78 (0.77) | 0.75 (0.75) | 0.74 (0.74) |
| | | | Unequal | 0.84 (0.70) | 0.80 (0.70) | 0.78 (0.72) | 0.76 (0.70) | 0.79 (0.66) | 0.82 (0.68) | 0.80 (0.70) | 0.77 (0.70) |
| | | High | Equal | 1.00 (1.00) | 1.00 (1.00) | 0.98 (0.73) | 1.00 (0.97) | 0.77 (0.78) | 0.82 (0.80) | 0.74 (0.74) | 0.73 (0.73) |
| | | | Unequal | 0.96 (0.86) | 0.94 (0.92) | 0.94 (0.82) | 0.90 (0.86) | 0.76 (0.73) | 0.87 (0.79) | 0.72 (0.70) | 0.74 (0.70) |

Table 20: Mean recovery of the clustering with seven objects with CCA and DBM with restricted and and unrestricted spread parameter in between brackets, 1/2.

| | | | | 2 centers Complete rankings | | 2 centers Tied rankings | | 3 centers Complete rankings | |
|---|---|---|---|---|---|---|---|---|---|
| Algorithm | Correlation | Spread | Cluster / Size | Small | Large | Small | Large | Small | Large |
| CCA | Negative | Low | Equal | 1.00 | 1.00 | 0.99 | 1.00 | | |
| | | | Unequal | 1.00 | 1.00 | 0.96 | 1.00 | | |
| | | Varying | Equal | 1.00 | 1.00 | 0.99 | 1.00 | | |
| | | | Unequal | 1.00 | 1.00 | 1.00 | 1.00 | | |
| | | High | Equal | 1.00 | 1.00 | 1.00 | 1.00 | | |
| | | | Unequal | 1.00 | 1.00 | 1.00 | 1.00 | | |
| | Uncorrelated | Low | Equal | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 1.00 |
| | | | Unequal | 1.00 | 1.00 | 0.97 | 1.00 | 0.99 | 1.00 |
| | | Varying | Equal | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | Unequal | 1.00 | 1.00 | 0.99 | 1.00 | 0.99 | 1.00 |
| | | High | Equal | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | Unequal | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | Positive | Low | Equal | 1.00 | 1.00 | 0.96 | 0.99 | 0.99 | 1.00 |
| | | | Unequal | 0.95 | 1.00 | 0.95 | 0.95 | 0.92 | 1.00 |
| | | Varying | Equal | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 |
| | | | Unequal | 0.95 | 0.98 | 0.94 | 0.95 | 0.89 | 0.95 |
| | | High | Equal | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | | | Unequal | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| DBM ($d_{\text{Kem}}$) | Negative | Low | Equal | 0.50 (0.86) | 0.29 (0.72) | 0.30 (0.65) | 0.29 (0.36) | | |
| | | | Unequal | 0.96 (0.65) | 0.92 (0.67) | 0.69 (0.57) | 0.70 (0.58) | | |
| | | Varying | Equal | 0.93 (1.00) | 1.00 (1.00) | 0.64 (0.76) | 0.36 (1.00) | | |
| | | | Unequal | 0.90 (0.71) | 0.97 (0.60) | 0.63 (0.64) | 0.79 (0.65) | | |
| | | High | Equal | 1.00 (0.93) | 1.00 (1.00) | 0.36 (0.43) | 0.65 (0.79) | | |
| | | | Unequal | 1.00 (0.82) | 1.00 (1.00) | 0.87 (0.61) | 0.90 (0.62) | | |
| | Uncorrelated | Low | Equal | 0.67 (0.70) | 1.00 (0.71) | 0.52 (0.53) | 0.52 (0.52) | 0.82 (0.61) | 0.90 (0.61) |
| | | | Unequal | 0.96 (0.65) | 0.95 (0.77) | 0.81 (0.58) | 0.89 (0.61) | 0.71 (0.49) | 0.53 (0.57) |
| | | Varying | Equal | 1.00 (0.97) | 1.00 (1.00) | 0.90 (0.94) | 1.00 (1.00) | 1.00 (0.87) | 1.00 (1.00) |
| | | | Unequal | 1.00 (0.71) | 1.00 (1.00) | 0.56 (0.57) | 0.59 (0.58) | 0.71 (0.73) | 0.78 (0.95) |
| | | High | Equal | 1.00 (1.00) | 1.00 (1.00) | 0.67 (0.66) | 0.86 (0.86) | 0.91 (0.90) | 0.72 (1.00) |
| | | | Unequal | 1.00 (1.00) | 1.00 (1.00) | 1.00 (0.72) | 0.86 (0.81) | 0.66 (0.70) | 0.68 (0.84) |
| | Positive | Low | Equal | 0.92 (0.76) | 1.00 (0.77) | 0.80 (0.72) | 0.72 (0.72) | 0.83 (0.72) | 0.72 (0.71) |
| | | | Unequal | 0.89 (0.74) | 0.93 (0.72) | 0.86 (0.75) | 0.75 (0.73) | 0.81 (0.70) | 0.81 (0.71) |
| | | Varying | Equal | 1.00 (0.98) | 1.00 (1.00) | 0.98 (0.93) | 1.00 (0.94) | 0.92 (0.90) | 0.86 (0.89) |
| | | | Unequal | 0.79 (0.79) | 0.71 (0.71) | 0.80 (0.74) | 0.85 (0.74) | 0.86 (0.80) | 0.89 (0.93) |
| | | High | Equal | 1.00 (1.00) | 1.00 (1.00) | 1.00 (1.00) | 0.97 (0.92) | 0.76 (0.77) | 0.72 (0.72) |
| | | | Unequal | 1.00 (1.00) | 1.00 (1.00) | 0.91 (0.93) | 0.89 (0.91) | 0.90 (0.85) | 0.83 (0.81) |

Table 21: Mean recovery of the clustering with seven objects with CCA and DBM with restricted and unrestricted spread parameter in between brackets, 2/2.

| | | | | 3 centers Tied rankings | | 4 centers Complete rankings | | 4 centers Tied rankings | |
|---|---|---|---|---|---|---|---|---|---|
| Algorithm | Correlation | Spread | Cluster / Size | Small | Large | Small | Large | Small | Large |
| CCA | Uncorrelated | Low | Equal | 0.97 | 1.00 | 0.98 | 1.00 | 0.94 | 0.99 |
| | | | Unequal | 0.94 | 0.98 | 0.97 | 1.00 | 0.85 | 0.97 |
| | | Varying | Equal | 0.99 | 1.00 | 1.00 | 1.00 | 0.96 | 1.00 |
| | | | Unequal | 0.85 | 0.97 | 1.00 | 0.94 | 0.76 | 0.80 |
| | | High | Equal | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 |
| | | | Unequal | 0.99 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 |
| | Positive | Low | Equal | 0.94 | 0.97 | 0.96 | 0.99 | 0.87 | 0.94 |
| | | | Unequal | 0.90 | 0.96 | 0.91 | 0.95 | 0.88 | 0.91 |
| | | Varying | Equal | 0.99 | 1.00 | 0.99 | 1.00 | 0.95 | 0.97 |
| | | | Unequal | 0.90 | 0.87 | 0.88 | 0.91 | 0.89 | 0.86 |
| | | High | Equal | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 |
| | | | Unequal | 0.97 | 1.00 | 1.00 | 1.00 | 0.96 | 0.99 |
| DBM ($d_{\text{Kem}}$) | Uncorrelated | Low | Equal | 0.68 (0.58) | 0.69 (0.56) | 0.67 (0.54) | 0.50 (0.39) | 0.66 (0.57) | 0.57 (0.51) |
| | | | Unequal | 0.71 (0.52) | 0.58 (0.52) | 0.70 (0.52) | 0.72 (0.56) | 0.63 (0.48) | 0.60 (0.51) |
| | | Varying | Equal | 0.94 (0.69) | 0.78 (0.76) | 0.65 (0.60) | 0.76 (0.67) | 0.75 (0.59) | 0.63 (0.54) |
| | | | Unequal | 0.68 (0.50) | 0.69 (0.69) | 0.60 (0.74) | 0.65 (0.55) | 0.63 (0.43) | 0.56 (0.43) |
| | | High | Equal | 0.73 (0.71) | 0.81 (0.78) | 0.78 (0.75) | 0.69 (0.80) | 0.67 (0.69) | 0.57 (0.66) |
| | | | Unequal | 0.79 (0.62) | 0.90 (0.68) | 0.72 (0.57) | 0.68 (0.65) | 0.62 (0.49) | 0.61 (0.56) |
| | Positive | Low | Equal | 0.75 (0.69) | 0.82 (0.65) | 0.82 (0.73) | 0.84 (0.70) | 0.78 (0.69) | 0.75 (0.69) |
| | | | Unequal | 0.83 (0.69) | 0.90 (0.69) | 0.78 (0.73) | 0.79 (0.71) | 0.77 (0.70) | 0.70 (0.63) |
| | | Varying | Equal | 0.96 (0.77) | 0.83 (0.65) | 0.88 (0.78) | 0.82 (0.82) | 0.84 (0.76) | 0.89 (0.76) |
| | | | Unequal | 0.81 (0.72) | 0.83 (0.73) | 0.82 (0.75) | 0.88 (0.85) | 0.76 (0.66) | 0.76 (0.70) |
| | | High | Equal | 0.94 (0.85) | 0.96 (1.00) | 0.96 (0.95) | 0.97 (0.96) | 0.91 (0.86) | 0.87 (0.88) |
| | | | Unequal | 0.78 (0.78) | 0.79 (0.79) | 0.87 (0.85) | 0.86 (0.85) | 0.78 (0.77) | 0.65 (0.68) |

# B Observed rankings of the real data sets

The observed rankings of the data sets in section 7.2 are given in the following tables. Table 22 lists the observed rankings and frequencies of the 'Voting' data set. Table 23 lists the observed rankings and frequencies of the 'Living places' data set. Table 24 lists the observed rankings and frequencies of the 'Political goals' data set. Table 25 lists the observed rankings and frequencies of the 'Song' data set. Table 26 lists the observed rankings and frequencies of the 'Idea' data set. Table 27 lists the observed rankings and frequencies of the 'Rice subset' data set. Table 28 lists the observed rankings and frequencies of the 'APA subset' data set. Table 29 lists the observed rankings and frequencies of the 'Sports' data set.

Table 22: Observed rankings and frequencies of the 'Voting' data set.

| Ranking | Frequency | Ranking | Frequency |
|---|---|---|---|
| $a\ b\ c$ | | $a\ b\ c$ | |
| (1 2 3) | 232 | (2 3 1) | 151 |
| (1 3 2) | 132 | (3 1 2) | 114 |
| (2 1 3) | 213 | (3 2 1) | 141 |

Table 23: Observed rankings and frequencies of the 'Living places' data set.

| Ranking | Frequency | Ranking | Frequency |
|---|---|---|---|
| $a\ b\ c$ | | $a\ b\ c$ | |
| (1 2 3) | 242 | (2 3 1) | 628 |
| (1 3 2) | 28 | (3 1 2) | 12 |
| (2 1 3) | 170 | (3 2 1) | 359 |

Table 24: Observed rankings and frequencies of the 'Political goals' data set.

| Ranking | Frequency | Ranking | Frequency |
|---|---|---|---|
| $a\ b\ c\ d$ | | $a\ b\ c\ d$ | |
| (1 2 3 4) | 137 | (3 1 2 4) | 330 |
| (1 2 4 3) | 29 | (3 1 4 2) | 294 |
| (1 3 2 4) | 309 | (3 2 1 4) | 117 |
| (1 3 4 2) | 255 | (3 2 4 1) | 69 |
| (1 4 2 3) | 52 | (3 4 1 2) | 70 |
| (1 4 3 2) | 93 | (3 4 2 1) | 34 |
| (2 1 3 4) | 48 | (4 1 2 3) | 21 |
| (2 1 4 3) | 23 | (4 1 3 2) | 30 |
| (2 3 1 4) | 61 | (4 2 1 3) | 29 |
| (2 3 4 1) | 55 | (4 2 3 1) | 52 |
| (2 4 1 3) | 33 | (4 3 1 2) | 35 |
| (2 4 3 1) | 59 | (4 3 2 1) | 27 |

Table 25: Observed rankings and frequencies of the 'Song' data set.

| Ranking | Frequency | Ranking | Frequency |
|---|---|---|---|
| $a\ b\ c\ d\ e$ | | $a\ b\ c\ d\ e$ | |
| (1 2 3 4 5) | 7 | (3 2 1 5 4) | 6 |
| (1 3 2 4 5) | 9 | (4 1 2 3 5) | 2 |
| (2 1 3 4 5) | 4 | (4 2 1 3 5) | 8 |
| (2 3 1 4 5) | 10 | (4 3 1 2 5) | 2 |
| (2 4 1 3 5) | 3 | (5 2 1 3 4) | 5 |
| (3 1 2 4 5) | 6 | (5 2 1 4 3) | 2 |
| (3 2 1 4 5) | 19 | | |

Table 26: Observed rankings and frequencies in the 'Idea' data set.

| Ranking | Frequency | Ranking | Frequency | Ranking | Frequency |
|---|---|---|---|---|---|
| $a\ b\ c\ d\ e$ | | $a\ b\ c\ d\ e$ | | $a\ b\ c\ d\ e$ | |
| (1 3 4 5 2) | 1 | (4 2 3 5 1) | 2 | (5 1 4 2 3) | 6 |
| (1 4 2 3 5) | 1 | (4 3 5 2 1) | 1 | (5 1 4 3 2) | 33 |
| (3 2 5 4 1) | 2 | (5 1 2 4 3) | 5 | (5 2 3 4 1) | 8 |
| (4 1 2 5 3) | 1 | (5 1 3 2 4) | 2 | (5 2 4 1 3) | 1 |
| (4 1 5 3 2) | 5 | (5 1 3 4 2) | 18 | (5 2 4 3 1) | 12 |

Table 27: Complete rankings and frequencies of the 'Rice subset' data set.

| Ranking | Frequency | Ranking | Frequency | Ranking | Frequency |
|---|---|---|---|---|---|
| $a\ b\ c\ d\ e$ | | $a\ b\ c\ d\ e$ | | $a\ b\ c\ d\ e$ | |
| (1 2 3 5 4) | 8 | (2 3 1 4 5) | 5 | (4 1 3 5 2) | 1 |
| (1 2 4 3 5) | 2 | (2 3 1 5 4) | 18 | (4 1 5 2 3) | 1 |
| (1 2 5 3 4) | 7 | (2 4 1 3 5) | 1 | (4 2 1 5 3) | 1 |
| (1 3 2 4 5) | 10 | (2 4 1 5 3) | 5 | (4 2 3 1 5) | 3 |
| (1 3 2 5 4) | 20 | (2 4 3 5 1) | 1 | (4 2 3 5 1) | 1 |
| (1 3 4 5 2) | 2 | (2 4 5 3 1) | 1 | (4 3 1 2 5) | 2 |
| (1 3 5 4 2) | 1 | (2 5 4 3 1) | 1 | (4 3 1 5 2) | 1 |
| (1 4 2 3 5) | 3 | (3 1 2 4 5) | 2 | (4 3 5 1 2) | 1 |
| (1 4 2 5 3) | 1 | (3 1 2 5 4) | 11 | (4 5 1 2 3) | 1 |
| (1 4 3 2 5) | 1 | (3 1 4 2 5) | 1 | (4 5 1 3 2) | 1 |
| (1 5 2 3 4) | 2 | (3 1 5 4 2) | 2 | (4 5 2 1 3) | 1 |
| (1 5 2 4 3) | 2 | (3 2 1 4 5) | 9 | (5 1 2 4 3) | 2 |
| (1 5 3 2 4) | 1 | (3 2 1 5 4) | 4 | (5 1 3 4 2) | 1 |
| (1 5 3 4 2) | 1 | (3 2 4 1 5) | 3 | (5 1 4 3 2) | 2 |
| (1 5 4 3 2) | 1 | (3 2 5 1 4) | 3 | (5 2 1 3 4) | 1 |
| (2 1 3 4 5) | 7 | (3 2 5 4 1) | 1 | (5 2 4 3 1) | 1 |
| (2 1 3 5 4) | 19 | (3 4 1 2 5) | 1 | (5 3 1 2 4) | 1 |
| (2 1 4 3 5) | 1 | (3 4 5 2 1) | 1 | (5 3 1 4 2) | 1 |
| (2 1 4 5 3) | 1 | (3 5 1 4 2) | 1 | (5 3 2 1 4) | 1 |
| (2 1 5 3 4) | 4 | (4 1 2 5 3) | 4 | (5 4 1 2 3) | 1 |
| (2 1 5 4 3) | 7 | (4 1 3 2 5) | 2 | (5 4 1 3 2) | 1 |

Table 28: Complete rankings and frequencies of the 'APA subset' data set.

| Ranking | Frequency | Ranking | Frequency | Ranking | Frequency |
|---|---|---|---|---|---|
| *a b c d e* | | *a b c d e* | | *a b c d e* | |
| (1 2 3 4 5) | 30 | (2 4 5 1 3) | 53 | (4 2 3 1 5) | 51 |
| (1 2 3 5 4) | 28 | (2 4 5 3 1) | 63 | (4 2 3 5 1) | 24 |
| (1 2 4 3 5) | 27 | (2 5 1 3 4) | 79 | (4 2 5 1 3) | 66 |
| (1 2 4 5 3) | 29 | (2 5 1 4 3) | 106 | (4 2 5 3 1) | 58 |
| (1 2 5 3 4) | 35 | (2 5 3 1 4) | 21 | (4 3 1 2 5) | 35 |
| (1 2 5 4 3) | 34 | (2 5 3 4 1) | 40 | (4 3 1 5 2) | 38 |
| (1 3 2 4 5) | 102 | (2 5 4 1 3) | 34 | (4 3 2 1 5) | 35 |
| (1 3 2 5 4) | 95 | (2 5 4 3 1) | 35 | (4 3 2 5 1) | 30 |
| (1 3 4 2 5) | 35 | (3 1 2 4 5) | 34 | (4 3 5 1 2) | 84 |
| (1 3 4 5 2) | 37 | (3 1 2 5 4) | 30 | (4 3 5 2 1) | 91 |
| (1 3 5 2 4) | 28 | (3 1 4 2 5) | 42 | (4 5 1 2 3) | 30 |
| (1 3 5 4 2) | 35 | (3 1 4 5 2) | 40 | (4 5 1 3 2) | 38 |
| (1 4 2 3 5) | 45 | (3 1 5 2 4) | 34 | (4 5 2 1 3) | 24 |
| (1 4 2 5 3) | 70 | (3 1 5 4 2) | 30 | (4 5 2 3 1) | 34 |
| (1 4 3 2 5) | 24 | (3 2 1 4 5) | 74 | (4 5 3 1 2) | 54 |
| (1 4 3 5 2) | 51 | (3 2 1 5 4) | 82 | (4 5 3 2 1) | 31 |
| (1 4 5 2 3) | 48 | (3 2 4 1 5) | 75 | (5 1 2 3 4) | 29 |
| (1 4 5 3 2) | 52 | (3 2 4 5 1) | 34 | (5 1 2 4 3) | 11 |
| (1 5 2 3 4) | 50 | (3 2 5 1 4) | 64 | (5 1 3 2 4) | 19 |
| (1 5 2 4 3) | 70 | (3 2 5 4 1) | 41 | (5 1 3 4 2) | 25 |
| (1 5 3 2 4) | 17 | (3 4 1 2 5) | 35 | (5 1 4 2 3) | 46 |
| (1 5 3 4 2) | 36 | (3 4 1 5 2) | 87 | (5 1 4 3 2) | 50 |
| (1 5 4 2 3) | 35 | (3 4 2 1 5) | 28 | (5 2 1 3 4) | 50 |
| (1 5 4 3 2) | 40 | (3 4 2 5 1) | 62 | (5 2 1 4 3) | 35 |
| (2 1 3 4 5) | 40 | (3 4 5 1 2) | 133 | (5 2 3 1 4) | 24 |
| (2 1 3 5 4) | 30 | (3 4 5 2 1) | 107 | (5 2 3 4 1) | 26 |
| (2 1 4 3 5) | 26 | (3 5 1 2 4) | 36 | (5 2 4 1 3) | 44 |
| (2 1 4 5 3) | 24 | (3 5 1 4 2) | 45 | (5 2 4 3 1) | 54 |
| (2 1 5 3 4) | 42 | (3 5 2 1 4) | 27 | (5 3 1 2 4) | 26 |
| (2 1 5 4 3) | 36 | (3 5 2 4 1) | 41 | (5 3 1 4 2) | 34 |
| (2 3 1 4 5) | 172 | (3 5 4 1 2) | 61 | (5 3 2 1 4) | 22 |
| (2 3 1 5 4) | 186 | (3 5 4 2 1) | 71 | (5 3 2 4 1) | 22 |
| (2 3 4 1 5) | 52 | (4 1 2 3 5) | 16 | (5 3 4 1 2) | 49 |
| (2 3 4 5 1) | 53 | (4 1 2 5 3) | 22 | (5 3 4 2 1) | 57 |
| (2 3 5 1 4) | 52 | (4 1 3 2 5) | 23 | (5 4 1 2 3) | 28 |
| (2 3 5 4 1) | 45 | (4 1 3 5 2) | 31 | (5 4 1 3 2) | 43 |
| (2 4 1 3 5) | 96 | (4 1 5 2 3) | 45 | (5 4 2 1 3) | 24 |
| (2 4 1 5 3) | 162 | (4 1 5 3 2) | 50 | (5 4 2 3 1) | 37 |
| (2 4 3 1 5) | 28 | (4 2 1 3 5) | 40 | (5 4 3 1 2) | 67 |
| (2 4 3 5 1) | 44 | (4 2 1 5 3) | 52 | (5 4 3 2 1) | 29 |

Table 29: Observed rankings and frequencies of the 'Sports' data set.

| Ranking | Frequency | Ranking | Frequency | Ranking | Frequency |
|---|---|---|---|---|---|
| $a\ b\ c\ d\ e\ f\ g$ | | $a\ b\ c\ d\ e\ f\ g$ | | $a\ b\ c\ d\ e\ f\ g$ | |
| (1 2 3 4 5 6 7) | 1 | (3 1 2 5 6 7 4) | 1 | (5 6 3 7 1 4 2) | 1 |
| (1 2 3 4 5 7 6) | 1 | (3 1 2 7 4 6 5) | 1 | (5 6 4 2 3 1 7) | 1 |
| (1 2 3 4 6 5 7) | 1 | (3 1 2 7 5 4 6) | 1 | (5 6 4 3 1 2 7) | 1 |
| (1 2 3 5 4 7 6) | 1 | (3 2 1 4 5 6 7) | 1 | (5 6 4 3 2 1 7) | 1 |
| (1 2 5 6 3 4 7) | 1 | (3 2 1 4 6 5 7) | 1 | (5 7 3 4 2 1 6) | 1 |
| (1 2 7 5 3 4 6) | 1 | (3 2 1 4 7 5 6) | 1 | (5 7 4 1 2 6 3) | 1 |
| (1 3 2 5 4 7 6) | 2 | (3 2 4 1 5 6 7) | 1 | (5 7 6 1 4 2 3) | 1 |
| (1 3 2 5 6 4 7) | 1 | (3 4 5 7 6 2 1) | 1 | (5 7 6 4 1 3 2) | 1 |
| (1 3 2 5 7 6 4) | 1 | (3 4 6 7 2 5 1) | 1 | (6 1 2 5 3 4 7) | 1 |
| (1 3 2 6 5 7 4) | 1 | (3 5 2 1 7 6 4) | 1 | (6 1 5 3 4 2 7) | 1 |
| (1 3 4 5 6 2 7) | 1 | (3 6 5 4 1 2 7) | 1 | (6 2 4 5 1 3 7) | 1 |
| (1 3 4 6 2 5 7) | 1 | (3 7 2 4 6 1 5) | 1 | (6 2 4 7 3 1 5) | 1 |
| (1 3 4 7 5 6 2) | 1 | (3 7 4 5 2 1 6) | 1 | (6 3 1 5 4 2 7) | 1 |
| (1 3 7 2 4 5 6) | 1 | (3 7 4 6 2 1 5) | 1 | (6 3 1 5 7 4 2) | 1 |
| (1 4 3 2 5 7 6) | 1 | (3 7 5 4 2 6 1) | 1 | (6 3 4 7 2 1 5) | 1 |
| (1 4 5 6 7 3 2) | 1 | (3 7 6 5 4 2 1) | 1 | (6 4 5 2 1 3 7) | 1 |
| (1 4 7 3 2 5 6) | 1 | (4 1 2 3 5 6 7) | 1 | (6 7 2 4 1 3 5) | 1 |
| (1 5 2 7 3 6 4) | 1 | (4 1 3 2 6 5 7) | 1 | (6 7 3 1 2 4 5) | 1 |
| (1 5 3 2 4 6 7) | 1 | (4 1 3 5 2 7 6) | 1 | (6 7 4 1 2 5 3) | 1 |
| (1 5 4 3 2 6 7) | 1 | (4 1 6 2 5 3 7) | 1 | (6 7 4 1 3 2 5) | 1 |
| (1 5 4 6 2 3 7) | 1 | (4 1 7 2 6 5 3) | 1 | (6 7 4 3 1 2 5) | 1 |
| (1 6 2 3 4 7 5) | 1 | (4 2 1 3 5 6 7) | 1 | (6 7 4 3 2 1 5) | 1 |
| (1 7 4 3 5 2 6) | 1 | (4 3 5 7 2 1 6) | 1 | (6 7 5 3 2 1 4) | 1 |
| (1 7 6 5 3 2 4) | 1 | (4 5 3 1 6 7 2) | 1 | (6 7 5 3 4 1 2) | 1 |
| (2 1 3 5 6 7 4) | 1 | (4 5 7 1 3 6 2) | 1 | (6 7 5 4 1 2 3) | 1 |
| (2 1 3 7 5 6 4) | 1 | (4 5 7 6 2 1 3) | 1 | (6 7 5 4 2 3 1) | 3 |
| (2 1 4 3 6 7 5) | 1 | (4 6 1 5 3 2 7) | 1 | (7 1 3 2 5 4 6) | 1 |
| (2 1 5 6 3 4 7) | 1 | (4 6 1 7 3 2 5) | 1 | (7 2 5 3 6 1 4) | 1 |
| (2 1 6 7 3 5 4) | 1 | (4 6 3 5 2 7 1) | 1 | (7 4 1 3 5 2 6) | 1 |
| (2 3 1 4 6 5 7) | 1 | (4 6 5 3 1 2 7) | 1 | (7 4 5 6 1 3 2) | 1 |
| (2 3 1 5 6 4 7) | 1 | (4 6 5 3 2 1 7) | 1 | (7 4 6 3 5 1 2) | 1 |
| (2 3 1 7 6 5 4) | 1 | (4 6 7 1 3 2 5) | 1 | (7 5 1 6 3 4 2) | 1 |
| (2 3 4 1 5 6 7) | 1 | (4 6 7 2 5 3 1) | 1 | (7 5 3 1 4 2 6) | 1 |
| (2 3 4 1 6 5 7) | 1 | (4 6 7 5 1 2 3) | 1 | (7 5 3 4 1 2 6) | 1 |
| (2 4 3 7 1 5 6) | 1 | (4 7 2 3 5 1 6) | 1 | (7 5 6 1 2 3 4) | 1 |
| (2 4 5 1 6 3 7) | 1 | (4 7 3 1 5 6 2) | 1 | (7 6 1 4 2 3 5) | 1 |
| (2 5 1 3 6 4 7) | 1 | (4 7 3 5 1 2 6) | 1 | (7 6 2 1 5 4 3) | 1 |
| (2 5 3 1 4 6 7) | 2 | (4 7 6 2 3 1 5) | 1 | (7 6 3 2 1 4 5) | 1 |
| (2 7 4 5 1 3 6) | 1 | (5 1 2 4 3 6 7) | 1 | (7 6 4 3 1 2 5) | 1 |
| (2 7 6 4 5 1 3) | 1 | (5 1 2 6 4 3 7) | 1 | (7 6 4 5 3 2 1) | 1 |
| (3 1 2 4 5 7 6) | 1 | (5 4 7 1 2 3 6) | 1 | (7 6 5 1 2 4 3) | 1 |
| (3 1 2 5 6 4 7) | 1 | (5 4 7 6 2 1 3) | 1 | (7 6 5 1 3 2 4) | 1 |