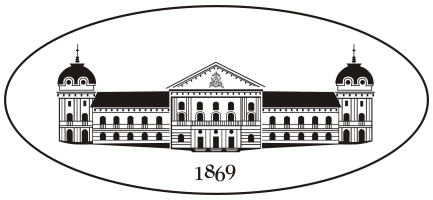


BULGARIAN ACADEMY OF SCIENCES

INSTITUTE OF MATHEMATICS AND INFORMATICS

METRIC METHODS FOR ANALYZING AND MODELING RANK DATA

NIKOLAY IVANCHEV NIKOLOV

THESIS

FOR CONFERRING OF ACADEMIC AND SCIENTIFIC DEGREE

DOCTOR

IN PROFESSIONAL FIELD 4.5 MATHEMATICS
(PROBABILITY THEORY AND MATHEMATICAL STATISTICS)

SUPERVISOR:
PROF. EUGENIA STOIMENOVA

Sofia, 2020

Acknowledgements

I wish to express my sincerely gratitude to my supervisor Professor Eugenia Stoimenova for her valuable help, motivation and support. I would also like to thank all the staff members of "Operations Research, Probability and Statistics" department for their generous attitude and friendly behaviour.

The research was supported by the Bulgarian Ministry of Education and Science under the National Research Programme "Young scientists and postdoctoral students" approved by DCM #577/17.08.2018.

Introduction

Rank data commonly arise from situations where it is desired to rank a set of individuals or objects in accordance with some criterion. Such data may be observed directly or it may come from a ranking of a set or subset of assigned scores. Alternatively, rank data may arise when transforming continuous or discrete data in a nonparametric analysis. Examples of rank data may be found in politics (Inglehart [35], Moors and Vermunt [53]), voting and elections (Diaconis [17], Koop and Poirier [40], Gormley and Murphy [29]), market research (Beggs et al. [4], Chapman and Staelin [8]), food preference (Kamishima and Akaho [38], Nombekela et al. [63]), phycology (Maydeu-Olivares and Böckenholt [50], Regenwetter et al. [70]), health economics (Salomon [71]) and medical treatments (Plumb et al. [69]). In general there are two types of rankings: complete and partial, depending on whether it is required to rank all of the objects or not. In this thesis, we will restrict our attention to the case where all objects are ranked, i.e. when the complete rankings are observed.

A complete ranking of N items simply assigns a full ordering to the items. Any such ranking vector can be viewed as an element π of the permutation group $\mathbf{S}_N$ generated by the first N positive integers. A permutation $\pi \in \mathbf{S}_N$ is a function from $\{1,\ldots,N\}$ onto itself, whose arguments are the items, and whose values are the ranks. If the items are labeled with the numbers $\{1,\ldots,N\}$, then $\pi(i)$ is the rank given to item i and $\pi^{-1}(i)$ is the item assigned the rank i. Thus

$$\pi^{-1} = \langle \pi^{-1}(1), \pi^{-1}(2), \ldots, \pi^{-1}(N) \rangle$$

is the permutation in $\mathbf{S}_N$ which corresponds to listing the objects in their ranked order. For a sample of n complete rankings we will use the notation $\pi^1, \pi^2, \ldots, \pi^n \in \mathbf{S}_N$.

Rank data have a natural structure that presents challenges and opportunities that are unavailable in typical multivariate samples. There is a rich body of work on analyzing rank data that includes the classical probabilistic models proposed by Thurstone [74], Bradley and Terry [5], Luce [44], Plackett [68] and Mallows [45]. Mallows' models are often convenient initial tool for analyzing a set of rank data. They capture the main structure of the data with only one parameter and could be the basis for further research. However, it is usually unrealistic to expect a one-parameter model to reveal all features of the data. One possible generalization of these models could be made by assuming that there are several latent groups in the population. The problem of finding the "consensus" ranking and clustering rankings has been widely studied by many authors, see Busse et al. [6], Klementiev et al. [39] and Murphy and Martin [55]. Most of these methods can be described in a form that involves distances on permutations, which are powerful tool for uncovering the hidden features of the rank data. Numerical characteristics, exact distributions, asymptotic approximations and statistical applications of the random variables based on the most commonly used distances on $\mathbf{S}_N$ can be found in Diaconis [17] and Marden [47]. An example of a more exotic distance is the Lee distance, which has been developed by Lee [42] as a generalization of the Hamming distance for error correcting coding in modulation. An application of Lee distance in a visual

recognition problem is given in Chan et al. [7]. However, the statistical properties of the Lee distance are not well-studied. In this thesis, certain asymptotic approximations for the random variable based on Lee distance are derived and applied to several probabilistic models for rank data and to other statistical problems involving rankings.

There are various applications of rankings in many applied scientific areas. One example can be found in the imperfect ranking analysis of the ranked set sampling (RSS) procedure. RSS can be used for creating more efficient methods for large range of statistical problems. McIntyre [51] first proposed the mean of a RSS as an estimator of the population mean. Later, Dell and Clutter [14] showed that the RSS mean is an unbiased estimator and is at least as precise as the simple random sample mean based on the same number of observations. Moreover, this remarkable fact is true regardless if the judgment ranking is perfect or not. However, the effectiveness of RSS mean depends directly on how well the judgment orderings within each set are obtained. More statistical developments based on RSS, such as variance estimation, quantiles estimation, density function estimation, M-estimates and distribution-free tests, are described in Chen et al. [9]. The benefit of using these RSS procedures is most significant when we have perfect ranking, but this is not always feasible. Hence, it is desirable to construct statistical models that capture the uncertainty of the judgment ordering process and test whether the rankings are perfect or not. These models can be defined by the ranking error probability matrix, which can be used to study the effect of imperfect ranking on the performance of the statistical procedures based on RSS, see Aragon et al. [2] and Section 3.1.2 in Chen et al. [9]. Nonparametric tests for null hypothesis of perfect rankings against a general alternative of imperfect ranking have been developed by Frey et al. [24], Li and Balakrishnan [43], Vock and Balakrishnan [79] and Zamanzade et al. [82]. Zamanzade and Vock [83] developed nonparametric tests of perfect ranking for judgment post stratification sampling scheme. In the case when the hypothesis of perfect ranking is rejected, the process of judgment ranking within the sets should be analyzed. Frey and Wang [25] considered four models for imperfect ranking: Bivariate normal model proposed by Dell and Clutter [14], Fraction of random rankings by Frey et al. [24], Fraction of inverse rankings by Frey et al. [24] and Fraction of neighbor rankings by Vock and Balakrishnan [79]. More models for imperfect ranking are presented in Frey [26] and Ozturk [65]. Furthermore, these models can be used to compare the ranking abilities of two judges or two ranking methods in order to increase the effectiveness of the RSS procedures for future observation measurements.

Rankings and distances on permutations find another application in one of the most important statistical problems: the comparison of two samples. If we assume that parent population distributions may differ only in location, there are many parametric and nonparametric tests at our disposal. The nonparametric approach requires few assumptions about the underlying distribution generating the data and gives us the ability to choose the test statistic that is best suited for the task at hand. Nonparametric tests have been applied in a variety of statistical procedures, for example in cluster analysis (Hubert and Levin [34]) and in Fourier analysis (Friedman and Lane [22]), and in a wide range of scientific areas: from anthropology (Fisher [20]) to atmospheric science (Tukey et al. [76]). Numerous statistical tools for nonparametric analysis are exhaustively described in Hollander and Wolfe [33] and Gibbons and Chakraborti [27]. There are various of techniques for constructing rank tests for hypothesis testing of two samples, see Hájek and Šidák [30] and Good [28]. Critchlow [11]

proposed a unified approach based on the minimum distance between two separate permutation sets corresponding to the null and the alternative hypothesis. By using different distances on permutations in Critchlow's method we obtain different test statistics. Some of the most popular rank statistics: Kolmogorov-Smirnov, Wilcoxon and Mann-Whitney statistics, can be derived by Ulam distance, Spearman's footrule and Kendall's tau, respectively. One of the benefits of having several test statistics is that they can be combined in order to produce more powerful procedures. Pesarin [67] developed an interesting theory, the Nonparametric Combination of Dependent Tests, which yields good results for many complex multivariate problems, including problems that have not yet been solved within a parametric setting.

The main objective of the thesis is to study the statistical properties of Lee distance and to explore its applications in several rank data models based on distances. In particular, our goals are to:

- Obtain an asymptotic result for the distribution of the random variable induced by Lee distance under uniformity of the rankings.
- Compare the Mallows' model based on Lee distance to other probability models for rank data.
- Propose an Expectation-Maximization algorithm for estimating the unknown parameters in Distance-based models for rank data with several latent groups.
- Give an approximation of the measure of "tightness" in the "K-means" clustering procedure for rank data based on Lee distance.
- Find an asymptotic approximation of the ranking error probability matrix based on Lee distance, Spearman's footrule and Spearman's rho in the framework of ranked set sampling.
- Present a procedure for estimating the unknown parameters in the Mallows' model for imperfect ranking by making use of the Expectation-Maximization technique.
- Derive a rank test statistic based on Critchlow's method and Lee distance for the two-sample location problem and study its distribution under the null hypothesis.

In Chapter 1, we present some properties of distances on permutations and the Lee distance in particular. In Section 1.1, we define eight commonly used distances and apply them to an illustrative example. Section 1.2 deals with the mean, variance and asymptotic distribution of the random variable induced by Lee distance. The results in Chapter 1 are based on Nikolov and Stoimenova [58, 59].

Chapter 2 is devoted to probabilistic models for rank data. In Sections 2.1 and 2.2, Distance-based models and their relation to Marginals models are studied. A description of some statistical tools for estimating the unknown parameters and testing the goodness-of-fit of the presented models are given in Section 2.3. For the case when there are latent groups in the population and the central rankings are unknown, an EM algorithm is proposed in Section 2.4. As an application of the obtained results, three illustrative examples are provided in Section 2.5. The study presented in Chapter 2 is based on Nikolov and Stoimenova [59, 60].

In Chapter 3, we consider the "K-means" clustering based on Lee distance. In Section 3.1, the "K-means" clustering procedure for complete rankings and its relation to distances on permutations are presented. Some properties and asymptotic results for Lee distance are given in Section 3.2. In Section 3.3, the presented clustering method is applied to the well-studied American Psychological Association election dataset. The results presented in Chapter 3 are based on Nikolov and Stoimenova [62].

In Chapter 4, we use the Distance-based models to describe the imperfect ranking in the framework of n-cycle balanced RSS. In Section 4.2, we discuss nonparametric methods for perfect rankings, present some of the existing test statistics and introduce new similar statistics. Distance-based models in the framework of RSS are studied in Section 4.3. In Section 4.4, we propose an EM algorithm for estimating the unknown parameter in the Mallows' model for imperfect ranking. In Section 4.5, models based on different distances are considered and some asymptotic results for the corresponding error matrix are derived. Power comparisons of the described tests for one-cycle RSS are provided in Section 4.6. An illustrative example and some concluding remarks are given in Section 4.7. The exposition of Chapter 4 is based on Nikolov and Stoimenova [61].

In Chapter 5, we study a rank test statistic induced by Lee distance. In Section 5.1, the Critchlow's approach is described and applied for the two-sample location problem. The test statistic obtained by the Critchlow's method and the Lee distance is derived in Section 5.2. The joint distribution of the statistics based on Hamming distance and Lee distance and its asymptotic properties under the null hypothesis are given in Section 5.3. In Section 5.4, the test statistic induced by Lee distance is compared to others through a simulation study for samples generated by t-distributions. The results in Chapter 5 are based on Nikolov [57].

The main contributions and accomplishments in the thesis due to the author are listed in Appendix A.

Contents

Chapter 1

Preliminaries

Ranking usually occurs when several raters determine the order of N items based on their preference on the items. Rank data commonly arises in a variety of areas ranging from preference rankings in psychology and social choice theory, to more modern learning tasks in online web search, crowd-sourcing and recommendation systems. Distances on permutations are often convenient tools for analyzing and modeling rank data. They measure the closeness between two rankings and can be very useful and informative for revealing the main structure and features of the data. In this chapter, distances on permutations are defined and several examples are considered. Some important statistical properties of Lee distance, which will be used in the next chapters, are derived.

1.1 Distances on permutations

A common goal with rank data is to obtain a central ranking and to study how close the observations are clustered around it. Although the componentwise average is a fine candidate for describing the center, it is not an actual rank vector unless all observed rankings are the same. Since the average minimizes the sum of squared Euclidian distances, a natural modification for obtaining a central ranking is to minimize the sum of the distances between the center and the observed rankings, which are elements of $\mathbf{S}_N$. In order to find the optimal central ranking first we need to define a distance between two rankings. The definitions and the exposition of this section are based on Chapters 2 and 3 of Marden [47].

Definition 1.1 (Distance on $\mathbf{S}_N$)**.** *A function $d : \mathbf{S}_N \times \mathbf{S}_N \to \mathbb{R}$ is a distance on $\mathbf{S}_N$ if it satisfies:*

(i) $d(\pi,\sigma) > 0$*, for $\pi \neq \sigma$ and $\pi,\sigma \in \mathbf{S}_N$;*

(ii) $d(\pi,\pi) = 0$ *for every $\pi \in \mathbf{S}_N$;*

(iii) $d(\pi,\sigma) = d(\sigma,\pi)$ *for every $\pi,\sigma \in \mathbf{S}_N$.*

Then for a sample of n complete rankings $\pi^1,\pi^2,\ldots,\pi^n \in \mathbf{S}_N$, one possible measure of cental location for a fixed distance $d(\cdot,\cdot)$ is the ranking $\hat{\sigma}$ that minimizes

$$\bar{d}(\sigma) = \frac{1}{n}\sum_{i=1}^{n} d\left(\pi^i,\sigma\right), \quad \text{for } \sigma \in \mathbf{S}_N.$$

Measures other than average, such as median, can be used. Notice that in general the minimizer $\hat{\sigma}$ is not unique. The spread of the data around the center $\hat{\sigma}$ is naturally measured by the quantity $\bar{d}(\hat{\sigma})$. However, this value highly depends on the choice of the distance $d(\cdot,\cdot)$. More

properties of the central ranking estimate and a normalization of the spread measurement are considered in details for more general settings in Chapter 3.

Let us focus on some properties of distances on the permutation group $\mathbf{S}_N$. Deza and Huang[16] considered some distances on $\mathbf{S}_N$ which are widely used in applied scientific and statistical problems.

$$
\begin{aligned}
d_F(\pi,\sigma) &= \sum_{i=1}^{N} \mid \pi(i)-\sigma(i) \mid && \text{Spearman's footrule}\\
d_R(\pi,\sigma) &= \sqrt{\sum_{i=1}^{N} \big(\pi(i)-\sigma(i)\big)^2} && \text{Spearman's rho}\\
d_M(\pi,\sigma) &= \max_{1\le i\le N} \mid \pi(i)-\sigma(i) \mid && \text{Chebyshev distance}\\
d_K(\pi,\sigma) &= \#\{(i,j) : 1\le i,j\le N, \pi(i)<\pi(j), \sigma(i)>\sigma(j)\} && \text{Kendall's tau}\\
d_C(\pi,\sigma) &= k \text{ minus number of cycles in } \sigma\circ\pi^{-1} && \text{Cayley distance}\\
d_U(\pi,\sigma) &= N \text{ minus the length of longest increasing} && \text{Ulam distance}\\
&\quad \text{subsequence in } \sigma\circ\pi^{-1} &&\\
d_H(\pi,\sigma) &= \#\{i\in\{1,2,\ldots,N\} : \pi(i)\neq\sigma(i)\} && \text{Hamming distance}\\
d_L(\pi,\sigma) &= \sum_{i=1}^{N} \min\big(\mid \pi(i)-\sigma(i) \mid, N- \mid \pi(i)-\sigma(i) \mid\big) && \text{Lee distance}
\end{aligned}
$$

Notice that the Spearman's footrule and the Spearman's rho are special cases (for $p=1$ and $p=2$) of the natural set of p-distances defined by

$$
d(\pi,\sigma) = \left(\sum_{i=1}^{N} \mid \pi(i)-\sigma(i) \mid^p\right)^{\frac{1}{p}} \quad \text{for } 0<p<\infty. \tag{1.1}
$$

The Spearman's rho (d_R) is the Euclidian distance and is commonly used in nonparametric statistics. Further, the standard normalization of d_R can be written as the average product-moment correlation coefficient between π and σ. The Chebyshev distance and the Hamming distance can be considered as limits of p-distances when $p\to\infty$ and $p\to 0$ respectively. It is not hard to see that the Hamming and Lee distances coincide when $N\le 3$. The Kendall's tau, Cayley and Ulam distances find application in various ranking problems. In order to illustrate some differences between the listed distances, we will consider the canonical example of arranging books on a shelf into alphabetic order. Let us start with five books in order BECAD, which corresponds to the ranking $\langle 4,1,3,5,2\rangle$. The steps that are required to order the books to ABCDE for six of the listed distances are given in Table 1.1. The distance between the rankings $\langle 4,1,3,5,2\rangle$ and $\langle 1,2,3,4,5\rangle$ is presented in each column as the number of steps needed to arrange BECAD to ABCDE.

From the definitions of the eight distances above we see that Kendall' tau is the number of discordant pairs, which for $\langle 1,2,3,4,5\rangle$ and $\langle 4,1,3,5,2\rangle$ are five: $(1,2)$; $(1,3)$; $(1,4)$; $(3,2)$ and $(4,5)$. Equivalently, Kendall' tau is the minimum number of simple transpositions that are needed to order BECAD to ABCDE. Instead of restricting to adjacent pairs, the Cayley distance counts the minimum number of arbitrary pairwise interchanges that are required to bring BECAD to ABCDE. That decreases the number of steps to 4. To describe the Ulam distance, suppose that the books can slide along the shelf. The "deletion-insertion" method of arranging chooses books one at a time, removing them and reinserting at a better place, while the remaining books are possibly slightly shifted to accommodate the insertion. The

Steps ↓	d_K	d_C	d_U	d_H	d_L	d_F
0	BECAD	BECAD	BECAD	␣␣C␣␣	(B)(E) C (A)(D)	[B][E] C [A][D]
1	BEACD	AECBD	ABECD	A␣C␣␣	A (E) C (A)(D)	A [E] C [A][D]
2	BAECD	ABCED	ABCDE	ABC␣␣	A (A) C (A)(D)	A [D] C [A][D]
3	ABECD	ABCDE		ABCD␣	A B C (A)(D)	A [C] C [A][D]
4	ABCED			ABCDE	A B C (E)(D)	A B C [A][D]
5	ABCDE				A B C D (D)	A B C [B][D]
6					A B C D E	A B C [C][D]
7						A B C D [D]
8						A B C D E

TABLE 1.1: Steps of arranging the books from BECAD to ABCDE

Ulam distance is the minimum number of such operations and equals 2 for the example in Table 1.1. The Hamming distance corresponds of the process of removing all books that are not on the "correct" position and then inserting them one at a time.

Since d_K, d_C, d_U and d_H measure the disorder between two rankings, they can be categorized as distances of disorder. Another type of distances are spatial distances, which measure the minimum distance between the pairwise elements of rankings on some geometric structure. For the example of arranging the books from BECAD to ABCDE, Spearman's footrule is the total sum of minimum distances between the pairs (B,A); (E,B); (A,D) and (D,E) if the books are connected on a straight line from A to E with a unit distance between them. For Lee distance, we can think that the books are connected not on a line, but on a simple cyclic graph with unit arcs. The two connection structures for d_L and d_F are illustrated on Figure 1.1.

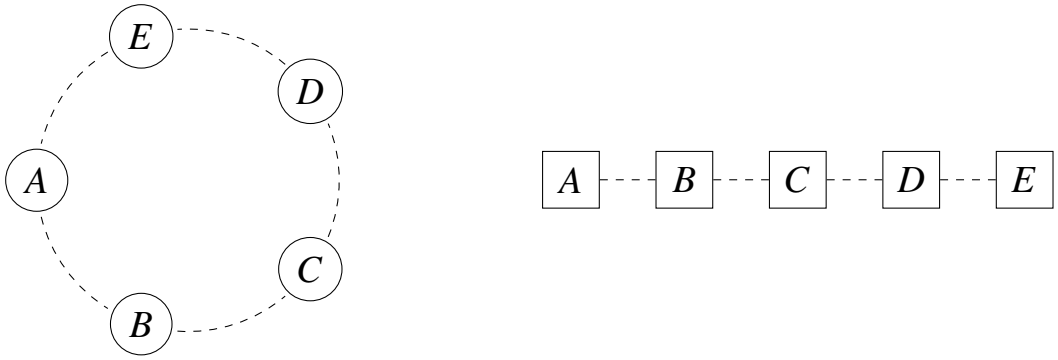


FIGURE 1.1: Connection structures of the books for Lee distance (on left) and Spearman's footrule (on right)

Distances between rankings which can be expressed as a total sum of componentwise distances are examples of one rich class of distances called Hoeffding distances.

Definition 1.2 (Hoeffding distance)**.** *A distance $d(\cdot,\cdot)$ on $\mathbf{S}_N$ is called Hoeffding distance, if*

$$d(\pi,\sigma) = \sum_{i=1}^{N} a\big(\pi(i),\sigma(i)\big),$$

where $a(\cdot,\cdot)$ is a function on $\{1,2,\ldots,N\}\times\{1,2,\ldots,N\}$ that satisfies $a(i,j)=a(j,i)$ and $a(i,i)=0$.

From the eight distances listed in this section d_F, d_H and d_L can be rewritten in the form of Hoeffding distances. We can obtain a Hoeffding distance if we take the square of d_R:

$$\left(d_R(\pi,\sigma)\right)^2 = d_{R^2}(\pi,\sigma) = \sum_{i=1}^{N}\left(\pi(i)-\sigma(i)\right)^2.$$

In a similar way all p-distances in (1.1) can be transformed as Hoeffding distances. However, with this method the transformed p-distances lose the following property.

Definition 1.3 (Metric on $\mathbf{S}_N$)**.** *A distance $d(\cdot,\cdot)$ is a metric on $\mathbf{S}_N$ if it satisfies the triangle inequality:*

$$d(\pi,\sigma)\le d(\pi,\tau)+d(\tau,\sigma),\quad \text{for every } \pi,\sigma,\tau\in\mathbf{S}_N.$$

It is not hard to show that all eight listed distances in this section are metrics on $\mathbf{S}_N$. The following are two other important properties of distances on $\mathbf{S}_N$.

Definition 1.4 (Right-invariance)**.** *A distance $d(\cdot,\cdot)$ on $\mathbf{S}_N$ is called right-invariant (label-invariant), if and only if $d(\pi,\sigma)=d(\pi\circ\tau,\sigma\circ\tau)$ for every $\pi,\sigma,\tau\in\mathbf{S}_N$.*

Definition 1.5 (Left-invariance)**.** *A distance $d(\cdot,\cdot)$ on $\mathbf{S}_N$ is called left-invariant (rank-invariant), if and only if $d(\pi,\sigma)=d(\tau\circ\pi,\tau\circ\sigma)$ for every $\pi,\sigma,\tau\in\mathbf{S}_N$.*

As remarked in Critchlow [10]: the right-invariance of a distance is necessary requirement since it ensures that the distance between rankings does not depend on the labelling of the objects. Deza and Huang [16] pointed that all eight listed distances are right-invariant. For example, if we change the labels of books A and E in Table 1.1, then we need to arrange the books from BACED to EBCDA. However, this will not affect the number of steps required and therefore will not change the distances between the two orderings.

The left-invariance of a distance means that the distance between two rankings does not use the numerical values (from 1 to N) of the rankings but only the way they are ordered. For example, the Hamming distance compares the elements of two rankings componentwise and does not depend on the integers from 1 to N. From the listed distances only d_C, d_U and d_H are left-invariant. If a distance is both right-invariant and left-invariant, it is called bi-invariant. More properties of distances on $\mathbf{S}_N$ can be found in Critchlow [10, 12], Deza and Huang [16], Diaconis [17], Diaconis and Graham [19], Kruskal [41] and Marden [47].

Let us consider again the problem of finding central ranking, which can be also called *modal* ranking. Assume that $\pi_0\in\mathbf{S}_N$ is fixed modal ranking. If π is randomly selected from $\mathbf{S}_N$ (i.e. $\pi\sim Uniform(\mathbf{S}_N)$), then $D=d(\pi,\pi_0)$ is a random variable. When the distance $d(\cdot,\cdot)$ is right-invariant

$$D=d(\pi,\pi_0)=d(\pi\circ\pi_0^{-1},e_N),$$

where e_N is the identity permutation ($e_N=\langle 1,2,\ldots,N\rangle$), and the distribution of D does not depend on π_0. In Chapter 2 we will consider probability models which are significantly simplified if the distribution of D is known. In the next section, we will study in more details the random variable D based on Lee distance.

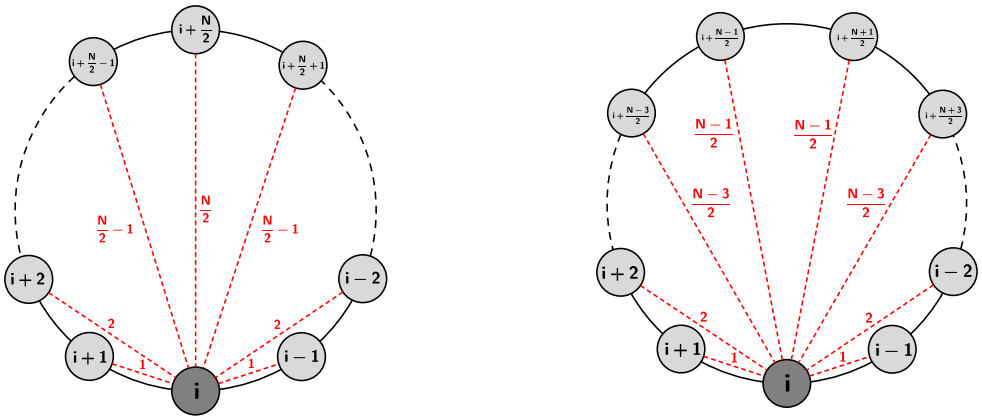


FIGURE 1.2: Graph G when N is even (on left) and when N is odd (on right)

1.2 Properties of Lee distance

Let us use the notation D_L for the random variable induced by Lee distance under uniformity of π. Notice that $D_L(\pi)$ can be decomposed linearly:

$$D_L(\pi) = d_L(\pi, e_N) = \sum_{i=1}^{N} \min\left(|\,\pi(i) - i\,|, N - |\,\pi(i) - i\,|\right) = \sum_{i=1}^{N} c_N(\pi(i), i)\,. \tag{1.2}$$

There is an interpretation of $c_N(i,j) := \min\left(|\,i-j\,|, N - |\,i-j\,|\right)$ in terms of graph theory. Let G be a simple cycle graph with nodes $\{i\}_{i=1}^{N}$ and edges $\bigcup_{i=1}^{N-1} \{i, i+1\}$ and $\{N, 1\}$. Then $c_N(i,j)$ is the minimum distance over G between the nodes i and j. On Figure 1.2, the graph G is illustrated separately when N is odd and when N is even. The quantities $c(i,j)$ for fixed node i are presented in red color.

Obviously, $0 \le c_N(i,j) \le N/2$ for even N and $0 \le c_N(i,j) \le (N-1)/2$ for odd N, i.e.

$$0 \le c_N(i,j) \le \left[\frac{N}{2}\right], \quad \text{for all } i,j = 1,2,\ldots,N\,, \tag{1.3}$$

where $[x]$ is the greatest integer less than or equal to x. From (1.2) and (1.3) it follows that

$$0 \le D_L(\pi) \le N\left[\frac{N}{2}\right], \quad \text{for all } \pi \in \mathbf{S}_N\,. \tag{1.4}$$

The lower limit in (1.4) is reached only for $\pi = e_N$. When N is even the upper limit is reached only for π equals to

$$e_N^* := \left\langle \frac{N}{2}+1, \frac{N}{2}+2, \ldots, N-1, N, 1, 2, \ldots, \frac{N}{2}-1, \frac{N}{2} \right\rangle,$$

and in the case of odd integers N the maximum value of D_L is reached when

$$\pi = e_N' := \left\langle \frac{N+1}{2}, \frac{N+1}{2}+1, \ldots, N-1, N, 1, \ldots, \frac{N+1}{2}-2, \frac{N+1}{2}-1 \right\rangle \text{ or}$$
$$\pi = e_N'' := \left\langle \frac{N+1}{2}+1, \frac{N+1}{2}+2, \ldots, N-1, N, 1, \ldots, \frac{N+1}{2}-1, \frac{N+1}{2} \right\rangle\,.$$

More properties of the permutations e_N^*, e_N' and e_N'' along with their interpretation are discussed at the end of Chapter 2.

Let N be an even positive integer. Then

$$c_N(\pi(i), e_N(i)) + c_N(\pi(i), e_N^*(i)) = \min(\mid \pi(i) - i \mid, N - \mid \pi(i) - i \mid) + \min\left(\mid \pi(i) - \frac{N}{2} - i \mid, N - \mid \pi(i) - \frac{N}{2} - i \mid\right) = \frac{N}{2} \text{ , for } i = 1, 2, \ldots, \frac{N}{2} \text{ ,}$$

and

$$c_N(\pi(i), e_N(i)) + c_N(\pi(i), e_N^*(i)) = \min(\mid \pi(i) - i \mid, N - \mid \pi(i) - i \mid) + \min\left(\mid \pi(i) - i + \frac{N}{2} \mid, N - \mid \pi(i) - i + \frac{N}{2} \mid\right) = \frac{N}{2} \text{ , for } i = \frac{N}{2} + 1, \ldots, N \text{ .}$$

Thus,

$$d_L(\pi, e_N) + d_L(\pi, e_N^*) = \sum_{i=1}^{N} c_N(\pi(i), e_N(i)) + c_N(\pi(i), e_N^*(i)) = \sum_{i=1}^{N} \frac{N}{2} = \frac{N^2}{2} \text{ ,} \qquad (1.5)$$

for all $\pi \in \mathbf{S}_N$. The right-invariant property of d_L implies that $d_L(\pi, e_N)$ and $d_L(\pi, e_N^*)$ have the same distribution when $\pi \sim Uniform(\mathbf{S}_N)$. From this fact and (1.5), it follows that

$$P(D_L = k) = P\left(D_L = \frac{N^2}{2} - k\right) \text{ , for } k = 0, 1, \ldots, \frac{N^2}{2} \text{ , i.e.} \qquad (1.6)$$

the distribution of D_L is symmetric when N is even. Furthermore, D_L can take only even values since for even integers N

$$D_L(\pi) \equiv \sum_{i=1}^{N} \min(\mid \pi(i) - i \mid, N - \mid \pi(i) - i \mid) \equiv \sum_{i=1}^{N} \mid \pi(i) - i \mid \pmod 2 \text{ ,}$$

where "$\equiv$" is mod equality with modulus 2. Hence

$$D_L(\pi) \equiv \sum_{i=1}^{N} (\pi(i) - i) \equiv 0 \pmod 2 \text{ .}$$

1.2.1 Mean and variance of D_L under uniformity of π

The mean, variance and asymptotic distribution of D_L can be derived from the combinatorial central limit theorem (CCLT), formulated and proved by Hoeffding [32].

Theorem 1.1 (Hoeffding's CCLT). *Let $\pi \sim Uniform(\mathbf{S}_N)$ and $D(\pi) = \sum_{i=1}^{N} a_N(\pi(i), i)$, where $a_N(i, j) \in \mathbf{R}$ for $i, j = 1, 2, \ldots, N$. Then the mean and variance of D are*

$$\mathbf{E}(D) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} a_N(i, j) \qquad (1.7)$$

$$\mathbf{Var}(D) = \frac{1}{N-1} \sum_{i=1}^{N} \sum_{j=1}^{N} b_N^2(i, j) \text{ ,} \qquad (1.8)$$

where

$$b_N(i,j) = a_N(i,j) - \frac{1}{N}\sum_{g=1}^{N} a_N(g,j) - \frac{1}{N}\sum_{h=1}^{N} a_N(i,h) + \frac{1}{N^2}\sum_{g=1}^{N}\sum_{h=1}^{N} a_N(g,h)$$

for $i, j = 1,2,\ldots,N$. Furthermore, the distribution of D is asymptotically normal if

$$\lim_{N\to\infty} \frac{\max\limits_{1\le i,j\le N} b_N^2(i,j)}{\frac{1}{N}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N} b_N^2(i,j)} = 0 \ . \tag{1.9}$$

For the random variable D_L induced by Lee distance

$$\begin{aligned} a_N(i,j) = c_N(i,j) &:= \min\left(\mid i-j \mid, N - \mid i-j \mid\right) \ \text{ and} \\ b_N(i,j) = c_N(i,j) - \frac{1}{N}\sum_{g=1}^{N} c_N(g,j) &- \frac{1}{N}\sum_{h=1}^{N} c_N(i,h) + \frac{1}{N^2}\sum_{g=1}^{N}\sum_{h=1}^{N} c_N(g,h) \ . \end{aligned} \tag{1.10}$$

Let i be an arbitrary integer from 1 to N. When N is even, the quantity $c_N(i,j)$ takes on the values $0,1,\ldots,\frac{N}{2}-1,\frac{N}{2},\frac{N}{2}-1,\ldots,2,1$ as j runs from 1 to N. In the case of odd number N, $c_N(i,j)$ takes on the values $0,1,\ldots,\frac{N-1}{2},\frac{N-1}{2},\ldots,2,1$ as j runs from 1 to N. Thus the sum over j of $c_N(i,j)$ is the same for each i. By using this fact, the expression (1.7) can be simplified to

$$\mathbf{E}\left(D_L\right) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} c_N(i,j) = \frac{1}{N}\sum_{i=1}^{N}\left(0 + 2\left(\sum_{k=1}^{\frac{N}{2}-1} k\right) + \frac{N}{2}\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{N^2}{4} = \frac{N^2}{4} \ ,$$

for even integers N and

$$\mathbf{E}\left(D_L\right) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} c_N(i,j) = \frac{1}{N}\sum_{i=1}^{N}\left(0 + 2\sum_{k=1}^{\frac{N-1}{2}} k\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{N^2-1}{4} = \frac{N^2-1}{4} \ ,$$

when N is odd. Hence, the mean of D_L is given by

$$\mathbf{E}\left(D_L\right) = \left\lfloor\frac{N+1}{2}\right\rfloor\left\lfloor\frac{N}{2}\right\rfloor. \tag{1.11}$$

From (1.10) and (1.11) it follows that

$$b_N(i,j) = c_N(i,j) - \frac{1}{N}\left\lfloor\frac{N+1}{2}\right\rfloor\left\lfloor\frac{N}{2}\right\rfloor. \tag{1.12}$$

Simplifying (1.8) for even N gives

$$\begin{aligned}\mathbf{Var}\,(D_L) &= \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{N} b_N^2(i,j) = \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(c_N(i,j)-\frac{N}{4}\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left((0-N/4)^2+2\sum_{k=1}^{\frac{N}{2}-1}(k-N/4)^2+(N/2-N/4)^2\right) \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\frac{N^3+8N}{48} = \frac{N^4+8N^2}{48(N-1)}\ ,\end{aligned}$$

and

$$\begin{aligned}\mathbf{Var}\,(D_L) &= \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{N} b_N^2(i,j) = \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(c_N(i,j)-\frac{(N+1)(N-1)}{4N}\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\left(0-\frac{(N+1)(N-1)}{4N}\right)^2+2\sum_{k=1}^{\frac{N-1}{2}}\left(k-\frac{(N+1)(N-1)}{4N}\right)^2\right) \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\frac{N^4+2N^2-3}{48N} = \frac{N^4+2N^2-3}{48(N-1)}\ ,\end{aligned}$$

for odd integers N. Thus,

$$\mathbf{Var}\,(D_L) = \begin{cases}\dfrac{N^4+8N^2}{48(N-1)}, & \text{for } N \text{ even} \\ \dfrac{N^4+2N^2-3}{48(N-1)}, & \text{for } N \text{ odd.}\end{cases} \tag{1.13}$$

1.2.2 Asymptotic distribution of D_L under uniformity of π

The asymptotic distribution of D_L can be obtained from Theorem 1.1 by checking condition (1.9) for the Lee distance.

Theorem 1.2. *The distribution of the random variable D_L is asymptotically normal for $N \to \infty$ with mean and variance given by* (1.11) *and* (1.13).

Proof. Using (1.12), the numerator of (1.9) takes the form

$$\max_{1\le i,j\le N} b_N^2(i,j) = \left(0-\frac{1}{N}\left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right]\right)^2 = N^2\left(\frac{1}{16}+O\left(\frac{1}{N}\right)\right),$$

where $\lim_{N\to\infty} O\left(\frac{1}{N}\right) = 0$. From (1.13) it follows that

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} b_N^2(i,j) = \frac{N-1}{N}\mathbf{Var}\,(D_L) = N^3\left(\frac{1}{48}+O\left(\frac{1}{N}\right)\right)\ .$$

Therefore, the condition (1.9) of Theorem 1.1 is fulfilled,

$$\lim_{N\to\infty} \frac{\max\limits_{1\le i,j\le N} b_N^2(i,j)}{\frac{1}{N}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N} b_N^2(i,j)} = \lim_{N\to\infty} \frac{N^2\left(\frac{1}{16}+O\left(\frac{1}{N}\right)\right)}{N^3\left(\frac{1}{48}+O\left(\frac{1}{N}\right)\right)} = 0,$$

and the distribution of D_L is asymptotically normal. □

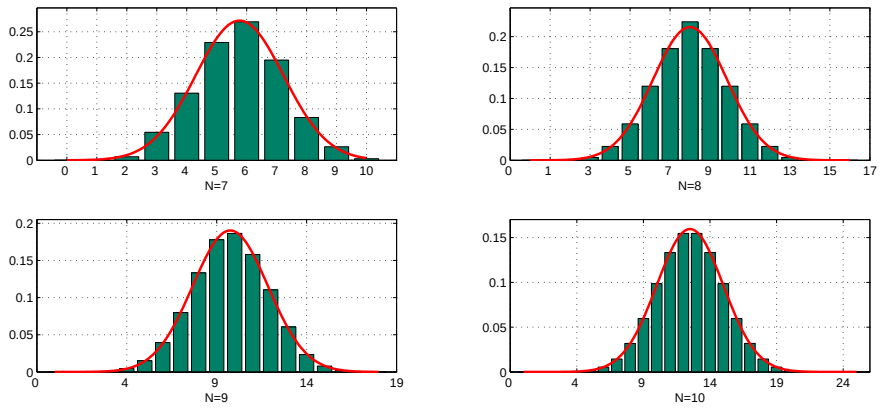


FIGURE 1.3: Probability mass function of $\left[\frac{D_L}{2}\right]$ for $N = 7,8,9,10$

When using the result of Theorem 1.2, it is important to keep in mind that D_L takes only even values for even numbers N. Hence, it is better to use normal approximation for the distribution of $\left[\frac{D_L}{2}\right]$ instead of D_L itself. Plots of the exact probability mass function of $\left[\frac{D_L}{2}\right]$ and the fitted normal curves are given on Figure 1.3 for $N = 7,8,9,10$.

Chapter 2

Probability models for rank data

One approach to analyze rank data is to construct a probability distribution $\mathbf{P}$ over the permutations in $\mathbf{S}_N$. A probability model is a family of probability distributions $\mathbf{P}$, i.e. a subset of the $(N!)$-dimensional simplex. Usually, this subset depends on some parameter vector $\vec{\theta}$, i.e. $\mathbf{P} = \mathbf{P}_{\vec{\theta}}$. The exponential family is among one of the most widely used sets of probability models for rankings and is defined by

$$\mathbf{P}_{\vec{\theta}}(\pi) = \exp\left(\sum_{i=1}^{r} \theta_i S_i(\pi) - \psi(\theta)\right) \quad \text{for } \pi \in \mathbf{S}_N, \tag{2.1}$$

where $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$ is a vector of r real parameters, $\{S_1(\cdot), S_2(\cdot), \ldots, S_r(\cdot)\}$ are real functions and $\psi(\cdot)$ is a normalizing function. Diaconis [17, p. 175] motivates the usefulness and the flexibility of this rich family. Two special cases of (2.1) are considered in the next subsections. More probability models for rankings can be found in Diaconis [17], Marden [47] and Alvo and Yu [1].

In this chapter, we compare the Distance-based probability model and the Marginals model for complete rankings. We propose an algorithm to find maximum likelihood estimates of the unknown parameters of the Latent-class Distance-based models by making use of the EM algorithm. Three rank datasets are analyzed as an illustration.

2.1 Distance-based models

In some situations, it is reasonable to assume that there is a fixed *modal* (or *antimodal*) ranking. An appropriate probability model, which assign larger (smaller) probabilities for rankings that are close to a *modal* (*antimodal*) ranking, can be defined as

$$\mathbf{P}_{\theta,\pi_0}(\pi) = \exp\left(\theta d(\pi, \pi_0) - \psi_N(\theta)\right) \quad \text{for } \pi \in \mathbf{S}_N, \tag{2.2}$$

where θ is a real parameter ($\theta \in \mathbf{R}$), $d(\cdot,\cdot)$ is a distance on $\mathbf{S}_N$, π_0 is a fixed ranking and $\psi_N(\theta)$ is a normalizing constant. When $\theta < 0$, π_0 is a *modal* ranking, for $\theta > 0$, π_0 is an *antimode*, and for $\theta = 0$, P_{θ,π_0} is the uniform distribution. Typically, the distance $d(\cdot,\cdot)$ is chosen in advance and the parameters θ and π_0 are to be estimated. The special cases of (2.2), when $d(\cdot,\cdot)$ is Kendall's tau and Spearman's rho, are first proposed by Mallows [45]. The case when the Hamming distance is being used is suggested by Fligner and Verducci [21] and more recently studied by Irurozki et al. [36]. Models based on Lee ditance are considered by Nikolov and Stoimenova [58].

The constant $\psi_N(\cdot)$ in (2.2) could be found by using the distribution of the random variable $D(\pi) = d(\pi, \pi_0)$ under uniformity of π. Let $g_N(t)$ be the moment generating function of D. Then, as shown by Fligner and Verducci [21],

$$\exp(\psi_N(\theta)) = \sum_{\pi \in \mathbf{S}_N} \exp(\theta D(\pi)) = N! \sum_{d_i} P(D = d_i) \exp(\theta d_i) = N!\, g_N(\theta) \tag{2.3}$$

and

$$\psi_N(\theta) = \log(N!\, g_N(\theta)).$$

This relation can be very useful for finding estimations of the unknown parameters θ and π_0. Theorem 1.2 and similar approximations for other distances can be applied in cases when N is too large and the exact computation of $g_N(t)$ is time-consuming.

By using the result of Theorem 1.2, the moment generating function $g_N(t)$ of the random variable D_L, induced by Lee distance, can be approximated with

$$\hat{g}_N(t) = \exp\left(t\mu + \frac{t^2 v^2}{2}\right),$$

where $\mu = \mathbf{E}(D_L)$ and $v^2 = \mathbf{Var}(D_L)$ are given in (1.11) and (1.13), respectively. Thus, for large values of N, the normalizing constant $\psi_N(\theta)$ in model (2.2) with $d = d_L$ can be approximated by

$$\hat{\psi}_N(\theta) = \log(N!\, \hat{g}_N(\theta)) = \log(N!) + \theta\mu + \frac{\theta^2 v^2}{2}. \tag{2.4}$$

The values of $\psi_N(\theta)$ and $\hat{\psi}_N(\theta)$ are given in Table 2.1 for $N \in \{4,5,6,7,8,9\}$ and $\theta \in \left\{-1, -\frac{3}{4}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{4}, 1\right\}$. It can be noticed that the percentage error, given in the last column, tends to zero as N increases or θ goes to zero. Even though the values of N in Table 2.1 are not too large, it looks reasonable to use $\hat{\psi}_N(\theta)$ as an approximation of $\psi_N(\theta)$ when $N \geq 8$ and $-\frac{3}{4} \leq \theta \leq \frac{3}{4}$.

The approximation $\hat{\psi}_N(\theta)$ can be useful in other applications of model (2.2). For example, Yu and Xu [81] propose a method for estimating the unknown parameters in a Latent-scale Distance-based model for the problem of rank aggregation. In one of the steps of their algorithm, it is required to find the value of the normalizing constant $\psi_N(\theta)$ for given value of θ. Since evaluating $\psi_N(\theta)$ by summing over all possible $N!$ rankings becomes computationally demanding for $N \geq 10$, the approximation (2.4) can be very helpful when the model is based on Lee distance.

One effective approach to extend the classical Distance-based model (2.2) is to assume that there are K latent groups (classes), $G_1, G_2, \ldots, G_K$, in the population and that the distributions of the rankings within each group can be described by (2.2), i.e.

$$\mathbf{P}_{\theta_j, \pi_{0,j}}(\pi) = \exp(\theta_j d(\pi, \pi_{0,j}) - \psi_N(\theta_j)), \quad \pi \in \mathbf{S}_N,$$

where θ_j is a real parameter and $\pi_{0,j}$ is the *modal* ranking in group G_j, for $j = 1,2,\ldots,K$. Then the overall density for this Latent-class Distance-based model is

$$\mathbf{P}_{\vec{\theta}, \vec{p}, \vec{\pi}_0}(\pi) = \sum_{j=1}^{K} p_j \exp(\theta_j d(\pi, \pi_{0,j}) - \psi_N(\theta_j)), \quad \pi \in \mathbf{S}_N, \tag{2.5}$$

N	θ	$\psi_N(\theta)$	$\hat{\psi}_N(\theta)$	**% error**	N	θ	$\psi_N(\theta)$	$\hat{\psi}_N(\theta)$	**% error**
4	-1.00	0.592	0.511	13.65%	7	-1.00	1.175	0.859	26.95%
4	-0.75	0.969	0.928	4.27%	7	-0.75	2.120	1.963	7.42%
4	-0.50	1.523	1.511	0.73%	7	-0.50	3.654	3.609	1.25%
4	0.50	5.523	5.511	0.20%	7	0.50	15.544	15.609	0.42%
4	0.75	6.969	6.928	0.59%	7	0.75	19.709	19.963	1.29%
4	1.00	8.592	8.511	0.94%	7	1.00	24.178	24.859	2.81%
5	-1.00	0.823	0.538	34.71%	8	-1.00	1.333	1.462	9.65%
5	-0.75	1.402	1.272	9.25%	8	-0.75	2.437	2.462	1.01%
5	-0.50	2.262	2.225	1.63%	8	-0.50	4.310	4.319	0.20%
5	0.50	8.192	8.225	0.41%	8	0.50	20.310	20.319	0.04%
5	0.75	10.152	10.272	1.19%	8	0.75	26.437	26.462	0.09%
5	1.00	12.237	12.538	2.46%	8	1.00	33.333	33.462	0.39%
6	-1.00	0.991	0.879	11.27%	9	-1.00	1.498	1.552	3.60%
6	-0.75	1.740	1.686	3.13%	9	-0.75	2.773	2.724	1.79%
6	-0.50	2.915	2.904	0.38%	9	-0.50	5.016	4.989	0.53%
6	0.50	11.915	11.904	0.09%	9	0.50	24.856	24.989	0.54%
6	0.75	15.240	15.186	0.36%	9	0.75	32.168	32.724	1.73%
6	1.00	18.991	18.879	0.59%	9	1.00	40.025	41.552	3.82%

TABLE 2.1: Values of $\psi_N(\theta)$ and $\hat{\psi}_N(\theta)$, induced by Lee distance

where $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$, $\vec{\pi}_0 = (\pi_{0,1}, \pi_{0,2}, \ldots, \pi_{0,K})$ and $\vec{p} = (p_1, p_2, \ldots, p_K)$ are vectors of unknown parameters. Since the probability p_j represents the proportion of the population in group G_j, for $j = 1, 2, \ldots, K$, the elements of $\vec{p}$ sum up to 1, i.e. $\sum_{j=1}^{K} p_j = 1$. More detailed description of model (2.5) can be found in *Chapter 10* of Marden [47].

2.2 Marginals model

For some rank datasets, models (2.2) and (2.5) are not rich enough to capture the full structure in the data. A submodel of (2.1), that includes the Lee distance model (2.2) with $d = d_L$, is the Marginals model,

$$\mathbf{P}_{\vec{\lambda}}(\pi) = \exp\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i^{(j)} \mathbf{I}[\pi(i) = j] - \psi(\vec{\lambda})\right) \quad \text{for } \pi \in \mathbf{S}_N, \tag{2.6}$$

where $\vec{\lambda} = \left\{\lambda_i^{(j)}\right\}_{i,j=1}^{N}$ are N^2 real parameters, $\mathbf{I}[\cdot]$ is the indicator function and $\psi(\vec{\lambda})$ is a normalizing constant. In contrast of models (2.2), the approach of constructing the Marginals model is more data-analytical in the sense that the aim of formula (2.6) is to explain the quantities

$$m_{ij} = \sum_{\pi(i)=j} \mathbf{P}_{\vec{\lambda}}(\pi), \text{ for } i, j = 1, 2, \ldots, N,$$

where the summation is over every permutation $\pi = (\pi(1), \pi(2), \ldots, \pi(N))$ in $\mathbf{S}_N$ such that $\pi(i) = j$. The matrix $M = \left\{m_{ij}\right\}_{i,j=1}^{N}$ is called *Marginals* matrix since the i-th row gives the theoretical marginal distribution of the ranks assigned to object i, and the j-th column gives

the theoretical marginal distribution of the objects given rank j. From

$$\sum_{\pi\in\mathbf{S}_N} \mathbf{P}_{\vec{\lambda}}(\pi) = 1$$

it follows that

$$\sum_{i=1}^{N} m_{ij} = 1 \quad \text{and} \quad \sum_{j=1}^{N} m_{ij} = 1.$$

Thus, there are only $(N-1)^2$ free parameters $\left\{\lambda_i^{(j)}\right\}_{i,j=1}^{N}$ of the Marginals model. An extension of model (2.6) with more free parameters is proposed by Diaconis [18] as an application of spectral analysis to permutation data.

The Marginals model is first proposed by Verducci [77] under the name *quasi-independence model*. Distance-based models (2.2) with Hoeffding distances (that include Spearman's footrule, Spearman's rho, Hamming distance and Lee distance) are also Marginals models (2.6). For example, model (2.2) induced by the Hamming distance coincides with (2.6) when

$$\lambda_i^{(j)} = \theta \mathbf{I}[j = \pi_0(i)], \text{ for } i, j = 1,2,\ldots,N.$$

If the model is based on Lee distance, then

$$\lambda_i^{(j)} = \theta \min(\mid j-\pi_0(i) \mid, N- \mid j-\pi_0(i) \mid), \text{ for } i, j = 1,2,\ldots,N.$$

Let us denote the Marginals matrix under the models (2.2) by $M(\theta,N)$. If the used distance is right-invariant, then without loss of generality it can be assumed that $\pi_0 = e_N$. Varying the permutation π_0 is equivalent to reordering the rows of the matrix. The elements of $M(\theta,N)$ can be expressed as

$$m_{ij}(\theta,N) = \sum_{\pi(i)=j} \mathbf{P}_{\theta,e_N}(\pi), \text{ for } i, j = 1,2,\ldots,N, \tag{2.7}$$

where $\mathbf{P}_{\theta,e_N}(\pi)$ is defined in (2.2) for $\pi_0 = e_N$. For $\theta = 0$ the matrix $M(\theta,N)$ has equal elements, i.e. $m_{ij}(0,N) = 1/N$ for $i, j = 1,2,\ldots,N$, and is associated with the uniform model. When $\theta \to -\infty$, the matrix $M(\theta,N)$ converges to the identity matrix $I_{N\times N}$, which corresponds to the identity ranking, i.e. $P_{\theta,e_N}(e_N) = 1$.

Notice from (2.7) that for Lee distance $m_{ij}(\theta,N)$ does not depend on both i and j, but only on the value of $c_N(i,j) := \min(\mid i-j \mid, N- \mid i-j \mid)$. Thus, there are only $\left[\frac{N+2}{2}\right]$ different elements of the Marginals matrix $M(\theta,N)$. For large values of N, we can use the following asymptotic approximation.

Theorem 2.1. *Let $M(\theta,N) = \left\{m_{ij}(\theta,N)\right\}_{i,j=1}^{N}$ be the Marginals matrix, based on Lee distance. Then*

$$m_{ij}(\theta,N)\frac{N}{\exp\left(\theta\mu + \frac{\theta^2\nu^2}{2}\right)} \xrightarrow[N\to\infty]{} 1, \quad \textit{for } i, j = 1,2,\ldots,N,$$

where

$$\mu = \frac{Nc_N(i,j)}{N-1} - \frac{1}{N-1}\left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right]$$

and

$$\nu^2 = \begin{cases} \dfrac{2N^2\left(c_N(i,j)\right)^2 - N^3 c_N(i,j)}{2(N-2)(N-1)^2} - \dfrac{N^2\left(N^2-2N+4\right)}{48(N-1)^2}, & \text{for } N \text{ even} \\ \dfrac{2N^2\left(c_N(i,j)\right)^2 - N\left(N^2-1\right)c_N(i,j)}{2(N-2)(N-1)^2} - \dfrac{N(N+1)(N-3)}{48(N-2)}, & \text{for } N \text{ odd.} \end{cases}$$

The proof of Theorem 2.1 is given in Appendix E. More properties of $M(\theta,N)$ based on other distances are studied in Chapter 4, where the Marginals matrix is called error probability matrix and denoted by $\mathbf{Q}$.

2.3 Statistical inference

Let $\pi^* = \left(\pi^1, \pi^2, \ldots, \pi^n\right)$ be a sample of n complete rankings and $\ell(\theta, \pi_0, \pi^*)$ be the loglikelihood function of model (2.2),

$$\ell(\theta, \pi_0, \pi^*) = \theta S(\pi_0, \pi^*) - n\psi_N(\theta),$$

where $S(\pi_0, \pi^*) = \sum_{k=1}^{n} d\left(\pi^k, \pi_0\right)$. Then for testing the hypothesis of uniform model ($\theta = 0$) against the alternative that $\theta \neq 0$, Marden [47, p. 144] suggested the likelihood ratio statistic,

$$LRS_d = 2\left[\ell(\hat{\theta}, \hat{\pi}_0, \pi^*) - \ell(0, \hat{\pi}_0, \pi^*)\right],$$

where $\left(\hat{\theta}, \hat{\pi}_0\right)$ are the maximum likelihood estimates (MLE's) of (θ, π_0). Various techniques for finding the MLE's $\left(\hat{\theta}, \hat{\pi}_0\right)$ are given in Marden [47, Chapter 6]. The likelihood function in the case of Latent-class model (2.5) is more complicated and it is not possible to find estimates of the unknown parameters $\vec{\theta}$, $\vec{\pi}_0$ and $\vec{p}$ in a similar way. However, an algorithm for finding MLE's of the parameters in model (2.5) is proposed in Section 2.4.

Notice that $S(\pi_0, \pi^*)$ is sufficient statistic for the Distance-based models (2.2) and gives a great reduction of the data. For the Marginals model (2.6) a sufficient statistic is the sample Marginals matrix $\hat{M} = \left\{\hat{m}_{ij}\right\}_{i,j=1}^{N}$ defined by

$$\hat{m}_{ij} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{I}\left[\pi^k(i) = j\right], \quad \text{for } i, j = 1, 2, \ldots, N.$$

Thus, the loglikelihood function of model (2.6) is

$$\ell(\lambda, \pi^*) = \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i^{(j)} \hat{m}_{ij} - n\psi(\lambda),$$

and

$$LRS_m = 2\left[\ell(\hat{\lambda}, \pi^*) - \ell(\vec{0}, \pi^*)\right]$$

can be used for testing the hypothesis of uniform model ($\lambda = \vec{0}$). The MLE's $\hat{\lambda} = \left\{\hat{\lambda}_i^{(j)}\right\}_{i,j=1}^{N}$ can be found by using the Newton-Raphson method or an algorithm based on minimum majorization decomposition and proposed by Verducci [78]. A general method for estimating

the unknown parameters of the exponential family models (2.1) is considered by Mukherjee [54].

For Distance-based models (2.2) that are submodels of the Marginals model (2.6), we have that $LRS_d \leq LRS_m$. The statistic

$$LRS_{diff} = LRS_m - LRS_d$$

can be used to test if LRS_m is significantly improved compared to LRS_d. If the *modal* ranking π_0 in model (2.2) is known and not estimated in LRS_d, then LRS_{diff} has chi-square distribution with $(N-1)^2 - 1$ degrees of freedom. This result is an implication of more general theorem for nested models, the proof of which is given in *Section 5.14.3* of Marden [47]. Another possibility to compare models (2.2) and (2.6) is to study the difference between $\hat{M}$, the sample Marginals matrix, and the matrix $M\left(\hat{\theta}, N\right)$, based on the distribution of model (2.2). Comparisons between the two models are presented in Section 2.5.

Let $f(\pi)$ be the frequency of a given permutation $\pi \in \mathbf{S}_N$, i.e. $f(\pi)$ is the number of observations that are equal to π. Then the empirical probability for π is $\dfrac{f(\pi)}{n}$ and a quantity, that measures the total nonuniformity of the data, could be defined as

$$TNU = 2 \sum_{\pi \in \mathbf{S}_N} f(\pi) \left[\log\left(\frac{f(\pi)}{n}\right) - \log\left(\frac{1}{N!}\right) \right].$$

One can use TNU together with LRS of a fitted model in order to test the goodness-of-fit of the model. Similarly to the multiple correlation coefficient in linear regression, Marden [47, p. 144] considered the coefficient

$$R^2 = \frac{LRS}{TNU}, \tag{2.8}$$

which measures the percentage of nonuniformity in the data that is explained by the fitted model. When $R^2 = 1$ the model exactly fits the data, and $R^2 = 0$ if it performs no better than the uniform model.

2.4 EM Estimation

It is not possible to estimate the unknown parameters $\vec{\theta}$, $\vec{p}$ and $\vec{\pi}_0$ in model (2.5) directly. However, the Expectation-Maximization (EM) algorithm proposed by Dempster et al. [15] can be applied. A complete description of the concept of the EM iteration procedure is given in [52]. Croon and Luijkx [13] considered similar algorithm in the case of Latent Models when $\vec{\pi}_0$ is known or can be approximated by other methods, for example the *"K-means" clustering* presented in Chapter 3.

2.4.1 Classical EM algorithm

The aim of the algorithm is to find the expected value of the group loglikelihood function $\ell(\vec{\theta}, \vec{p}, \vec{\pi}_0, \vec{\pi}^*, G)$ for given initial approximations of $\vec{\theta}$, $\vec{p}$ and $\vec{\pi}_0$ (*E-step*). This expectation is usually denoted by

$$Q^{(t)}\left(\vec{\theta}, \vec{p}, \vec{\pi}_0, \vec{\pi}^*\right) = \mathbf{E}_{G|\vec{\theta}^{(t)}, \vec{p}^{(t)}, \vec{\pi}_0{}^{(t)}, \vec{\pi}^*} \left[\ell(\vec{\theta}, \vec{p}, \vec{\pi}_0, \vec{\pi}^*, G) \right]$$

for some initial values $\vec{\theta}^{(t)}, \vec{p}^{(t)}$ and $\vec{\pi}_0{}^{(t)}$. From (2.5) it follows that

$$Q^{(t)}\left(\vec{\theta},\vec{p},\vec{\pi}_0,\vec{\pi}^*\right)=\sum_{i=1}^{n}\sum_{j=1}^{K}\frac{p_j^{(t)}P_{\theta_j^{(t)},\pi_{0,j}^{(t)}}\left(\pi^i\right)}{\sum_{s=1}^{K}p_s^{(t)}P_{\theta_s^{(t)},\pi_{0,s}^{(t)}}\left(\pi^i\right)}\left[\log(p_j)+\theta_j d\left(\pi^i,\pi_{0,j}\right)-\psi(\theta_j)\right].$$

The next step is to maximize $Q^{(t)}\left(\vec{\theta},\vec{p},\vec{\pi}_0,\vec{\pi}^*\right)$ with respect to $\vec{\theta}$, $\vec{p}$ and $\vec{\pi}_0$ (*M-step*), i.e.

$$\left(\vec{\theta}^{(t+1)},\vec{p}^{(t+1)},\vec{\pi}_0{}^{(t+1)}\right)=\underset{\left(\vec{\theta},\vec{p},\vec{\pi}_0\right)}{\operatorname{argmax}}\left\{Q^{(t)}\left(\vec{\theta},\vec{p},\vec{\pi}_0,\vec{\pi}^*\right)\right\}. \tag{2.9}$$

The optimal solution of (2.9) with respect to $\vec{p}$ is

$$p_j^{(t+1)}=\frac{1}{n}\sum_{i=1}^{n}\frac{p_j^{(t)}P_{\theta_j^{(t)},\pi_{0,j}^{(t)}}\left(\pi^i\right)}{\sum_{s=1}^{K}p_s^{(t)}P_{\theta_s^{(t)},\pi_{0,s}^{(t)}}\left(\pi^i\right)},\quad \text{for } j=1,2,\ldots,K, \tag{2.10}$$

which is independent of the values of $\vec{\theta}$ and $\vec{\pi}_0$.

The optimal value θ_j, for $j=1,2,\ldots,K$, is the solution of the equation

$$\sum_{i=1}^{n}\frac{p_j^{(t)}P_{\theta_j^{(t)},\pi_{0,j}^{(t)}}\left(\pi^i\right)}{\sum_{s=1}^{K}p_s^{(t)}P_{\theta_s^{(t)},\pi_{0,s}^{(t)}}\left(\pi^i\right)}\left[d\left(\pi^i,\pi_{0,j}\right)-\psi_N'(\theta_j)\right]=0, \tag{2.11}$$

which depends on $\pi_{0,j}$. Therefore, the values of $\vec{\theta}=(\theta_1,\theta_2,\ldots,\theta_K)$ should be calculated from (2.11) for every possible choice of $\vec{\pi}_0=(\pi_{0,1},\pi_{0,2},\ldots,\pi_{0,K})$, where $\pi_{0,j}\in\mathbf{S_N}$ for $j=1,2,\ldots,K$ and $\pi_{0,j}\neq\pi_{0,s}$ for $j\neq s$. Then the optimal solution $\left(\vec{\theta}^{(t+1)},\vec{\pi}_0{}^{(t+1)}\right)$ is the pair $\left(\vec{\theta},\vec{\pi}_0\right)$ that maximizes $Q^{(t)}\left(\vec{\theta},\vec{p}^{(t+1)},\vec{\pi}_0,\vec{\pi}^*\right)$.

After $\vec{\theta}^{(t+1)},\vec{p}^{(t+1)}$ and $\vec{\pi}_0{}^{(t+1)}$ are obtained, they are substituted as initial approximations in the *E-step* for calculating the new values of $Q^{(t+1)}\left(\vec{\theta},\vec{p},\vec{\pi}_0,\vec{\pi}^*\right)$ and so on. This procedure continues until some optimal criteria are met, for example the change of the likelihood function is relatively small or a prefixed number of iterations is reached.

The monotonicity and convergence of the described EM algorithm follows directly from *Theorem 3.2* in *Chapter 3* of McLachlan and Krishnan [52]. However, the convergence rate strongly depends on the initial values $\vec{\theta}^{(0)},\vec{p}^{(0)}$ and $\vec{\pi}_0{}^{(0)}$. It looks reasonable to assume that all elements of $\vec{p}$ are equal, i.e. $p_j^{(0)}=\frac{1}{K}$ for $j=1,2,\ldots,K$. The initial point $\vec{\pi}_0{}^{(0)}$ could be taken as a combination of permutations in $\mathbf{S_N}$ for which the empirical probability is large (*modal* rankings) or close to zero (*antimodes*). From the empirical experience it seems that $\theta_j^{(0)}=\frac{1}{2}$ is a good initial approximation, when the corresponding ranking $\pi_{0,j}^{(0)}$ is *modal*, and $\theta_j^{(0)}=-\frac{1}{2}$ when $\pi_{0,j}^{(0)}$ is an *antimode*.

2.4.2 Generalized EM algorithm

Since there are $\frac{N!}{(N-K)!}$ possible choices for the values of $\vec{\pi}_0$ and it is necessary to use some numerical method, for example Newton-Raphson method, to find the corresponding values of $\vec{\theta}^{(t+1)} = (\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_K^{(t+1)})$, the algorithm described in the previous subsection requires solving $\frac{K(N!)}{(N-K)!}$ equations of the form (2.11) at each iteration. To simplify the complexity of the procedure, a generalized version of the EM algorithm can be applied. Condition (2.9) for $\left(\vec{\theta}^{(t+1)}, \vec{p}^{(t+1)}, \vec{\pi}_0{}^{(t+1)}\right)$ could be relaxed and replaced by

$$Q^{(t)}\left(\vec{\theta}^{(t+1)}, \vec{p}^{(t+1)}, \vec{\pi}_0{}^{(t+1)}, \vec{\pi}^*\right) \geq Q^{(t)}\left(\vec{\theta}^{(t)}, \vec{p}^{(t)}, \vec{\pi}_0{}^{(t)}, \vec{\pi}^*\right). \tag{2.12}$$

It can be shown, see McLachlan and Krishnan [52, p. 78], that (2.12) is sufficient to ensure the monotonicity of the algorithm, i.e. the likelihood is not decreased after an EM iteration. Thus $\vec{\pi}_0{}^{(t+1)} = (\pi_{0,1}^{(t+1)}, \pi_{0,2}^{(t+1)}, \ldots, \pi_{0,K}^{(t+1)})$ can be defined as

$$\pi_{0,j}^{(t+1)} = \underset{\pi \in \mathbf{S_N}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \frac{p_j^{(t)} P_{\theta_j^{(t)}, \pi_{0,j}^{(t)}}\left(\pi^i\right)}{\sum_{s=1}^{K} p_s^{(t)} P_{\theta_s^{(t)}, \pi_{0,s}^{(t)}}\left(\pi^i\right)} \theta_j^{(t)} d\left(\pi^i, \pi\right) \right\}, \text{ for } j = 1, 2, \ldots, K, \tag{2.13}$$

where if $\pi_{0,j}^{(t+1)}$ is in $\left\{\pi_{0,j+1}^{(t)}, \pi_{0,j+2}^{(t)}, \ldots, \pi_{0,K}^{(t)}\right\}$ or $\left\{\pi_{0,1}^{(t+1)}, \pi_{0,2}^{(t+1)}, \ldots, \pi_{0,j-1}^{(t+1)}\right\}$, then $\pi_{0,j}^{(t+1)} = \pi_{0,j}^{(t)}$. The corresponding value of $\theta_j^{(t+1)}$ can be found as the solution of

$$\sum_{i=1}^{n} \frac{p_j^{(t)} P_{\theta_j^{(t)}, \pi_{0,j}^{(t)}}\left(\pi^i\right)}{\sum_{s=1}^{K} p_s^{(t)} P_{\theta_s^{(t)}, \pi_{0,s}^{(t)}}\left(\pi^i\right)} \left[d\left(\pi^i, \pi_{0,j}^{(t+1)}\right) - \psi_N'(\theta_j^{(t+1)}) \right] = 0, \text{ for } j = 1, 2, \ldots, K. \tag{2.14}$$

Proposition 2.1. *Let $\vec{p}^{(t+1)}$, $\vec{\pi}_0{}^{(t+1)}$ and $\vec{\theta}^{(t+1)}$ are given by (2.10), (2.13) and (2.14) respectively. Then condition (2.12) holds.*

Proof. Since $\vec{p}^{(t+1)}$ is the solution of (2.9) and is independent of $\left(\vec{\theta}, \vec{\pi}_0\right)$,

$$Q^{(t)}\left(\vec{\theta}^{(t)}, \vec{p}^{(t+1)}, \vec{\pi}_0{}^{(t)}, \vec{\pi}^*\right) \geq Q^{(t)}\left(\vec{\theta}^{(t)}, \vec{p}^{(t)}, \vec{\pi}_0{}^{(t)}, \vec{\pi}^*\right).$$

From (2.13) and the definition of $Q^{(t)}$ it follows that

$$\vec{\pi}_0{}^{(t+1)} = \underset{\vec{\pi}_0}{\operatorname{argmax}} \left\{ Q^{(t)}\left(\vec{\theta}^{(t)}, \vec{p}, \vec{\pi}_0, \vec{\pi}^*\right) \right\},$$

for every vector $\vec{p}$. Thus,

$$Q^{(t)}\left(\vec{\theta}^{(t)}, \vec{p}^{(t+1)}, \vec{\pi}_0{}^{(t+1)}, \vec{\pi}^*\right) \geq Q^{(t)}\left(\vec{\theta}^{(t)}, \vec{p}^{(t+1)}, \vec{\pi}_0{}^{(t)}, \vec{\pi}^*\right).$$

Finally,

$$Q^{(t)}\left(\vec{\theta}^{(t+1)}, \vec{p}^{(t+1)}, \vec{\pi}_0{}^{(t+1)}, \vec{\pi}^*\right) \geq Q^{(t)}\left(\vec{\theta}^{(t)}, \vec{p}^{(t+1)}, \vec{\pi}_0{}^{(t+1)}, \vec{\pi}^*\right),$$

since from (2.14) we have that $\vec{\theta}^{(t+1)}$ is the local maximum of $Q^{(t)}$ when $\vec{\pi}_0 = \vec{\pi}_0{}^{(t+1)}$. □

Since condition (2.12) is sufficient only for convergence of the algorithm to a local maximum of the likelihood function, the resulting points $\vec{\theta}, \vec{p}$ and $\vec{\pi}_0$ could differ from the actual MLE's. When the parameter $\vec{\pi}_0$ is fixed there is a convergence of the EM sequence to a stationary point, see [13]. Therefore, it is recommended to run the generalized EM procedures several times with different initial approximations of $\vec{\pi}_0$. One possibility is to simulate various values from the set of initial points $\vec{\pi}_0{}^{(0)}$ as described in Subsection 2.4.1.

2.4.3 Simulation study

In this study, a comparison between Latent-class models based on different distances is made. The comparison is constructed from 800 Monte Carlo simulations of data samples with size $n = 1000$ from model (2.5) for $N = 4$, i.e. 100 data samples for each of the listed eight distances in Section 1.1. The used theoretical parameters are: $K = 3$, $\vec{p} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, $\pi_{0,1} = \langle 1,2,3,4\rangle$, $\pi_{0,2} = \langle 4,3,2,1\rangle$, $\pi_{0,3} = \langle 3,1,4,2\rangle$ and $\vec{\theta} = \left(-1, -\frac{1}{2}, \frac{1}{4}\right)$.

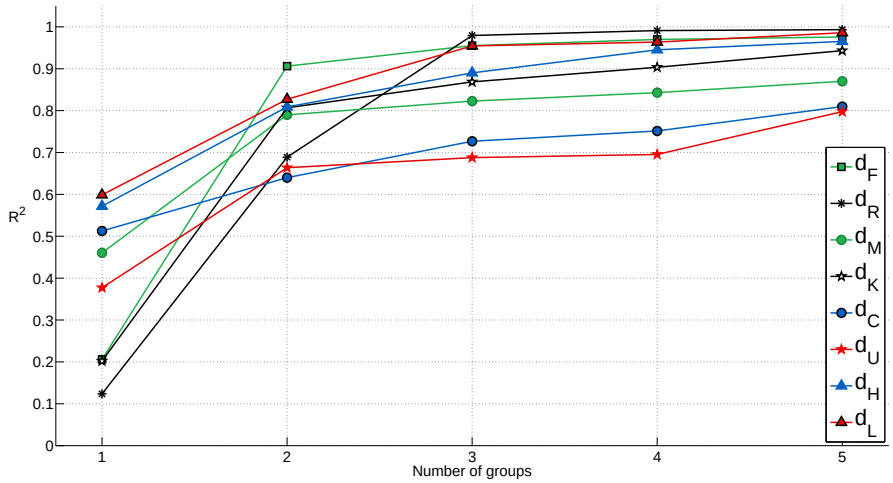


FIGURE 2.1: The average percentage of nonuniformity R^2 for models based on different distances

By applying the proposed EM algorithm in Subsection 2.4.2, the values of the percentage of nonuniformity R^2, defined by (2.8), are estimated for every sample and in cases when the number of the underlying groups K is from 1 to 5. The average values of R^2 for each metric and for each K are presented on Figure 2.1. As it is expected, there is a significant improvement in the goodness of fit of the models when $K \geq 2$. Models base on d_R, d_F and d_L explain most of the nonuniformity of the data for $K \geq 3$. Even when $K = 1$, models based on d_H and d_L perform well for the chosen theoretical parameters $\vec{p}$, $\vec{\pi}_0$ and $\vec{\theta}$.

2.5 Illustrative examples

The goal of Subsection 2.5.1 is to compare the classical Distance-based models (2.2) induced by Hamming distance, Lee distance and Kendall's tau to the Marginals model (2.6). In Subsection 2.5.2 it is illustrated how the contrast between the model (2.2) based on Lee distance and the Marginals model (2.6) can be explored only through analyzing the sample Marginals matrix and the Marginals matrix induced by Lee distance. An application of models (2.2)

and (2.5) to the popular APA election data are given in Subsection 2.5.3. The section ends with a short discussion about the situations, where the Lee distance is most appropriate. The datasets given in Subsection 2.5.1 and Subsection 2.5.2 are available at PrefLib [49], an online library of datasets concerning preferences. The APA election data in Subsection 2.5.3 can be found in Diaconis [17, p. 96].

2.5.1 Pictures of dots

Mao et al. [46] consider the human computation problem of counting pseudo-randomly distributed dots in images as a proxy for noisy comparisons of items in ranking tasks. Each voting task in their study involve sorting four pictures from fewest dots to most dots. One of the used datasets consists of $n = 794$ complete rankings of $N = 4$ types of pictures with 200, 205, 210 and 215 dots. In this subsection, models (2.2) and (2.6) are fitted to the data and compared in regard to their explanatory power.

The settings of the problem suggest that the *modal* ranking in model (2.2) is known and equals the identity permutation, i.e. $\pi_0 = \langle 1,2,3,4 \rangle$. The total nonuniformity of the data is $TNU = 248.55$. The MLE of the unknown parameter θ in (2.2) for the Distance-based models induced by Hamming distance, Lee distance and Kendall's tau are $\hat{\theta}_H = -0.3475$, $\hat{\theta}_L = -0.2654$ and $\hat{\theta}_K = -0.3549$, respectively. The corresponding values of LRS_d are $LRS_H = 120.75$, $LRS_L = 153.89$ and $LRS_K = 207.92$, which explain $R^2_H = 0.4859$, $R^2_L = 0.6192$ and $R^2_K = 0.8366$ of the TNU. These results imply that the fitted models perform much better than the uniform model. However, the obtained statistics can be improved by fitting the Marginals model (2.6) to the sample Marginals matrix $\hat{M}$. The next equation gives the values of $\hat{M}$ in percentages, i.e. $100 \times \hat{M}$.

$$\hat{M} = \begin{pmatrix} 42.57 & 26.45 & 14.86 & 16.12 \\ 27.08 & 28.97 & 25.06 & 18.89 \\ 16.50 & 25.69 & 31.36 & 26.45 \\ 13.85 & 18.89 & 28.72 & 38.54 \end{pmatrix}$$

The Marginals model fits fairly well with $LRS_m = 235.31$ and $R^2_m = 0.9468$. All differences $LRS_m - LRS_H = 114.56$, $LRS_m - LRS_L = 81.42$ and $LRS_m - LRS_K = 27.39$ are significant, since the gain in the explanatory power is based on $(N-1)^2 - 1 = 8$ degrees of freedom with critical value 20.09 at significance level of 0.01.

Another way to compare the Distance-based models and the Marginals model is to investigate the differences between $\hat{M}$ and the Marginals matrices induced by the distances. The values of M_H, M_L and M_K, based on $\hat{\theta}_H$, $\hat{\theta}_L$ and $\hat{\theta}_K$ respectively, are given in percentages.

$$M_H = \begin{pmatrix} 35.36 & 21.55 & 21.55 & 21.55 \\ 21.55 & 35.36 & 21.55 & 21.55 \\ 21.55 & 21.55 & 35.36 & 21.55 \\ 21.55 & 21.55 & 21.55 & 35.36 \end{pmatrix}, \quad M_L = \begin{pmatrix} 35.21 & 23.83 & 17.14 & 23.83 \\ 23.83 & 35.21 & 23.83 & 17.14 \\ 17.14 & 23.83 & 35.21 & 23.83 \\ 23.83 & 17.14 & 23.83 & 35.21 \end{pmatrix},$$

$$M_K = \begin{pmatrix} 39.40 & 27.63 & 19.38 & 13.59 \\ 27.63 & 28.51 & 24.48 & 19.38 \\ 19.38 & 24.48 & 28.51 & 27.63 \\ 13.59 & 19.38 & 27.63 & 39.40 \end{pmatrix}, \quad Z = \begin{pmatrix} 2.02 & -0.85 & -3.98 & 2.17 \\ -0.39 & 0.32 & 0.42 & -0.39 \\ -2.44 & 0.87 & 1.92 & -0.85 \\ 0.24 & -0.39 & 0.76 & -0.56 \end{pmatrix}.$$

The matrix Z is the difference between $\hat{M}$ and M_K, divided by the standard error of M_K under model (2.2) based on Kendall's tau. There are significant elements in the first and the third rows of Z, meaning that the images with 200 dots are ranked higher and the images with 215 dots are ranked lower than the expected by the Distance-based model. Nevertheless, the model based on Kendall's tau fits well and the Marginals model is not overwhelmingly better.

Notice that the variety of elements in $\hat{M}$ can't be explained by the two-element matrix M_H and the three-element matrix M_L. This is the reason for the lack of explanatory power of the models based on Hamming distance and Lee distance in this example. As illustrated in the next subsection, the number of elements in M_L increases with N and the model induced by Lee distance becomes more flexible.

2.5.2 Courses

Skowron et al. [72] study the course preferences of $n = 146$ students at AGU University of Science and Technology, Krakow, in 2003. Each student provided a complete rank ordering over 9 courses. Since *Course 9* has rank 1 in every observation, only the preferences over the remaining $N = 8$ courses are considered in this subsection.

The values of LRS_d for the models based on Hamming distance, Lee distance and Kendall's tau are $LRS_H = 124.11$, $LRS_L = 285.01$ and $LRS_K = 118.65$, respectively. The corresponding R^2 coefficients are $R^2_H = 0.0725$, $R^2_L = 0.1665$ and $R^2_K = 0.0693$. These results show that the one-parameter distance models fits better than the uniform model, but are not rich enough to capture the structure in the data.

The sample Marginals matrix, given in percentages, is

$$\hat{M} = \begin{pmatrix} 11.64 & 2.05 & 3.42 & 0.68 & 17.12 & 9.59 & 19.86 & 35.62 \\ 28.77 & 5.48 & 3.42 & 3.42 & 13.70 & 17.81 & 17.81 & 9.59 \\ 31.51 & 30.14 & 4.79 & 6.85 & 8.90 & 6.16 & 7.53 & 4.11 \\ 11.64 & 28.77 & 19.86 & 5.48 & 8.90 & 10.96 & 7.53 & 6.85 \\ 2.05 & 20.55 & 24.66 & 13.01 & 15.07 & 10.96 & 9.59 & 4.11 \\ 12.33 & 9.59 & 30.82 & 31.51 & 6.85 & 6.16 & 2.05 & 0.68 \\ 1.37 & 2.05 & 8.90 & 21.23 & 11.64 & 17.12 & 13.01 & 24.66 \\ 0.68 & 1.37 & 4.11 & 17.81 & 17.81 & 21.23 & 22.60 & 14.38 \end{pmatrix}.$$

The goodness-of-fit statistics for the Marginals model are $LRS_m = 554.46$ and $R^2_m = 0.3239$ obtained by estimating $(N-1)^2 = 49$ free parameters. Thus, there are significant differences between LRS_m and the LRS_H, LRS_L and LRS_K, based on the three distances. In this case, the critical difference value at level 0.01 is 73.68 on $(N-1)^2 - 1 = 48$ degrees of freedom. Even though, model (2.6) performs considerably better than model (2.2), the most part of $TNU = 1711.69$ is still not explained. As in the example in the next section, when the classical Distance-based model (2.2) induced by Lee distance has larger R^2 coefficient compared to the models induced by other distances, it is worth to consider the Latent-class model (2.5). In the case when it is assumed that there are $K = 2$ latent groups in model (2.5), the R^2 coefficient of the fitted models based on Hamming distance, Lee distance and Kendall's tau are $R^2_{H,2} = 0.1396$, $R^2_{L,2} = 0.2036$ and $R^2_{K,2} = 0.1013$, respectively. Thus, the explanatory power of model (2.2) can be significantly improved by modifying it to model (2.5) with only 3 additional unknown parameters (for $K = 2$).

For large values of N, the calculation of LRS_m could require a lot of time and computer resources, since there are $(N-1)^2$ unknown parameters in model (2.6) that have to be estimated. In these cases, the Marginals matrices can reveal some differences between models (2.2) and (2.6) more efficiently. The values of M_L, induced by the Lee distance model with parameters $\pi_0 = \langle 8,5,1,2,3,4,6,7\rangle$ and $\hat{\theta}_L = -0.3790$, are given below. Just by comparing the matrices $\hat{M}$ and M_L, it can be concluded that the Marginals model fits significantly better than the model based on Lee distance.

$$M_L = \begin{pmatrix} 10.57 & 6.93 & 4.64 & 6.93 & 16.52 & 10.57 & 16.52 & 27.33 \\ 16.52 & 10.57 & 6.93 & 4.64 & 27.33 & 6.93 & 10.57 & 16.52 \\ 27.33 & 16.52 & 10.57 & 6.93 & 16.52 & 4.64 & 6.93 & 10.57 \\ 16.52 & 27.33 & 16.52 & 10.57 & 10.57 & 6.93 & 4.64 & 6.93 \\ 10.57 & 16.52 & 27.33 & 16.52 & 6.93 & 10.57 & 6.93 & 4.64 \\ 6.93 & 10.57 & 16.52 & 27.33 & 4.64 & 16.52 & 10.57 & 6.93 \\ 4.64 & 6.93 & 10.57 & 16.52 & 6.93 & 27.33 & 16.52 & 10.57 \\ 6.93 & 4.64 & 6.93 & 10.57 & 10.57 & 16.52 & 27.33 & 16.52 \end{pmatrix}$$

The computation of the matrix M_L itself could be time-consuming for large values of N. However, M_L can be approximated by using the asymptotic result from Theorem 2.1. The elements of $\tilde{M}_L$ are the approximated values of M_L, divided by the sum of the first row in the approximation, i.e. $\tilde{M}_L$ is normalized by keeping the proportions of its elements.

$$\tilde{M}_L = \begin{pmatrix} 10.61 & 6.99 & 4.75 & 6.99 & 16.61 & 10.61 & 16.61 & 26.85 \\ 16.61 & 10.61 & 6.99 & 4.75 & 26.85 & 6.99 & 10.61 & 16.61 \\ 26.85 & 16.61 & 10.61 & 6.99 & 16.61 & 4.75 & 6.99 & 10.61 \\ 16.61 & 26.85 & 16.61 & 10.61 & 10.61 & 6.99 & 4.75 & 6.99 \\ 10.61 & 16.61 & 26.85 & 16.61 & 6.99 & 10.61 & 6.99 & 4.75 \\ 6.99 & 10.61 & 16.61 & 26.85 & 4.75 & 16.61 & 10.61 & 6.99 \\ 4.75 & 6.99 & 10.61 & 16.61 & 6.99 & 26.85 & 16.61 & 10.61 \\ 6.99 & 4.75 & 6.99 & 10.61 & 10.61 & 16.61 & 26.85 & 16.61 \end{pmatrix}$$

Similarly to the results for $\hat{\psi}_N(\theta)$ in Table 2.1, even when N is not too large ($N = 8$) the approximation matrix $\tilde{M}_L$ looks reasonably close to M_L and could be used to compare the

Marginals model and the Lee distance model.

2.5.3 APA election

In 1980, the American Psychological Association (APA) conducted an election in which $N = 5$ candidates were running for president and voters were asked to rank order all of the candidates. The complete rankings of $n = 5738$ voters are listed in Diaconis [17, p. 96]. The average ranks received by the five candidates A, B, C, D and E are 2.84, 3.16, 2.92, 3.09, and 2.99, respectively, and the total nonuniformity of the data is $TNU = 1717.51$. The results obtained by fitting models (2.2) based on eight commonly used distances are given in Table 2.2.

Distance name	**Notation**	$\hat{\theta}$	$\hat{\pi}_0$	**Ordering**	*LRS*	R^2
Spearman's footrule	d_F	0.0828	$\langle 5,1,3,2,4\rangle$	BDCEA	282.26	0.1643
Spearman's rho	d_R	-0.0163	$\langle 1,5,2,4,3\rangle$	ACEDB	150.78	0.0878
Kendall's tau	d_K	-0.0722	$\langle 1,5,2,4,3\rangle$	ACEDB	124.28	0.0723
Chebyshev metric	d_M	-0.2639	$\langle 1,5,2,4,3\rangle$	ACEDB	379.54	0.2210
Cayley distance	d_C	-0.2483	$\langle 2,3,1,5,4\rangle$	CABED	304.21	0.1771
Ulam distance	d_U	-0.2505	$\langle 2,3,1,5,4\rangle$	CABED	181.52	0.1057
Hamming distance	d_H	0.2437	$\langle 5,1,3,2,4\rangle$	BDCEA	290.16	0.1689
Lee distance	d_L	0.1656	$\langle 5,1,3,2,4\rangle$	BDCEA	524.39	0.3053

TABLE 2.2: Results of fitting model (2.2) to the APA data

All models explain less than a third of the nonuniformity, where the model based on d_L (Lee distance) has the highest $R^2_L = 0.3053$, and the lowest $R^2_K = 0.0723$ is obtained when using d_K (Kendall's tau). The estimated *modal* rankings (antimodes for $\hat{\theta} > 0$) are given in the forth column, while the corresponding orderings are in the fifth. The orderings of d_R, d_M and d_K coincide with the *modal* ordering based on the average ranks. There are definite camps within APA: candidates A and C are research psychologists, D and E are clinical psychologists, and B is a community psychologist. These groups can also be noticed in the orderings of d_R, d_M and d_K, where the group {A,C} is ranked lower than {D,E}, which are followed by candidate B. In the orderings of d_C and d_U, candidate B separates {A,C} and {D,E}. Since $\hat{\theta} > 0$ for d_F, d_H and d_L, their orderings corresponds to *antimodal* rankings. Some properties of d_L can be helpful to interpret the ordering BDCEA. Since $N = 5$ is odd, from Section 1.2 it follows that there are two "opposite" (*modal*) orderings – EABDC and CEABD. In the first one, candidates {D,E} have lower ranks compared to {A,C}, while in the second one it is quite the opposite. Moreover, the corresponding rankings $\langle 2,3,5,4,1\rangle$ and $\langle 3,4,1,5,2\rangle$ are not close in d_L sense. Hence, the model (2.2) based on Lee distance gives larger probability to orderings that are close to either EABDC or CEABD, i.e. the model describes two groups – one that favors candidates {D,E} and one that supports {A,C}. The fact that the model, which most clearly distinguishes the groups {A,C} and {D,E}, has the greatest explanatory power $\left(R^2_L = 0.3053\right)$ indicates that the Latent-class models (2.5) with two classes ($K = 2$) might fit the data better than models (2.2). From the definitions of models (2.2) and (2.5), it is easy to see that they coincide when $K = 1$. The values of R^2 for models (2.5) based on the eight considered distances are given in Table 2.3 in the case when there are one, two or there latent classes ($K = 1,2,3$).

	d_F	d_R	d_K	d_M	d_C	d_U	d_H	d_L
$K=1$	0.164	0.088	0.072	0.221	0.177	0.106	0.169	0.305
$K=2$	0.609	0.669	0.657	0.499	0.387	0.190	0.357	0.419
$K=3$	0.682	0.716	0.676	0.551	0.447	0.191	0.384	0.556

TABLE 2.3: Values of R^2 for models (2.5) for $K = 1,2,3$

It can be noticed that even for $K = 2$ there is a significant increase of R^2 for all models (2.2) based on the eight distances. As it was expected, the model (2.5) based on Lee distance has one of the least improvements from $K = 1$ to $K = 2$ since even when $K = 1$ there are two groups that can be distinguished in the model. For $K = 2$, the model that fits the data best is based on Spearman's rho (d_R) with $R^2 = 0.669$ on 3 unknown parameters and performs even better than the fitted Marginals model (2.6) with $R^2 = 0.567$ on 16 unknown parameters. For most of models (2.5) the improvement from $K = 2$ to $K = 3$ is not as drastic as from $K = 1$ to $K = 2$ and the estimated proportion coefficients $\hat{p_3}$ for the third group in (2.5) are close to zero. This shows that most likely there are only two major classes in the observed rankings and they are related to the groups {A,C} and {D,E}. The influence of candidate B over the complete rankings is studied in Nikolov and Stoimenova [58].

As shown in this subsection and Subsection 2.5.2, there are examples where the Distance-based model (2.2) induced by Lee distance fits better than the models based on other distances. One possible explanation for this can be found in the structure of the random variable induced by Lee distance and more specifically in the quantities $c_N(i, j)$ in (1.2). As described in Section 1.2, $c_N(i, j)$ is the minimum distance between i and j on a simple cyclic graph and $c_N(i, j)$ is not nondecreasing as j moves away from i, i.e. it is tent-shaped. Thus the "opposite" ordering is not close to the inverse ordering in the case of Lee distance. Therefore, using Lee distance in model (2.2) is appropriate in situations where there are multiple groups in the observed rankings and the *modal* ordering of one group is not the inverse ordering of another. Furthermore, models based on Lee distance can be used to detect if there are more than one groups or clusters in the data.

Chapter 3

Rank data clustering

Clustering of rank data aims to identify groups of rankers with a common or typical preference behavior. Marden [47, p. 33] considered unsupervised clustering for complete rankings based on "K-means" procedure and distances on permutations. Among recent work, Busse et al. [6] presented a method for clustering heterogeneous rank data based on the standard Mallows' model. In this chapter, the "K-means" procedure based on Lee distance is studied in details and several asymptotical results for large values of N are derived. An algorithm for approximating the normalizing constant in the clustering procedure is proposed by using some properties of Lee distance. In order to compare the clustering method based on Lee distance to those based on other distances on permutations, we apply the presented procedure to the APA dataset.

3.1 "K-means" clustering for rank data

Hartigan [31] presents clustering method for continuous data based on "K-means" procedure such that observations are grouped into K clusters by finding K means and assigning each observation to the group indexed by the closest mean. Marden [47, p. 33] considered a similar clustering procedure for rank data that aims to find K centers (rankings) about which the observations are clustered in K groups.

Suppose that there are n observations of complete rankings $\pi^1, \pi^2, \ldots, \pi^n \in \mathbf{S}_N$. For fixed number of groups K, Marden [47, p. 33] suggests the cluster centers $\hat{\sigma}^{(1)}, \hat{\sigma}^{(2)}, \ldots, \hat{\sigma}^{(K)} \in \mathbf{S}_N$ to be those rankings that minimize the mean distance between the observations and the closest corresponding group center, i.e.

$$\left(\hat{\sigma}^{(1)}, \hat{\sigma}^{(2)}, \ldots, \hat{\sigma}^{(K)}\right) = \underset{\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(K)}}{\operatorname{argmin}} C_K\left(\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(K)}\right), \tag{3.1}$$

where

$$C_K\left(\sigma^{(1)}, \sigma^{(2)}, \ldots, \sigma^{(K)}\right) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \le j \le K} d\left(\pi^i, \sigma^{(j)}\right), \tag{3.2}$$

for some distance $d(\cdot, \cdot)$ on $\mathbf{S}_N$. Usually, the number of clusters K and the distance $d(\cdot, \cdot)$ are chosen in advance and only the cluster centers in (3.1) are to be estimated by (3.2). The quantity $C_K\left(\hat{\sigma}^{(1)}, \hat{\sigma}^{(2)}, \ldots, \hat{\sigma}^{(K)}\right)$ indicates how tightly the data are clustered about the estimated centers. When the data size n, the number of ranked items N and the number of clusters K are small the rankings given in (3.1) can be found by exhaustive search. However, the estimated cluster centers (3.1) are highly dependent on the used distance $d(\cdot, \cdot)$ and their interpretation may vary based on the properties of $d(\cdot, \cdot)$. Therefore, the cluster analysis is very sensitive to the used distance on permutations.

From (3.2) it follows that the value of C_K for the estimated cluster centers decreases when the number of groups K increases, i.e. $C_K\left(\hat{\sigma}^{(1)},\hat{\sigma}^{(2)},\ldots,\hat{\sigma}^{(K)}\right)$ is decreasing in K. Thus, the quantity C_K is not appropriate for comparing the results obtained from several clustering analysis based on different values of K. In order to adjust C_K to account the number of clusters K, Marden [47, p. 34] considered the coefficient

$$T_K = 1 - \frac{C_K\left(\hat{\sigma}^{(1)},\hat{\sigma}^{(2)},\ldots,\hat{\sigma}^{(K)}\right)}{C_K^0}, \tag{3.3}$$

where $K = 1,2,\ldots$ and

$$C_K^0 = \min_{\sigma^{(1)},\sigma^{(2)},\ldots,\sigma^{(K)}} \frac{1}{N!}\sum_{\pi\in\mathbf{S}_N} \min_{1\le j\le K} d\left(\pi,\sigma^{(j)}\right) \tag{3.4}$$

is the value of C_K under uniform distribution over all possible $N!$ rankings. It is not difficult to see that $T_K \le 1$, $T_K = 0$ when the observations are uniformly distributed over the $N!$ rankings, i.e. there are no groups in the data, and $T_K = 1$ when the observations coincide with the cluster centers, i.e. the cluster centers are sufficient to describe the variety in the data. Thus, the coefficient T_K indicates how tightly the observations are clustered and can be considered as a measure of "tightness". Since the normalizing constant C_K^0 in (3.4) depends on the distance $d(\cdot,\cdot)$, it is not reasonable to compare the values of T_K which are based on different distances on $\mathbf{S}_N$. However, it is useful to study T_K for fixed distance and different values of K.

The value of C_K^0 does not depend on the observations and can be computed for fixed distance $d(\cdot,\cdot)$ and value K. Since there are $\binom{N!}{K}$ possible choices for cluster centers, the complete search in (3.4) becomes computationally demanding for $N \ge 10$. Therefore, it is helpful to approximate C_K^0 by using some properties of $d(\cdot,\cdot)$. For instance, when $K = 2$ and $d(\cdot,\cdot)$ is Kendall's tau, the value of C_K^0 can be calculated by an iterative procedure described in [47]. In the next section we make use of the properties of Lee distance and drive some approximations for the value of C_K^0.

3.2 Measure of "tightness" based on Lee distance

The asymptotic result in Theorem 1.2 can be used to find an approximation of the normalizing constant C_K^0 in the "K-means" clustering procedure based on Lee distance.

Corollary 3.1. *Define the constant C_2^0 by* (3.4) *for $K = 2$ and by using Lee distance $d_L(\cdot,\cdot)$. Then for large and even values of N the constant C_2^0 is approximated by*

$$\hat{C}_2^0 = \frac{N^2}{4} - \sqrt{\frac{N^4+8N^2}{24\pi(N-1)}}. \tag{3.5}$$

Proof. Since Lee distance is right-invariant, without loss of generality we can assume that $\sigma^{(1)} = e_N$ in (3.4). From (1.5) we know that when N is even

$$d_L(\pi,e_N) + d_L(\pi,e_N^*) = \frac{N^2}{2},$$

for all $\pi \in \mathbf{S}_N$. We will prove that $\sigma^{(2)} = e_N^*$ is an optimal solution of (3.4), i.e.

$$C_2^0 = \frac{1}{N!} \sum_{\pi \in \mathbf{S}_N} \min\left(d_L(\pi, e_N), d_L(\pi, e_N^*)\right) = \mathbf{E}\left(D_L^*\right),$$

where the random variable D_L^* is defined as $D_L^* = \min\left(D_L, \frac{N^2}{2} - D_L\right)$ and the random variable $D_L = d_L(\pi, e_N)$ is defined in Section 1.2.

Let us consider the set $\mathbf{S}_N^{(\sigma)} = \{\pi \in \mathbf{S}_N : d_L(\pi, e_N) \leq d_L(\pi, \sigma)\}$ for fixed $\sigma \in \mathbf{S}_N$. Then,

$$\begin{aligned} C_2^0 &= \min_{\sigma \in \mathbf{S}_N} \left[\frac{1}{N!} \sum_{\pi \in \mathbf{S}_N} \min\left(d_L(\pi, e_N), d_L(\pi, \sigma)\right) \right] \\ &= \frac{1}{N!} \min_{\sigma \in \mathbf{S}_N} \left[\sum_{\pi \in \mathbf{S}_N} d_L(\pi, \sigma) + \sum_{\pi \in \mathbf{S}_N^{(\sigma)}} \left(d_L(\pi, e_N) - d_L(\pi, \sigma)\right) \right] \\ &= \frac{1}{N!} \sum_{\pi \in \mathbf{S}_N} d_L(\pi, e_N^*) + \frac{1}{N!} \min_{\sigma \in \mathbf{S}_N} \left[\sum_{\pi \in \mathbf{S}_N^{(\sigma)}} \left(d_L(\pi, e_N) - d_L(\pi, \sigma)\right) \right]. \end{aligned} \tag{3.6}$$

From the triangular inequality for $d_L(\cdot, \cdot)$ and the definition of $\mathbf{S}_N^{(\sigma)}$ it follows that for $\pi \in \mathbf{S}_N^{(\sigma)}$

$$d_L(\pi, e_N) \leq d_L(\pi, \sigma) \leq d_L(\pi, e_N^*) + d_L(e_N^*, \sigma) = \frac{N^2}{2} - d_L(\pi, e_N) + \frac{N^2}{2} - d_L(e_N, \sigma), \quad \text{i.e.}$$

$$0 \leq d_L(\pi, e_N) \leq \frac{N^2}{2} - \frac{d_L(\sigma, e_N)}{2} \qquad \text{and}$$

$$d_L(\pi, e_N) \leq d_L(\pi, \sigma) \leq N^2 - d_L(\sigma, e_N) - d_L(\pi, e_N).$$

Therefore,

$$C_2^0 = \frac{1}{N!} \sum_{\pi \in \mathbf{S}_N} d_L(\pi, e_N^*) + \frac{1}{N!} \min_{\sigma \in \mathbf{S}_N} \left[\sum_{s=0}^{s_1} \sum_{t=t_0}^{t_1} (s - t) g(s, t, \sigma) \right], \tag{3.7}$$

where $s_1 = \frac{N^2}{2} - \frac{d_L(\sigma, e_N)}{2}$, $t_0 = \max\left(s, d_L(\sigma, e_N) - s\right)$, $t_1 = \min\left(\frac{N^2}{2}, N^2 - d_L(\sigma, e_N) - s\right)$ and $g(s, t, \sigma)$ is the number of permutations $\pi \in \mathbf{S}_N$ such that $d_L(\pi, e_N) = s$ and $d_L(\pi, \sigma) = t$.

Let $h_N : \left\{0, 1, \ldots, \frac{N^2}{2}\right\} \to [0, 1]$ be the probability mass function (pmf) of the random variable $D_L = d_L(\pi, e_N)$ and H_N be the set of all joint probability mass functions over $\left\{0, 1, \ldots, \frac{N^2}{2}\right\} \times \left\{0, 1, \ldots, \frac{N^2}{2}\right\}$ with marginal pmf $h_N(\cdot)$ for both variables. The right-invariant property of $d_L(\cdot, \cdot)$ implies that the random variables $d_L(\pi, e_N)$ and $d_L(\pi, \sigma)$ have the same distribution when $\pi \sim Uniform(\mathbf{S}_N)$. Thus, $\frac{g(\cdot, \cdot, \sigma)}{N!}$ is the joint pmf of $d_L(\pi, e_N)$ and $d_L(\pi, \sigma)$ for every permutation $\sigma \in \mathbf{S}_N$ and $\frac{g(\cdot, \cdot, \sigma)}{N!} \in H_N$, i.e. in (3.7) we are looking for optimal joint distribution $\frac{g(\cdot, \cdot, \sigma)}{N!} \in H_N$ such that

$$C_2^0 = \min_{\sigma \in \mathbf{S}_N} \left\{ \mathbf{E}\Big[\min\left(d_L(\pi, e_N), d_L(\pi, \sigma)\right) \Big] \right\}, \quad \text{for} \quad \pi \sim Uniform(\mathbf{S}_N).$$

From (3.7) it follows that

$$C_2^0 \geq \frac{1}{N!}\sum_{\pi\in\mathbf{S}_N} d_L(\pi,e_N^*) + \min_{\sigma\in\mathbf{S}_N}\left[\sum_{s=0}^{\frac{N^2}{2}}\sum_{t=s}^{\frac{N^2}{2}}(s-t)\frac{g(s,t,\sigma)}{N!}\right]$$
$$\geq \frac{1}{N!}\sum_{\pi\in\mathbf{S}_N} d_L(\pi,e_N^*) + \min_{f\in H_N}\left[\sum_{s=0}^{\frac{N^2}{2}}\sum_{t=s}^{\frac{N^2}{2}}(s-t)f(s,t)\right]. \tag{3.8}$$

Since $h_N(\cdot)$ is symmetric about $\frac{N^2}{4}$ when N is even (see the properties of D_L studied in Section 1.2), the chain of inequalities (3.8) can be continued by

$$C_2^0 \geq \frac{1}{N!}\sum_{\pi\in\mathbf{S}_N} d_L(\pi,e_N^*) - \frac{N^2}{2}h_N(0) + \min_{f\in H_N^*}\left[\sum_{s=1}^{\frac{N^2}{2}-1}\sum_{t=s}^{\frac{N^2}{2}-1}(s-t)f(s,t)\right]$$
$$\geq \frac{1}{N!}\sum_{\pi\in\mathbf{S}_N} d_L(\pi,e_N^*) + \sum_{s=0}^{1}\left(2s-\frac{N^2}{2}\right)h_N(s) + \min_{f\in H_N^{**}}\left[\sum_{s=2}^{\frac{N^2}{2}-2}\sum_{t=s}^{\frac{N^2}{2}-2}(s-t)f(s,t)\right]$$
$$\geq \ldots \geq \frac{1}{N!}\sum_{\pi\in\mathbf{S}_N} d_L(\pi,e_N^*) + \sum_{s=0}^{\frac{N^2}{4}}\left(2s-\frac{N^2}{2}\right)h_N(s), \tag{3.9}$$

where

$$H_N^* = \left\{ f\in H_N : \begin{array}{c} f\left(0,\frac{N^2}{2}\right) = h_N(0) \text{ and} \\ f(0,t) = f\left(s,\frac{N^2}{2}\right) = 0 \text{ for } s\neq 0, t\neq\frac{N^2}{2} \end{array} \right\},$$

$$H_N^{**} = \left\{ f\in H_N^* : \begin{array}{c} f\left(1,\frac{N^2}{2}-1\right) = h_N(1) \text{ and} \\ f(1,t) = f\left(s,\frac{N^2}{2}-1\right) = 0 \text{ for } s\neq 1, t\neq\frac{N^2}{2}-1 \end{array} \right\}, \text{ etc.}$$

Thus, one possible optimal joint pmf $\frac{g(s,t,\sigma)}{N!}$ in (3.7) takes non-zero values only on the diagonal $s+t=\frac{N^2}{2}$. The solution of the problem of finding optimal rankings in (3.6) may not be unique. However, by combining (3.9) with

$$\sum_{s=0}^{\frac{N^2}{4}}\left(2s-\frac{N^2}{2}\right)h_N(s) = \frac{1}{N!}\sum_{\pi\in\mathbf{S}_N^{(e_N^*)}}\left(d_L(\pi,e_N)-d_L(\pi,e_N^*)\right),$$

we conclude that

$$C_2^0 \geq \frac{1}{N!}\sum_{\pi\in\mathbf{S}_N}\min\left(d_L(\pi,e_N),d_L(\pi,e_N^*)\right) = \mathbf{E}\left(D_L^*\right),$$

i.e. $C_2^0 = \mathbf{E}\left(D_L^*\right)$ and one possible solution for σ in (3.6) is e_N^*.

The random variable $D_L^* = \min\left(D_L, \frac{N^2}{2} - D_L\right)$ is obtained by bounding D_L from above by $\frac{N^2}{4}$. Theorem 1.2 states that the distribution of D_L is asymptotically normal with mean and

variance

$$\mathbf{E}(D_L) = \frac{N^2}{4} \quad \text{and} \quad \mathbf{Var}(D_L) = \frac{N^4 + 8N^2}{48(N-1)}. \tag{3.10}$$

Therefore, the asymptotic distribution of D_L^* is truncated normal that is obtained from normal distribution with mean and variance given in (3.10) with upper bound $\frac{N^2}{4}$. The approximation $\hat{C}_2^0$ in (3.5) is derived by using $C_2^0 = \mathbf{E}(D_L^*)$ and formula (13.134) for the expectation of truncated normal random variable in Johnson et al. [37, p. 156]. □

The distribution of the random variable $D_L = d_L(\pi, e_N)$, where $\pi \sim Uniform(\mathbf{S}_N)$, is not symmetric when N is odd. However, from Theorem 1.2 it follows that this distribution is asymptotically normal for large values of N. Thus, (3.5) can be used as an approximation of C_K^0 for the case when $K = 2$ and N is odd. The values of C_2^0 and $\hat{C}_2^0$ induced by Lee distance are given in Table 3.1 for $4 \leq N \leq 10$. From the relative errors in percentage, presented in the last row, we can notice that the approximation $\hat{C}_2^0$ looks reasonable for $N \geq 7$.

N	4	5	6	7	8	9	10
C_2^0	3.000	4.667	6.883	9.758	13.128	16.737	20.973
$\hat{C}_2^0$	2.697	4.596	6.950	9.765	13.045	16.793	21.011
% error	10.097	1.513	0.972	0.072	0.632	0.335	0.181

TABLE 3.1: Values of C_2^0 and $\hat{C}_2^0$ based on Lee distance

The proof of Corollary 3.1 suggests that the values of C_K^0, induced by Lee distance and for $2 \leq K \leq N$, could be approximated by

$$\tilde{C}_K^0 = \frac{1}{N!} \sum_{\pi \in \mathbf{S}_N} \min_{1 \leq j \leq K} d\left(\pi, \sigma^{(j)}\right),$$

where

$$\begin{aligned}
\sigma^{(1)} &= \langle 1, 2, \ldots, N-1, N \rangle, \\
\sigma^{(2)} &= \left\langle 1 + \left[\frac{N}{K}\right], 2 + \left[\frac{N}{K}\right], \ldots, N, 1, \ldots, \left[\frac{N}{K}\right] \right\rangle, \\
&\ldots \\
\sigma^{(j)} &= \left\langle 1 + \left[\frac{(j-1)N}{K}\right], 2 + \left[\frac{(j-1)N}{K}\right], \ldots, N, 1, 2, \ldots, \left[\frac{(j-1)N}{K}\right] \right\rangle, \\
&\ldots \\
\sigma^{(K)} &= \left\langle 1 + \left[\frac{(K-1)N}{K}\right], 2 + \left[\frac{(K-1)N}{K}\right], \ldots, N, 1, 2, \ldots, \left[\frac{(K-1)N}{K}\right] \right\rangle,
\end{aligned}$$

and $[x]$ is the integer part of x. The values of C_K^0 and this approximation are given in Table 3.2 for $4 \leq N \leq 6$ and $2 \leq K \leq 4$. In this case however, we can see that the relative error, given in the last column, increases when the values N and K increase. Thus, $\tilde{C}_K^0$ doesn't approximate very well the values of C_K^0 for large number of items N and large number of groups K.

N	K	C_K^0	$\tilde{C}_K^0$	**% error**
4	2	3.000	3.000	0.000
4	3	2.500	2.500	0.000
4	4	2.000	2.000	0.000
5	2	4.667	4.667	0.000
5	3	4.133	4.250	2.823
5	4	3.817	3.958	3.710
6	2	6.883	6.883	0.000
6	3	6.400	6.500	1.562
6	4	5.944	6.267	5.422

TABLE 3.2: Values of C_K^0 and $\tilde{C}_K^0$ based on Lee distance

3.3 Illustrative example

Let us consider again the American Psychological Association (APA) election data used in Subsection 2.5.3. The values of the coefficient T_K, defined in (3.3) and based on the eight distances from Section 1.2, are given in Table 3.3 for $1 \leq K \leq 4$.

Distance name	**Notation**	$K=1$	$K=2$	$K=3$	$K=4$
Spearman's footrule	d_F	0.0693	0.1469	0.1149	0.1004
Spearman's rho	d_R	0.0810	0.1958	0.1776	0.1324
Chebyshev metric	d_M	0.0809	0.1318	0.1260	0.0945
Kendall's tau	d_K	0.0601	0.1573	0.1405	0.1165
Cayley distance	d_C	0.0777	0.1158	0.1119	0.1178
Ulam distance	d_U	0.0490	0.0524	0.0570	0.0812
Hamming distance	d_H	0.0556	0.0958	0.0920	0.0893
Lee distance	d_L	0.0827	0.0938	0.1000	0.1113

TABLE 3.3: Values of T_K for the APA data

It can be noticed that there is a significant increase of the values of T_K from the clustering procedures with $K = 1$ group to the clustering with $K = 2$ groups. The reason for this is that there are definite camps within APA: candidates A and C are research psychologists, D and E are clinical psychologists, and B is a community psychologist. The "K-means" clustering based on most of the distances does not improve T_K when K increases to 3 and 4. This shows that most likely there are only two major classes in the observed rankings. The same dichotomy is revealed by the probabilistic models (2.5) and in Diaconis [18] by using models based on the distributions of pairs of candidates.

The measure of "tightness" T_K for the clustering procedure based on Lee distance does not change significantly for different values of K. That can be explained by the fact that when $K = 1$ the clustering based on Lee distance performs relatively better than the procedures based on other distances. This phenomenon is caused by the structure of $d_L(\cdot,\cdot)$ and can be useful in situations where it is desirable to construct models with less number of groups K and respectively with fewer unknown parameters. An example of such models are the Latent-class Distance-based models presented in Section 2.1.

Chapter 4

Imperfect ranking in ranked set sampling

Ranked set sampling (RSS) is a scheme which was first proposed by McIntyre [51] and can be useful in settings where small sets of observations can be accurately or approximately ranked at a cost that is negligible compared to the cost of making formal measurements. By using the additional information from the units that are ranked, but not actually measured, RSS typically outperforms simple random sampling for a wide variety of testing and estimation problems. An elaborate review and various applications of RSS can be found in Chen et al. [9] and Wolfe [80].

In this chapter, we consider some statistical measures of deviation from the perfect ranking in the framework of ranked set sampling. We use nonparametric approach for testing the null hypothesis for perfect ranking. The Distance-based Mallows' models (2.2) with appropriate distance on permutations are suggested in the case of imperfect ranking. Some asymptotic results for the corresponding error probability matrix are derived for the models based on Spearman's footrule and Spearman's rho. We propose an EM algorithm for estimating the unknown parameter in the Mallows' models in order to compare the power of the presented test statistics. Since in the literature for RSS the sample size is commonly denoted by k, we will consider rankings of k items, which are elements of $\mathbf{S_k}$.

4.1 Ranked set sampling scheme

There are two types of RSS: balanced and unbalanced. To obtain a balanced ranked set sample, first it is necessary to draw a random sample (set) of size k and order the observations from smallest to largest. The ranking can be done by judgment and without actual measurement. Then, only the observation with the smallest rank is selected for measurement. Next, another random sample (set) of size k is drawn and ordered, but only the second smallest observation is measured and the rest are not measured. The procedure is continued until the largest observation of the k-th random sample of size k is measured. This process yields a sample of k independent values and is referred as a *one-cycle ranked set sample*. Let us denote the measured quantities by $X_{RSS} = \left\{X_{[1]}, X_{[2]}, \ldots, X_{[k]}\right\}$. The steps of obtaining X_{RSS} are illustrated on the following scheme, where the measured observation in each set (row) is

underlined and is chosen in accordance with the judgment order.

$$\begin{array}{cccccc} \underline{X_{1:k}} & X_{2:k} & \dots & X_{k:k} & \to & X_{[1]} \\ X_{1:k} & \underline{X_{2:k}} & \dots & X_{k:k} & \to & X_{[2]} \\ \dots & \dots & \dots & \dots & \to & \dots \\ X_{1:k} & X_{2:k} & \dots & \underline{X_{k:k}} & \to & X_{[k]} \end{array}$$

To obtain a *n-cycle ranked set sample*, this procedure is repeated n times and the observed data is denoted by $X_{RSS} = \left\{X_{1[1]}, X_{1[2]}, \dots, X_{1[k]}, X_{2[1]}, \dots, X_{n[1]}, \dots, X_{n[k]}\right\}$. To draw an *unbalanced ranked set sample*, we remove the constraint that the number of measured observations with in-set rank i must be the same for $i = 1, 2, \dots, k$. Instead, we determine a set size k and a vector $n = (n_1, n_2, \dots, n_m)$ such that n_i is the number of measured observations with in-set rank i, for $i = 1, 2, \dots, k$. Then the total measured sample size is $N = \sum_{i=1}^{k} n_i$.

In this study, we focus on analyzing n-cycle balanced RSS and assume that the random variable of interest X has a continuous distribution. This assures that there are no ties of the measured observations in each cycle. Since we consider models based on orderings of the measurements, the requirement that X has a continuous distribution is a necessary assumption that plays a key role in our analysis. For applications in which ties are inevitable, Ozturk [66] and Frey [23] proposed two modifications of RSS that allow the judge to declare ties in the ranking process.

Consider a n-cycle balanced RSS based on random samples (sets) of size k, $X_{RSS} = \left\{X_{1[1]}, X_{1[2]}, \dots, X_{1[k]}, X_{2[1]}, \dots, X_{n[1]}, \dots, X_{n[k]}\right\}$. In the case when the judgment ranking within each set is correct and the ranks of the measured observations coincide with their true ranks, we say that *perfect ranking* is obtained. However, since the ordering within a set is not based on actual measurement, but on some judgment criteria, it may contain errors and the judgment ranking could be inaccurate. Thus, it is possible that the judgment rank of a measured observation differs from its actual rank within the set. In this case we have *imperfect ranking*. For example, imperfect ranking could be obtained if $X_{1[1]}$ is not the smallest observation in the first random set of the first cycle.

4.2 Hypotheses testing problem

Let $\left\{X_{[1]}, X_{[2]}, \dots, X_{[k]}\right\}$ be a one-cycle balanced ranked set sample of size k from a continuous population. By arranging the $X_{[i]}$'s in an increasing order, we obtain the *ordered ranked set sample* (ORSS) $X_{1:k}^{ORSS} \le X_{2:k}^{ORSS} \le \dots X_{k:k}^{ORSS}$ introduced by Balakrishnan and Li [3]. Suppose that the observed r-th ORSS comes from the i_r-th ordered set, i.e. $X_{r:k}^{ORSS} = X_{[i_r]}$ for $r = 1, 2, \dots, k$. Let denote by $p(i_1, i_2, \dots, i_k)$ the probability that $X_{[i_1]} \le X_{[i_2]} \le \dots \le X_{[i_k]}$ under the assumption of perfect ranking. Li and Balakrishnan [43] proved that under the hypothesis of perfect ranking

$$\begin{aligned} p(i_1, i_2, \dots, i_k) &= \mathbf{P}\left(X_{[i_1]} \le X_{[i_2]} \le \dots \le X_{[i_k]} \mid \textit{Perfect ranking}\right) \\ &= \prod_{l=1}^{k-1} \sum_{j_l = j_{l-1}+i_l}^{lk} \frac{\binom{lk}{j_l}\binom{k}{i_{l+1}}}{\binom{(l+1)k}{j_l + i_{l+1}}} \left(\frac{i_{l+1}}{j_l + i_{l+1}}\right), \end{aligned} \tag{4.1}$$

where $j_0 \equiv 0$. Let assume that there is a judgement error in ranking and denote by p_{ij} the probability of the event that the j-th order statistic is judged as having rank i for $i, j = 1,2,\ldots,k$. Notice that $\sum_{j=1}^{k} p_{ij} = \sum_{i=1}^{k} p_{ij} = 1$. More details about the probabilities p_{ij} can be found in Aragon et al. [2]. Li and Balakrishnan [43] showed that

$$\mathbf{P}\left(X_{[i_1]} \leq X_{[i_2]} \leq \ldots \leq X_{[i_k]} \mid \textit{Judgement error}\right) = \sum_{l_1=1}^{k}\sum_{l_2=1}^{k}\cdots\sum_{l_k=1}^{k}\left\{\left(\prod_{r=1}^{k} p_{i_r l_r}\right) p\left(l_1,l_2,\ldots,l_k\right)\right\}, \tag{4.2}$$

where $j_0 \equiv 0$ and the probabilities $p\left(l_1,l_2,\ldots,l_k\right)$ are given in (4.1). In order to test the hypothesis of perfect ranking, Li and Balakrishnan [43] suggested to use test statistics based on distance measures between the observed rank vector of ORSS $\langle i_1,i_2,\ldots,i_k\rangle$ and the identity $e_k = \langle 1,2,\ldots,k\rangle$, which is associated with perfect ranking. Since we assume that there are no ties, both vectors are elements of the permutation group $\mathbf{S_k}$, generated by the first k natural integers. The test statistics proposed by Li and Balakrishnan [43] are

$$N_k = \sum_{r=1}^{k}\sum_{s=1}^{r-1} I\left(i_r < i_s\right), \quad S_k = \sum_{r=1}^{k}\left(i_r - r\right)^2 \quad \text{and} \quad A_k = \sum_{r=1}^{k} \mid i_r - r \mid, \tag{4.3}$$

where $I\left(\cdot\right)$ is an indicator function. The hypothesis of perfect ranking is rejected when these nonparametric test statistics are too large. Their exact null distributions can be derived from (4.1) by computing the probabilities for all possible $k!$ rankings $\langle i_1,i_2,\ldots,i_k\rangle$. Similarly, formula (4.2) can be used to calculate their power for given alternative model for imperfect ranking.

From the definitions of distances on rankings in Section 1.1, it is clear that the test statistics N_k, S_k and A_k measure the distance between $\langle i_1,i_2,\ldots,i_k\rangle$ and $\langle 1,2,\ldots,k\rangle$ by Kendall's tau, Spearman's rho and Spearman's footrule, respectively. In a similar way we can define more test statistics by using other distances on $\mathbf{S_k}$. In the next sections, we will consider the test statistics

$$M_k = \max_{1\leq r\leq k} \mid i_r - r \mid \quad \text{and} \quad L_k = \sum_{r=1}^{k} \min\left(\mid i_r - r \mid, k - \mid i_r - r \mid \right), \tag{4.4}$$

based on Chebyshev's distance and Lee's distance, respectively. Power comparisons show that the test statistics constructed by using other distances listed in Section 1.1 are less powerful than N_k, S_k, A_k, M_k and L_k. Thus, in Section 4.6 are presented results only for the test statistics defined in (4.3) and (4.4).

Notice that the distributions of test statistics based on distances on $\mathbf{S_k}$ depend not only on the probabilities $p(i_1,i_2,\ldots,i_k)$ in (4.1) under the null hypothesis, but also on the properties of the used distance. For some distances in Section 1.1, there are exact or approximate results for the distributions of the corresponding statistics under uniformly distributed permutations $\langle i_1,i_2,\ldots,i_k\rangle$. For example, A_k is approximately normally distributed with expectation $\frac{k^2-1}{3}$ and variance $\frac{(k+1)(2k^2+7)}{45}$, when $\langle i_1,i_2,\ldots,i_k\rangle$ is uniformly chosen from $\mathbf{S_k}$. However, since the probabilities $p(i_1,i_2,\ldots,i_k)$ in (4.1) are not in closed form and cannot be expressed as a function of the distance between $\langle i_1,i_2,\ldots,i_k\rangle$ and $\langle 1,2,\ldots,k\rangle$, the distributions under the null hypothesis for these statistics cannot be given explicitly. The same problem occurs for the distributions under the alternative probabilities, given in (4.2).

The test statistics for the one-cycle RSS given in (4.3) and (4.4) can be extended to the case of multi-cycle balanced RSS. Let $\mathbf{X}_{RSS} = \left\{X_{i[j]}, i = 1,2,\ldots,n; j = 1,2,\ldots,k\right\}$ be a n-cycle balanced RSS of size k, where $X_{i[j]}$ is the j-th observation in the i-th cycle. Li and Balakrishnan [43] considered the test statistics

$$\begin{aligned} N_{k,n} &= \sum_{i=1}^{n} N_k^{(i)}, & N_{k,n}^* &= \max\left(N_k^{(1)}, N_k^{(2)}, \ldots, N_k^{(n)}\right), \\ S_{k,n} &= \sum_{i=1}^{n} S_k^{(i)}, & S_{k,n}^* &= \max\left(S_k^{(1)}, S_k^{(2)}, \ldots, S_k^{(n)}\right), \\ A_{k,n} &= \sum_{i=1}^{n} A_k^{(i)}, & A_{k,n}^* &= \max\left(A_k^{(1)}, A_k^{(2)}, \ldots, A_k^{(n)}\right), \end{aligned}$$

where $N_k^{(i)}$, $S_k^{(i)}$ and $A_k^{(i)}$ are the values the test statistics in (4.3) for the i-th cycle of RSS, for $i = 1,2,\ldots,n$. Simulation results in Li and Balakrishnan [43] indicate that the tests based on $N_{k,n}$, $S_{k,n}$ and $A_{k,n}$ are more powerful than the tests based on $N_{k,n}^*$, $S_{k,n}^*$ and $A_{k,n}^*$, respectively. Pesarin and Salmaso [67, p.128] presented several functions for combining nonparametric tests and also suggested the direct method (sum of the one-cycle test statistics) in the case when all partial test statistics are homogeneous. Therefore, in this study only $N_{k,n}$, $S_{k,n}$ and $A_{k,n}$ are considered. Similarly,

$$M_{k,n} = \sum_{i=1}^{n} M_k^{(i)} \quad \text{and} \quad L_{k,n} = \sum_{i=1}^{n} L_k^{(i)}$$

can be defined as an extension of (4.4) for testing the hypothesis of perfect ranking for a n-cycle balanced RSS. Since all cycles are independent of each other, the null distributions of the test statistics $N_{k,n}$, $S_{k,n}$, $A_{k,n}$, $M_{k,n}$ and $L_{k,n}$ can be obtained from the null distributions of the statistics defined in (4.3) and (4.4).

4.3 Mallows' models for imperfect ranking

In order to compare the test statistics described in the previous section, it is necessary to fix an alternative model for the imperfect judgment ranking. Since the probability in (4.2) depends on p_{ij}, we need to specify the probability with which the event that the j-th order statistic is judged as having rank i or to specify the probability of ranking the observations within each random sample (set) with a permutation $\pi = \langle \pi(1), \pi(2), \ldots, \pi(k) \rangle \in \mathbf{S_k}$, where $\pi(j)$ is the rank of the j-th order statistic, for $j = 1,2,\ldots,k$. In this case the identity $e_k = \langle 1,2,\ldots,k \rangle$ corresponds to the perfect ranking.

Distance-based models, defined in Chapter 2, are appropriate probability models that assign larger probabilities for rankings that are *close* to a *modal* ranking. Since the judgment ranking is expected to be close to the perfect ranking, in model (2.2) we can assume that π_0 is the identity $e_k \in \mathbf{S}_k$ and $\theta \leq 0$. Then, we can define a model for imperfect ranking by

$$\mathbf{P}(\pi \mid \theta) = \exp\left(\theta d(\pi, e_k) - \psi_k(\theta)\right) \quad \text{for } \pi \in \mathbf{S_k}, \tag{4.5}$$

where $\theta \leq 0$ is a real parameter, $d(\cdot,\cdot)$ is a distance on $\mathbf{S_k}$ and $\psi_k(\theta)$ is a normalizing constant. For a chosen distance $d(\cdot,\cdot)$, the value of $\psi_k(\theta)$ can be found by using (2.3).

To distinguish models (4.5) from models (2.2) we will refer to (4.5) as Mallows' models for imperfect ranking.

By using the probabilities $\mathbf{P}(\pi \mid \theta)$ in (4.5) for all permutations $\pi \in \mathbf{S}_k$, we can derive the probability p_{ij} of the event that the j-th order statistic is judged as having rank i, for $i, j = 1,2,\ldots,k$. Since p_{ij} depends on k and θ, let us denote the ranking error probability matrix by $\mathbf{Q}(k,\theta) = \{q(i,j,k,\theta)\}_{i,j=1}^{k}$. Notice that in Chapter 2 the matrix $\mathbf{Q}(k,\theta)$ is referred as the Marginals matrix. Similar to formula (2.7), the elements of $\mathbf{Q}(k,\theta)$ can be expressed as

$$q(i,j,k,\theta) = \sum_{\pi \in \mathbf{S}_k,\, \pi(j)=i} \mathbf{P}(\pi \mid \theta), \quad \text{for } i,j = 1,2,\ldots,k, \tag{4.6}$$

where the summation is over all permutations $\pi = \langle \pi(1), \pi(2), \ldots, \pi(k) \rangle$ such that $\pi(j) = i$. From (4.5), (2.3) and (4.6) it is easy to see that $\{q(i,j,k,\theta)\}_{i,j=1}^{k}$ are continuous with respect to θ. For $\theta = 0$ the matrix $\mathbf{Q}(k,\theta)$ has equal elements, i.e. $q(i,j,k,0) = 1/k$ for $i,j = 1,2,\ldots,k$, and is associated with the uniform model. When $\theta \to -\infty$ the matrix $\mathbf{Q}(k,\theta)$ converges to the identity matrix $\mathbf{I}_{k\times k}$, which corresponds to the perfect ranking. An example of $\mathbf{Q}(k,\theta)$ based on Kendall's tau for $k = 5$ and $\theta = -0.5$ is given below.

$$\mathbf{Q}\left(5, -\frac{1}{2}\right) = \begin{pmatrix} 0.4287 & 0.2600 & 0.1577 & 0.0956 & 0.0580 \\ 0.2600 & 0.2810 & 0.2158 & 0.1476 & 0.0956 \\ 0.1577 & 0.2158 & 0.2530 & 0.2158 & 0.1577 \\ 0.0956 & 0.1476 & 0.2158 & 0.2810 & 0.2600 \\ 0.0580 & 0.0956 & 0.1577 & 0.2600 & 0.4287 \end{pmatrix}.$$

From (4.2) it follows that the alternative model for imperfect judgment ranking is completely specified by the matrix $\mathbf{Q}(k,\theta)$. The most common models for imperfect ranking, used in the literature, are: Bivariate normal model proposed by Dell and Clutter [14], Fraction of random rankings by Frey et al. [24], Fraction of inverse rankings by Frey et al. [24] and Fraction of neighbor rankings by Vock and Balakrishnan [79]. All of these four models depend on a parameter, corresponding to the magnitude of the judgment ranking error. However, it is not clear how to specify the parameters in these models. For the imperfect ranking model based on Mallows' models the unknown parameter θ can be estimated by maximizing the probability of observing $\mathbf{X}_{RSS} = \left\{X_{i[j]}, i = 1,2,\ldots,n; j = 1,2,\ldots,k\right\}$ under the hypothesis of judgment error. Moreover, the Mallows' models (4.5) give the probability of the event that the judge orders the observations within each random sample (set) to a given permutation $\pi \in \mathbf{S_k}$, i.e. the probability $\mathbf{P}\left(\pi \mid \hat{\theta}\right)$, where $\hat{\theta}$ is the maximum likelihood estimation of θ. The probability for the perfect ranking that corresponds to $\pi = e_k$ is then presented by

$$\mathbf{P}\left(e_k \mid \hat{\theta}\right) = \exp\left(-\psi_k\left(\hat{\theta}\right)\right).$$

Therefore, similarly to the interpretation of the correlation coefficient ρ in the bivariate normal model, the parameter θ can be considered as a measure of the accuracy of the judgment ranking. In addition, the error probability matrix for the bivariate normal model with $\rho = 0.5$ in Li and Balakrishnan [43] has a similar structure and is "close" to the matrix $\mathbf{Q}\left(5, -\frac{1}{2}\right)$ in the example above.

4.4 Maximum likelihood estimation of the parameter θ

Suppose that $\mathbf{X}_{RSS} = \left\{X_{i[j]}, i = 1,2,\ldots,n; j = 1,2,\ldots,k\right\}$ is an n-cycle balanced RSS and $R_{i[j]}$ is the number of the set in the i-th cycle from which comes the j-th ordered statistic, i.e. $X^{ORSS}_{i[j:k]} = X_{i\left[R_{i[j]}\right]}$ for $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,k$. Let consider Mallows' model for imperfect ranking associated with ranking error probability matrix $\mathbf{Q}(k,\theta)$ and free parameter $\theta \leq 0$. For each cycle $i = 1,2,\ldots,n$, the probability of observing $R_i = \left\langle R_{i[1]}, R_{i[2]}, \ldots, R_{i[k]} \right\rangle$ under judgment error can be calculated from formula (4.2). By using the independence of the measurements, the likelihood function can be expressed as

$$L(R \mid \theta) = \prod_{i=1}^{n} \left\{ \sum_{l} \left[\left(\prod_{j=1}^{k} q\left(R_{i[j]}, l_{[j]}, k, \theta\right) \right) p\left(l_{[1]}, l_{[2]}, \ldots, l_{[k]}\right) \right] \right\}, \tag{4.7}$$

where $\sum_{l}$ denotes a summation over all possible vectors $l = \left(l_{[1]}, l_{[2]}, \ldots, l_{[k]}\right)$ such that $l_{[j]} \in \{1,2,\ldots,k\}$ for $j = 1,2,\ldots,k$. Here $R = (R_1, R_2, \ldots, R_n)$ is the vector of the observed ORSS, $q\left(R_{i[j]}, l_{[j]}, k, \theta\right)$ are elements of $\mathbf{Q}(k,\theta)$ and the probabilities $p\left(l_{[1]}, l_{[2]}, \ldots, l_{[k]}\right)$ are calculated from (4.1).

In order to find the maximum likelihood estimate (MLE) it is required to maximize (4.7) with respect to θ. In general, there is no closed expression for the elements of the matrix $\mathbf{Q}(k,\theta)$ and it is not possible to estimate θ directly. However, the Expectation-Maximization (EM) algorithm proposed by Dempster et al. [15] can be applied. Similar to the EM algorithms in Section 2.4, we need to maximize the complete likelihood function on the value of an unknown latent random variable as well as on the unknown parameter. The latent variable of the model for imperfect ranking is the vector $Z = \left\{Z_{i[j]}, i = 1,2,\ldots,n; j = 1,2,\ldots,k\right\}$, where the element $Z_{i[j]}$ is the true rank of $X_{i[j]}$ in the j-th random sample (set) of the i-th cycle. Then from (4.7) it follows that the joint likelihood function has the form

$$L(R,Z \mid \theta) = \prod_{i=1}^{n} \left[\left(\prod_{j=1}^{k} q\left(R_{i[j]}, Z_{i[j]}, k, \theta\right) \right) p\left(Z_{i[1]}, Z_{i[2]}, \ldots, Z_{i[k]}\right) \right]$$

and the loglikelihood is

$$\log\left(L(R,Z \mid \theta)\right) = \sum_{i=1}^{n} \log\left(p\left(Z_{i[1]}, Z_{i[2]}, \ldots, Z_{i[k]}\right)\right) + \sum_{i=1}^{n}\sum_{j=1}^{k} \log\left(q\left(R_{i[j]}, Z_{i[j]}, k, \theta\right)\right).$$

E-step. The first step of the algorithm is to find the expected value of the loglikelihood function for a given initial approximation of θ. This expectation is usually denoted by

$Q\left(\theta \mid \theta^{(t)}\right)$ for a given initial value $\theta^{(t)}$. Then, for the imperfect ranking model

$$
\begin{aligned}
Q\left(\theta \mid \theta^{(t)}\right) &= \mathbf{E}_{Z|R,\theta^{(t)}}\left[\log\left(L(R,Z \mid \theta)\right)\right] \\
&= \sum_{z} \frac{\prod_{i=1}^{n}\left[\left(\prod_{j=1}^{k} q\left(R_{i[j]}, z_{i[j]}, k, \theta^{(t)}\right)\right) p\left(z_{i[1]}, z_{i[2]}, \ldots, z_{i[k]}\right)\right]}{\sum_{l}\left\{\prod_{i=1}^{n}\left[\left(\prod_{j=1}^{k} q\left(R_{i[j]}, l_{i[j]}, k, \theta^{(t)}\right)\right) p\left(l_{i[1]}, l_{i[2]}, \ldots, l_{i[k]}\right)\right]\right\}} \times \\
&\quad \times \left[\sum_{i=1}^{n} \log\left(p\left(z_{i[1]}, z_{i[2]}, \ldots, z_{i[k]}\right)\right) + \sum_{i=1}^{n}\sum_{j=1}^{k} \log\left(q\left(R_{i[j]}, z_{i[j]}, k, \theta\right)\right)\right], \qquad (4.8)
\end{aligned}
$$

where the summation $\sum_{z}$ is over all vectors $z = \left(z_{1[1]}, z_{1[2]}, \ldots, z_{n[k]}\right)$ and $\sum_{l}$ is over all vectors $l = \left(l_{1[1]}, l_{1[2]}, \ldots, l_{n[k]}\right)$ such that $z_{i[j]} \in \{1,2,\ldots,k\}$ and $l_{i[j]} \in \{1,2,\ldots,k\}$ for $i = 1,2,\ldots,k$ and $j = 1,2,\ldots,k$.

M-step. The next step is to maximize $Q\left(\theta \mid \theta^{(t)}\right)$ with respect to θ, i.e.

$$\theta^{(t+1)} = \underset{\theta \le 0}{\operatorname{argmax}}\, Q\left(\theta \mid \theta^{(t)}\right).$$

From (4.8) it follows that $\theta^{(t+1)}$ is the solution of the equation

$$
\begin{aligned}
&\sum_{z} \frac{\prod_{i=1}^{n}\left[\left(\prod_{j=1}^{k} q\left(R_{i[j]}, z_{i[j]}, k, \theta^{(t)}\right)\right) p\left(z_{i[1]}, z_{i[2]}, \ldots, z_{i[k]}\right)\right]}{\sum_{l}\left\{\prod_{i=1}^{n}\left[\left(\prod_{j=1}^{k} q\left(R_{i[j]}, l_{i[j]}, k, \theta^{(t)}\right)\right) p\left(l_{i[1]}, l_{i[2]}, \ldots, l_{i[k]}\right)\right]\right\}} \times \\
&\qquad \times \sum_{i=1}^{n}\sum_{j=1}^{k} \frac{q'\left(R_{i[j]}, z_{i[j]}, k, \theta\right)}{q\left(R_{i[j]}, z_{i[j]}, k, \theta\right)} = 0, \qquad (4.9)
\end{aligned}
$$

where $q'(r,z,k,\theta)$ is the derivative of $q(r,z,k,\theta)$ with respect to θ and can be expressed from (4.6) as

$$q'(r,z,k,\theta) = \sum_{\pi \in \mathbf{S}_k,\, \pi(z)=r} \left[d(\pi,e_k) - \psi_k'(\theta)\right] \exp\left(\theta d(\pi,e_k) - \psi_k(\theta)\right),$$

for $r,z = 1,2,\ldots,k$. The obtained maximum likelihood estimator $\theta^{(t+1)}$ is then substituted in the E-step for calculating the new values of $Q(\theta \mid \theta^{(t+1)})$ and so on. This procedure continues until some optimal criteria is reached, for example if the change of the likelihood function is relatively small.

From (4.8) and the fact that $\{q(i,j,k,\theta)\}_{i,j=1}^{k}$ are continuous in θ it follows that $Q\left(\theta \mid \theta^{(t)}\right)$ is continuous in both θ and $\theta^{(t)}$. Thus, by using *Theorem 3.2* in McLachlan and Krishnan [52, p. 82] we have that the proposed EM algorithm converges monotonically to some stationary point of $L(R \mid \theta)$. If $L(R \mid \theta)$ is unimodal, i.e. $L(R \mid \theta)$ has only one stationary point, then the EM algorithm converges to the unique MLE of θ, see the *Corollary* of *Theorem 3.5* in McLachlan and Krishnan [52, p. 84]. However, the rate of convergence and the resulting

stationary point depend on the initial value $\theta^{(0)}$. Therefore, in order to find the MLE of θ one has to take $\theta^{(0)}$ to be close to the global maximum of $L(R \mid \theta)$. From our practical experience we noticed that it is very rare for $L(R \mid \theta)$ to has more than one stationary point. Nevertheless, we recommend to apply the EM algorithm for several initial values $\theta^{(0)}$ from -10 to 0 and take the result that maximizes $L(R \mid \theta)$.

4.5 Error probability matrix based on different distances

Even though, the elements of the error probability matrix $\mathbf{Q}(k,\theta)$ can be calculated from (4.6), this may take a lot of computational time and resources when k is too large, for example $k \geq 10$. In this section, some properties of $\mathbf{Q}(k,\theta)$ based on the eight distances listed in Section 1.1 are discussed.

Let us consider the matrix $\mathbf{Q}(k,\theta)$ for Cayley and Hammning distances. From the eight distances in Section 1.1, they are the only ones that possess the bi-invariant property. Marden [47] showed that for models based on a bi-invariant distance there exists a constant a such that

$$q(i,j,k,\theta) = \begin{cases} a, & \text{for } i = j \\ \dfrac{1-a}{k-1}, & \text{for } i \neq j, \end{cases} \tag{4.10}$$

i.e. $\mathbf{Q}(k,\theta)$ has equal diagonal values and equal off-diagonal values. Since $d_C(\pi,\sigma)$ is the minimum number of transpositions needed to obtain π from σ, it follows from (4.6) that

$$q(i,j,k,\theta) = \begin{cases} \exp\big(\psi_{k-1}(\theta) - \psi_k(\theta)\big), & \text{for } i = j \\ \exp\big(\theta + \psi_{k-1}(\theta) - \psi_k(\theta)\big), & \text{for } i \neq j. \end{cases}$$

Combining the last expression with (4.10) gives

$$q(i,j,k,\theta) = \begin{cases} \dfrac{1}{1+(k-1)\exp(\theta)}, & \text{for } i = j \\ \dfrac{\exp(\theta)}{1+(k-1)\exp(\theta)}, & \text{for } i \neq j. \end{cases}$$

For Hamming distance, formula (4.6) can be represented by the recursive equation

$$q(i,j,k,\theta) = \begin{cases} \exp\big(\psi_{k-1}(\theta) - \psi_k(\theta)\big), & \text{for } i = j \\ \dfrac{\exp(\theta) + (k-2)q(i,j,k-2,\theta)}{\exp\big(\psi_k(\theta) - \psi_{k-2}(\theta) - \theta\big)}, & \text{for } i \neq j. \end{cases} \tag{4.11}$$

It follows from (4.10) and (4.11) that

$$q(i,j,k,\theta) = \begin{cases} \exp\big(\psi_{k-1}(\theta) - \psi_k(\theta)\big), & \text{for } i = j \\ \dfrac{1-\exp\big(\psi_{k-1}(\theta) - \psi_k(\theta)\big)}{k-1}, & \text{for } i \neq j. \end{cases}$$

Alternative proofs of these results for $\mathbf{Q}(k,\theta)$ based on Cayley and Hamming distances and an efficient algorithm for computing $\mathbf{Q}(k,\theta)$ based on Kendall's tau are given in Marden [47, p.165].

In Section 2.2 we have already considered the probability error matrix based on Lee distance. The asymptotic approximation in Theorem 2.1 is very helpful, since the exact calculation of $\mathbf{Q}(k,\theta)$ from (4.6) by summing over all possible $k!$ rankings becomes computationally demanding for large values of k. Similar asymptotic results for the ranking error probability matrix based on Spearman's footrule and Spearman's rho are given in Theorem 4.1 and Theorem 4.2, which are proved in Appendix E.

Theorem 4.1. *Let $\mathbf{Q}(k,\theta)$ be the ranking error probability matrix based on the Spearman's footrule. Then*

$$q(i,j,k,\theta)\frac{k}{\exp\left(\theta\mu+\frac{\theta^2\nu^2}{2}\right)} \xrightarrow[k\to\infty]{} 1,$$

where

$$\mu=\frac{k+1}{3}-\frac{f(i)+f(j)-|\,i-j\,|}{k-1}+|\,i-j\,|,$$

$$\nu^2=\frac{1}{k-2}\left\{\sum_{\substack{r=1\\ r\neq i}}^{k}\sum_{\substack{s=1\\ s\neq j}}^{k}\left[|\,r-s\,|+\frac{k(k+1)}{3(k-1)}-\frac{f(r)+f(s)-|\,i-s\,|-|\,r-j\,|}{k-1}\right.\right.$$
$$\left.\left.-\frac{f(i)+f(j)-|\,i-j\,|}{(k-1)^2}\right]^2\right\}-\frac{(k+1)(2k^2+7)}{45}$$

and

$$f(x)=\frac{x(x-1)+(k-x)(k-x+1)}{2}.$$

Theorem 4.2. *Let $\mathbf{Q}(k,\theta)$ be the ranking error probability matrix based on the Spearman's rho. Then*

$$q(i,j,k,\theta)\frac{k}{\exp\left(\theta\mu+\frac{\theta^2\nu^2}{2}\right)} \xrightarrow[k\to\infty]{} 1,$$

where

$$\mu=\frac{k(k+1)}{6}-\frac{h(i)+h(j)-(i-j)^2}{k-1}+(i-j)^2,$$

$$\nu^2=\frac{1}{k-2}\left\{\sum_{\substack{r=1\\ r\neq i}}^{k}\sum_{\substack{s=1\\ s\neq j}}^{k}\left[(r-s)^2+\frac{k^2(k+1)}{6(k-1)}-\frac{h(r)+h(s)-(i-s)^2-(r-j)^2}{k-1}\right.\right.$$
$$\left.\left.-\frac{h(i)+h(j)-(i-j)^2}{(k-1)^2}\right]^2\right\}-\frac{k^2(k-1)(k+1)^2}{36}$$

and

$$h(x)=\frac{x(x-1)(2x-1)+(k-x)(k-x+1)(2k-2x+1)}{6}.$$

As we showed in Section 2.1, the asymptotic result of Theorem 1.2 can be applied even for relatively small values of k. The approximations in Theorem 2.1, Theorem 4.1 and Theorem 4.2 have similar accuracy and look reasonably close to the exact values of $q(i,j,k,\theta)$ for $k\geq 8$. Thus, these results can be used for computing (4.9) in the EM algorithm when $k\geq 8$. As an illustration, consider the matrix $\mathbf{Q}(k,\theta)$ based on Spearman's footrule for $k=8$ and $\theta=-1/3$ and its asymptotic approximation $\hat{\mathbf{Q}}(k,\theta)$ presented in Theorem 4.1:

$$\mathbf{Q}\left(8,-\frac{1}{3}\right)=\begin{pmatrix}0.415 & 0.213 & 0.132 & 0.087 & 0.059 & 0.042 & 0.030 & 0.023\\ 0.213 & 0.298 & 0.175 & 0.114 & 0.077 & 0.054 & 0.039 & 0.030\\ 0.132 & 0.175 & 0.256 & 0.160 & 0.107 & 0.075 & 0.054 & 0.042\\ 0.087 & 0.114 & 0.160 & 0.241 & 0.156 & 0.107 & 0.077 & 0.059\\ 0.059 & 0.077 & 0.107 & 0.156 & 0.241 & 0.160 & 0.114 & 0.087\\ 0.042 & 0.054 & 0.075 & 0.107 & 0.160 & 0.256 & 0.175 & 0.132\\ 0.030 & 0.039 & 0.054 & 0.077 & 0.114 & 0.175 & 0.298 & 0.213\\ 0.023 & 0.030 & 0.042 & 0.059 & 0.087 & 0.132 & 0.213 & 0.415\end{pmatrix},$$

$$\hat{\mathbf{Q}}\left(8,-\frac{1}{3}\right)=\begin{pmatrix}0.410 & 0.210 & 0.131 & 0.089 & 0.062 & 0.043 & 0.031 & 0.025\\ 0.210 & 0.273 & 0.169 & 0.114 & 0.079 & 0.054 & 0.039 & 0.031\\ 0.131 & 0.169 & 0.247 & 0.164 & 0.112 & 0.077 & 0.054 & 0.043\\ 0.089 & 0.114 & 0.164 & 0.247 & 0.165 & 0.112 & 0.079 & 0.062\\ 0.062 & 0.079 & 0.112 & 0.165 & 0.247 & 0.164 & 0.114 & 0.089\\ 0.043 & 0.054 & 0.077 & 0.112 & 0.164 & 0.247 & 0.169 & 0.131\\ 0.031 & 0.039 & 0.054 & 0.079 & 0.114 & 0.169 & 0.273 & 0.210\\ 0.025 & 0.031 & 0.043 & 0.062 & 0.089 & 0.131 & 0.210 & 0.410\end{pmatrix}.$$

Since the differences between the corresponding elements of the two matrices are relatively small, the matrix $\hat{\mathbf{Q}}\left(8,-\frac{1}{3}\right)$ can be used as an approximation of $\mathbf{Q}\left(8,-\frac{1}{3}\right)$.

4.6 Power comparisons

Let us consider again the problem of testing the hypothesis of perfect ranking versus the general alternative of imperfect ranking. For a given nominal level, the critical values for the nonparametric test statistics for one-cycle RSS, defined in (4.3) and (4.4), does not depend on the model for imperfect ranking and can be calculated by using (4.1). For example, when $k=5$ and the nominal level is 0.05, the critical values of N_k, S_k, A_k, M_k and L_k are 4, 12, 8, 3 and 7, respectively. Since the test statistics have discrete distributions, it is clear that the nominal level 0.05 cannot be achieved exactly. Thus, we can use a standard randomization in order to fix the significance to be exactly 0.05. For example, when $k=5$ and significance level is 0.05, the two possible critical values of N_k are 4 with exact level of 0.03345 and 3 with exact level of 0.12687. When the observed value of N_k is 4, the hypothesis of perfect ranking is rejected. If the observed value of N_k is 3, then the hypothesis of perfect ranking is rejected with probability $\frac{0.05-0.03345}{0.12687-0.03345}\approx 0.17716$.

The powers of the presented tests depend on the alternative model for imperfect ranking and can be expressed as a sum of the probabilities (4.2) for all permutations $\langle i_1,i_2,\ldots,i_k\rangle$ in the critical region. Power comparisons between N_k, S_k and A_k under bivariate normal models are presented in Li and Balakrishnan [43]. Similar results for the powers of N_k, S_k, A_k, M_k and L_k are given in Table 4.1. The powers are determined from 100 000 Monte Carlo simulations based on an bivariate normal alternative with correlation coefficient ρ varying from 0.50 to 1.00. The significance is fixed to be exactly 0.05 by using a standard randomization. Hence,

the powers under perfect ranking ($\rho = 1$) coincide with the nominal level and are presented in the last column of Table 4.1.

k	Test	ρ					
		0.50	0.60	0.70	0.80	0.90	1.00
4	N_4	0.4024	0.3523	0.2934	0.2310	0.1615	0.0500
	S_4	0.4035	0.3534	0.2943	0.2317	0.1620	0.0500
	A_4	0.3728	0.3259	0.2709	0.2139	0.1508	0.0500
	M_4	0.3315	0.2913	0.2444	0.1955	0.1410	0.0500
	L_4	0.2471	0.2234	0.1945	0.1621	0.1237	0.0500
5	N_5	0.5310	0.4642	0.3879	0.3033	0.2029	0.0500
	S_5	0.5529	0.4859	0.4079	0.3208	0.2152	0.0500
	A_5	0.5100	0.4458	0.3732	0.2930	0.1978	0.0500
	M_5	0.5029	0.4385	0.3656	0.2861	0.1927	0.0500
	L_5	0.3854	0.3419	0.2924	0.2370	0.1680	0.0500
6	N_6	0.6526	0.5787	0.4900	0.3848	0.2530	0.0500
	S_6	0.6674	0.5939	0.5049	0.3984	0.2626	0.0500
	A_6	0.6622	0.5899	0.5021	0.3966	0.2624	0.0500
	M_6	0.5724	0.5035	0.4246	0.3347	0.2243	0.0500
	L_6	0.5781	0.5183	0.4452	0.3565	0.2413	0.0500

TABLE 4.1: Simulated powers under bivariate normal model with ρ from 0.50 to 1.00 and nominal level 0.05

From the results in Table 4.1, it can be concluded that S_k is more powerful than the other test statistics, whereas M_k and L_k are less powerful when $k \leq 6$. However, it is worth comparing the powers under the Mallows' models based on the eight distances listed in Section 1.1 for all test statistics defined in (4.3) and (4.4). The alternative model in this case is specified by the parameter θ and the MLE $\hat{\theta}$ depends on the observed ORSS permutation $\langle i_1, i_2, \ldots, i_k \rangle$. Therefore, the power of the tests N_k, S_k, A_k, M_k and L_k depends on $\langle i_1, i_2, \ldots, i_k \rangle$ and can be computed by using the EM algorithm in Section 4.4. The randomized powers of the considered test statistics when $k = 5$ and the nominal level is 0.05 are given in Table 4.2 for some key permutations $\langle i_1, i_2, \ldots, i_5 \rangle \in \mathbf{S_5}$. The last section of Table 4.2 prints the powers under the perfect ranking $\langle 1, 2, 3, 4, 5 \rangle$.

Similar to the powers under bivariate normal model, S_5 is most powerful under all Mallows' models for the presented ORSS permutations in Table 4.2. However, under Mallows' alternative, the power of the test statistic L_5 is much closer to the power of N_5 and A_5, and it is not clear which one of those three is more powerful. The results in Table 4.2 show that M_5 is much less powerful in this case.

4.7 Illustrative example

To illustrate the use of Mallows' alternative for imperfect ranking in n-cycle RSS, the models in Section 4.5 are applied here to an example. Murrary et al. [56] compared the effect of four different sprayer settings on the amount of spray deposit on the leaves of apple trees. In order to estimate the mean amount of spray deposit, which is measured by the percentage of the upper leaf surface covered with deposit, Murrary et al. [56] collected a RSS by first

ORSS	Test	Distance							
		d_F	d_R	d_M	d_K	d_C	d_U	d_H	d_L
$\langle 5,4,3,2,1\rangle$	N_5	0.780	0.780	0.780	0.780	0.780	0.780	0.780	0.780
	S_5	0.798	0.798	0.798	0.798	0.798	0.798	0.798	0.798
	A_5	0.756	0.756	0.756	0.756	0.756	0.756	0.756	0.756
	M_5	0.419	0.419	0.419	0.419	0.419	0.419	0.419	0.419
	L_5	0.755	0.755	0.755	0.755	0.755	0.755	0.755	0.755
$\langle 4,1,2,3,5\rangle$	N_5	0.451	0.494	0.494	0.468	0.448	0.659	0.448	0.571
	S_5	0.473	0.512	0.513	0.488	0.479	0.677	0.479	0.596
	A_5	0.433	0.476	0.474	0.450	0.425	0.636	0.425	0.548
	M_5	0.241	0.273	0.269	0.256	0.215	0.345	0.215	0.282
	L_5	0.432	0.462	0.465	0.441	0.451	0.633	0.451	0.564
$\langle 3,2,1,5,4\rangle$	N_5	0.478	0.387	0.427	0.454	0.643	0.533	0.643	0.629
	S_5	0.500	0.401	0.445	0.473	0.669	0.551	0.669	0.652
	A_5	0.459	0.375	0.411	0.437	0.616	0.514	0.616	0.605
	M_5	0.254	0.228	0.240	0.250	0.320	0.281	0.320	0.314
	L_5	0.458	0.357	0.401	0.427	0.629	0.514	0.629	0.616
$\langle 2,3,1,4,5\rangle$	N_5	0.183	0.176	0.179	0.182	0.212	0.217	0.212	0.253
	S_5	0.194	0.179	0.186	0.191	0.235	0.219	0.235	0.275
	A_5	0.178	0.176	0.177	0.179	0.202	0.205	0.202	0.245
	M_5	0.121	0.136	0.130	0.127	0.114	0.128	0.114	0.138
	L_5	0.177	0.159	0.168	0.172	0.224	0.213	0.224	0.265
$\langle 1,2,3,4,5\rangle$	N_5	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	S_5	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	A_5	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	M_5	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
	L_5	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050

TABLE 4.2: Powers under the Mallows' models for $k = 5$ and nominal level 0.05

spraying the leaves with a fluorescent tracer dye and then visually ranking them under ultra-violet light. Perfect ranking in this experiment would be obtained if the ordering based on the visual ranking of the leaves coincide with the true ordering based on the amount of spray deposit on the leaves. In this section, we use and compare the RSS for the first low-volume sprayer settings and for the second high-volume sprayer settings which are given as an example in Ozturk [64]. The data consists of measurements of the percentage of leaf surface covered with deposit. The observations for each settings are obtained from five-cycle RSS with each cycle being of size 5, i.e. $k = 5$ and $n = 5$. The measured percentages of cover $\left\{X_{[1]}, X_{[2]}, X_{[3]}, X_{[4]}, X_{[5]}\right\}$, the corresponding ORSS and the observed values of the test statistics (4.3) and (4.4) for the low-volume settings are presented in Table 4.3 for each of the five cycles.

For testing the hypothesis of perfect ranking, we use the statistics $N_{k,n}$, $S_{k,n}$, $A_{k,n}$, $M_{k,n}$ and $L_{k,n}$, defined in Section 4.2. Their values from the observed data in Table 4.3 are as follows:

$$N_{5,5} = 8, \quad S_{5,5} = 22, \quad A_{5,5} = 16, \quad M_{5,5} = 7, \quad L_{5,5} = 16.$$

The critical values of $N_{5,5}$, $S_{5,5}$, $A_{5,5}$, $M_{5,5}$ and $L_{5,5}$ at 0.05 significance level are 12, 34, 22, 9

Cycle	$X_{[1]}$	$X_{[2]}$	$X_{[3]}$	$X_{[4]}$	$X_{[5]}$	ORSS	N_5	S_5	A_5	M_5	L_5
1	0.3	2.8	24.4	5.7	14.3	$\langle 1,2,4,5,3\rangle$	2	6	4	2	4
2	3.9	11.9	12.6	10.5	56.5	$\langle 1,4,2,3,5\rangle$	2	6	4	2	4
3	3.4	11.8	13.0	21.8	29.6	$\langle 1,2,3,4,5\rangle$	0	0	0	0	0
4	5.1	10.4	19.3	21.0	15.0	$\langle 1,2,5,3,4\rangle$	2	6	4	2	4
5	3.2	14.1	13.0	25.0	22.9	$\langle 1,3,2,5,4\rangle$	2	4	4	1	4

TABLE 4.3: RSS of the percentage of leaf surface covered with deposit for the low-volume settings

and 20, respectively. Hence, there is not enough evidence to conclude that the ranking in the RSS data in Table 4.3 is not perfect. The powers of the considered tests under the Mallows' alternative for imperfect ranking, calculated by the MLE of the unknown parameter θ, are presented in Table 4.4. The interpretation of the powers based on the low-volume settings example are similar to the ones obtained from the results in Table 4.2. As in the one-cycle example, $S_{5,5}$ is more powerful than the other test statistics, whereas $M_{5,5}$ and $L_{5,5}$ are less powerful. The test statistics $N_{5,5}$ and $A_{5,5}$ have similar powers under all eight Mallows' models.

Test statistic	Distance							
	d_F	d_R	d_M	d_K	d_C	d_U	d_H	d_L
$N_{5,5}$	0.818	0.760	0.809	0.784	0.963	0.926	0.963	0.950
$S_{5,5}$	0.830	0.760	0.814	0.791	0.970	0.934	0.970	0.959
$A_{5,5}$	0.807	0.758	0.801	0.776	0.956	0.917	0.955	0.940
$M_{5,5}$	0.775	0.698	0.755	0.732	0.949	0.900	0.948	0.933
$L_{5,5}$	0.748	0.721	0.755	0.728	0.904	0.850	0.902	0.866

TABLE 4.4: Estimated powers under the Mallows' alternative for imperfect ranking

The main advantage of using Mallows' models as an alternative for imperfect ranking is that it is possible to estimate the unknown parameter θ by applying the EM algorithm in Section 4.4. The estimated value of θ can be used not only for comparing the powers of the presented nonparametric statistics, but is helpful for measuring the judgment ranking ability. For the example in Table 4.3 and Mallows' alternative model based on Spearman's footrule d_F, the estimated parameter is $\hat{\theta}_F^{low} = -0.494$, which is not close to 0 and indicates that the observations in each set of the RSS are not randomly ranked. The corresponding error probability matrix is

$$\mathbf{Q}\left(5,\hat{\theta}_F^{low}\right) = \begin{pmatrix} 0.577 & 0.215 & 0.110 & 0.061 & 0.037 \\ 0.215 & 0.431 & 0.190 & 0.103 & 0.061 \\ 0.110 & 0.190 & 0.400 & 0.190 & 0.110 \\ 0.061 & 0.103 & 0.190 & 0.431 & 0.215 \\ 0.037 & 0.061 & 0.110 & 0.215 & 0.577 \end{pmatrix} \tag{4.12}$$

and can be used for further investigation of the effect of imperfect ranking on the performance of some statistical procedures based on RSS, see Aragon et al. [2] and Section 3.1.2 in Chen et al. [9]. Furthermore, the ranking abilities of two judges (or ranking methods) can be

compared just by considering their MLEs of θ in the Mallows' model. For example, we can check if the ranking procedure based on the fluorescent tracer method orders the leaves in the same way for the low-volume and the high-volume settings.

Cycle	$X_{[1]}$	$X_{[2]}$	$X_{[3]}$	$X_{[4]}$	$X_{[5]}$	ORSS	N_5	S_5	A_5	M_5	L_5
1	4.2	8.9	19.9	26.9	39.5	$\langle 1,2,3,4,5\rangle$	0	0	0	0	0
2	4.4	8.3	22.7	17.7	74.4	$\langle 1,2,4,3,5\rangle$	1	2	2	1	2
3	4.4	17.1	6.7	19.2	33.6	$\langle 1,3,2,4,5\rangle$	1	2	2	1	2
4	0.9	1.7	21.7	43.8	54.4	$\langle 1,2,3,4,5\rangle$	0	0	0	0	0
5	7.1	13.2	31.0	34.3	37.9	$\langle 1,2,3,4,5\rangle$	0	0	0	0	0

TABLE 4.5: RSS of the percentage of leaf surface covered with deposit for the high-volume settings

The measured observations of percentages of cover for the high-volume settings are presented in Table 4.5 for each of the five cycles. From the corresponding ORSS and the observed values of N_5, S_5, A_5, M_5 and L_5, given in Table 4.5, it seems that the ranking of the leaves for the high-volume settings is closer to the perfect ranking compared to the ranking in the low-volume case. Similar results are obtained by comparing the MLEs of the unknown parameter in the Mallows' models for imperfect ranking. For the example in Table 4.5 and model based on Spearman's footrule d_F, the estimated parameter is $\hat{\theta}_F^{high} = -0.696$, which is smaller than the corresponding estimation $\hat{\theta}_F^{low} = -0.494$ for the RSS in the low-volume settings. Since the perfect ranking is associated with $\theta \to -\infty$, we conclude that the ranking of the leaves based on the fluorescent tracer method is better in the high-volume case. In order to study in more details the differences between the ranking abilities of the method in the two settings, we can use the error probability matrix, which for the example in Table 4.5 is

$$\mathbf{Q}\left(5, \hat{\theta}_F^{high}\right) = \begin{pmatrix} 0.714 & 0.178 & 0.068 & 0.028 & 0.012 \\ 0.178 & 0.565 & 0.164 & 0.065 & 0.028 \\ 0.068 & 0.164 & 0.536 & 0.164 & 0.068 \\ 0.028 & 0.065 & 0.164 & 0.565 & 0.178 \\ 0.012 & 0.028 & 0.068 & 0.178 & 0.714 \end{pmatrix}, \tag{4.13}$$

when the used distance is Spearman's footrule d_F. From (4.12) and (4.13) we see that $\mathbf{Q}\left(5, \hat{\theta}_F^{high}\right)$ is closer to the identity matrix compared to $\mathbf{Q}\left(5, \hat{\theta}_F^{low}\right)$. For example, the probability that rank 1 is correctly assigned to the smallest observation is 0.577 in $\mathbf{Q}\left(5, \hat{\theta}_F^{low}\right)$ and 0.714 in $\mathbf{Q}\left(5, \hat{\theta}_F^{high}\right)$. Therefore, the ranking based on fluorescent tracer method is better in the high-volume settings and the performance of the mean estimator based on RSS is more efficient for the data in Table 4.5.

Chapter 5

Lee distance in two-sample rank tests

Nonparametric rank tests have proved their useful in a wide range of applications, including many which are beyond the reach of conventional parametric statistics. For example, they can be applied to continuous, ordered and categorical data, and to values that are normal, almost normal, and non-normally distributed. Flexible, robust in the face of missing data and violations of assumptions, the rank tests are among the most powerful statistical procedures. They have been developed for a multitude of hypothesis testing situations, such as the two-sample and multi-sample location problems, the two-sample dispersion problem with equal medians and problems of testing for trend and for independence, see Hollander and Wolfe [33] and Gibbons and Chakraborti [27].

In this chapter, we apply Critchlow's [11] unified approach to the two-sample location problem. The test statistic induced by Lee distance is studied in details. The distribution of the test statistic under the null hypothesis is derived and an asymptotic approximation is proposed for large sample sizes. A comparison between the considered test statistic and other statistics for two samples is made via simulation study.

5.1 Critchlow's method for two-sample location problem

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two independent random samples with continuous distribution functions $F(x)$ and $G(x)$, respectively. We consider rank tests for the two-sample location problem of testing the null hypothesis H_0 against the alternative H_1

$$H_0 : F(x) \equiv G(x) \tag{5.1}$$

$$H_1 : F(x) \geq G(x), \tag{5.2}$$

with strict inequality for some x. Let $\alpha(i)$ be the rank of X_i for $i = 1, 2, \ldots, m$ and $\alpha(m+j)$ be the rank of Y_j for $j = 1, 2, \ldots, n$ among $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$. Then, the rank vector of all observations is $\alpha = \langle \alpha(1), \alpha(2), \ldots, \alpha(m+n) \rangle$ and $\alpha \in \mathbf{S}_{m+n}$, where $\mathbf{S}_{m+n}$ is the permutation group generated by the first $m+n$ natural integers. The class of permutations, which are most in agreement with the alternative H_1 is $E = \mathbf{S}_m \times \mathbf{S}_n = \{\pi \in \mathbf{S}_{m+n} : \pi(i) \leq m, \forall i \leq m\}$. The left coset $[\alpha] = \alpha(\mathbf{S}_m \times \mathbf{S}_n) = \{\alpha \circ \pi : \pi \in \mathbf{S}_m \times \mathbf{S}_n\}$ consists of all permutations in $\mathbf{S}_{m+n}$ which are equivalent to α. Many rank statistics could be obtained by using distances between sets of permutation. Critchlow [11] proposed a unified approach to constructing nonparametric tests which produces many well-known rank statistics. The method is based on finding the minimum interpoint distance between the class of equivalence $[\alpha]$ and

the extremal set E

$$d([\alpha],E) = \min_{\substack{\pi\in[\alpha]\\ \sigma\in E}} d(\pi,\sigma) \tag{5.3}$$

where d is an arbitrary distance on $\mathbf{S}_{m+n}$. The proposed test rejects the null hypothesis H_0 for small values of the statistic $d([\alpha],E)$. This contrasts with the structure of some parametric test, where H_0 is rejected if the distance from H_0 is large. Since the minimal value of the proposed test statistic is zero and $d([\alpha],E) = 0$ if and only if $d(\alpha,\sigma) = 0$ for some $\sigma \in E$, the strongest evidence for rejecting H_0 occurs if and only if the observed permutation α is equivalent to some extremal permutation $\sigma \in E$.

Critchlow[11] obtained the minimal value defined by (5.3) for four of the eight listed distances in Section 1.1 and proved that the induced test statistics are equivalent to some familiar rank test statistics: $d_F([\alpha],E) \leftrightarrow$ Wilcoxon test statistic; $d_U([\alpha],E) \leftrightarrow$ Kolmogorov-Smirnov test statistic; $d_K([\alpha],E) \leftrightarrow$ Mann-Whitney test statistic; $d_H([\alpha],E) \leftrightarrow$ Mood "median test" statistic for equal sample sizes $(m = n)$. For the Chebyshev distance Stoimenova [73] derived $d_M([\alpha],E) = \max\{a_m - m, m+1-a_{m+1}\}$, where a_m is the maximal rank in $\{\alpha(1),\alpha(2),\ldots,\alpha(m)\}$ and a_{m+1} is the minimal rank in $\{\alpha(m+1),\alpha(m+2),\ldots,\alpha(m+n)\}$.

5.2 Rank test statistic based on Lee distance

The goal of this section is to derive and study the rank test statistic in (5.3) induced by the Lee distance. Since $d_L(\cdot,\cdot)$ is right-invariant, it follows that

$$\begin{aligned} d_L([\alpha],E) &= \min_{\substack{\pi\in[\alpha]\\ \sigma\in E}} d_L(\pi,\sigma) = \min_{\pi\in[\alpha]} d_L(\pi,e) \\ &= \min_{\pi\in[\alpha]} \left\{ \sum_{i=1}^{m+n} \min\left(|\, a(i)-i\,|, m+n-|\, a(i)-i\,|\right) \right\}, \end{aligned} \tag{5.4}$$

where $e = \langle 1,2,\ldots,m+n\rangle$ is the identity permutation. After solving the optimal problem (5.4), $d_L([\alpha],E)$ can be expressed as

$$\begin{aligned} &d_L([\alpha],E) = \\ &\sum_{i\in K_m} \min\left(|\, \alpha(i) - \gamma_n^{-1}(k+1-\gamma_m(\alpha(i)))\,|, m+n-|\, \alpha(i)-\gamma_n^{-1}(k+1-\gamma_m(\alpha(i)))\,|\right) \\ &+ \sum_{i\in K_n} \min\left(|\, \alpha(i) - \gamma_m^{-1}(k+1-\gamma_n(\alpha(i)))\,|,\right. \\ &\qquad\qquad \left. m+n-|\, \alpha(i)-\gamma_m^{-1}(k+1-\gamma_n(\alpha(i)))\,|\right) \\ &= 2\sum_{i\in K_m} \min\left(|\, \alpha(i) - \gamma_n^{-1}(k+1-\gamma_m(\alpha(i)))\,|,\right. \\ &\qquad\qquad \left. m+n-|\, \alpha(i)-\gamma_n^{-1}(k+1-\gamma_m(\alpha(i)))\,|\right) \end{aligned} \tag{5.5}$$

where

$$\begin{aligned} K_m &= \{i \in \{1,2,\ldots,m\} : \alpha(i) > m\}\ , \\ K_n &= \{i \in \{m+1,m+2,\ldots,m+n\} : \alpha(i) \le m\}\ , \end{aligned} \tag{5.6}$$

k is the number of elements of K_m $(k = \mid K_m \mid = \mid K_n \mid)$, $\gamma_m(\alpha(i))$ is the rank of $\alpha(i)$ among $\{\alpha(i) : i \in K_m\}$, $\gamma_n(\alpha(i))$ is the rank of $\alpha(i)$ among $\{\alpha(i) : i \in K_n\}$ and γ^{-1} is the inverse of γ, i.e. $\gamma^{-1}(\gamma(\alpha(i))) = \alpha(i)$. Since $d_L([\alpha], E)$ is equivalent to the rank statistic

$$L_{m,n} := \frac{d_L([\alpha], E)}{2}, \tag{5.7}$$

$L_{m,n}$ can be used for testing H_0 against the alternative H_1.

5.3 Properties of $L_{m,n}$

There is an interpretation of the test statistic $L_{m,n}$ in terms of graph theory. Let C be a simple cycle graph with vertices $\{i\}_{i=1}^{m+n}$ and edges $\bigcup_{i=1}^{m+n-1}\{i, i+1\}$ and $\{m+n, 1\}$. Then $L_{m,n}$ is the minimum sum of distances over C between the elements of K_m and the elements of K_n. An example when $m = 6$, $n = 4$, $K_m = \{3,5\}$ and $K_n = \{8,9\}$ is illustrated on Figure 5.1. In this case $L_{m,n} = (10 - \mid 3-9 \mid) + \mid 5-8 \mid = 4+3 = 7$.

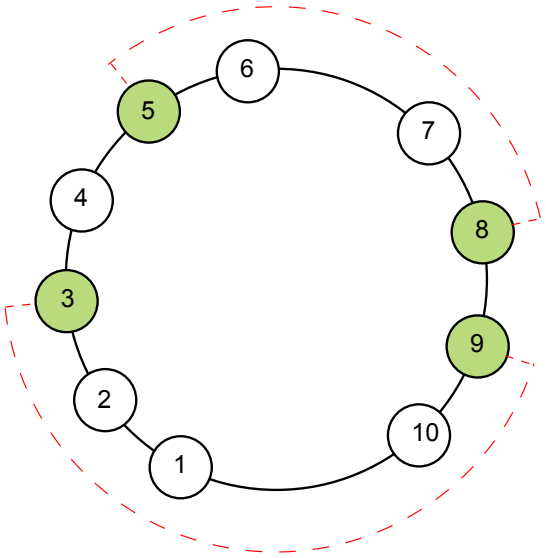


FIGURE 5.1: Cyclic graph and Lee distance

The value of $L_{m,n}$ depends not only on the elements in K_m and K_n, but also on the way in which their elements are paired. Formula (5.5) gives that the minimal sum of distances between pairwise elements of K_m and K_n is obtained when the smallest element of K_m is combined with the largest element of K_n, the second smallest element of K_m is combined with the second largest element of K_n, ..., the largest element of K_m is combined with the smallest element of K_n. Using this fact the distribution of the test statistic could be calculated for fixed number k of elements in K_m and K_n, $k = \mid K_m \mid = \mid K_n \mid$. Let $[K_m \times K_n]^*$ be the described above set of pairs and $s-1$ be the number of pairs $(x,y) \in [K_m \times K_n]^*$ for which the shortest path on C goes over the edge $\{m, m+1\}$. Obviously, s is between 1 and $k+1$. If for some pair $(x,y) \in [K_m \times K_n]^*$ the paths over $\{m, m+1\}$ and over $\{m+n, 1\}$ are with the same length, then the path over $\{m+n, 1\}$ is considered to be the shortest. For $i = 0, 1, \ldots, s-1$ let $a_i^{(m)}$ be the number of elements in $\{1, 2, \ldots, m\} \backslash K_m$ which are in the shortest path of exactly i pairs $(x,y) \in [K_m \times K_n]^*$ connected by the edge $\{m, m+1\}$. For $j = 1, 2, \ldots, k-s+1$ let $b_j^{(m)}$ be the number of elements in $\{1, 2, \ldots, m\} \backslash K_m$ which are in the shortest path of exactly j pairs $(x,y) \in [K_m \times K_n]^*$ connected by the edge $\{m+n, 1\}$. Similarly the numbers $\{a_i^{(n)}\}_{i=0}^{s-1}$ and

$\{b_j^{(n)}\}_{j=1}^{k-s+1}$ are defined for the set $\{m+1, m+2, \ldots, m+n\} \backslash K_n$. An illustration of the used notation is shown on Figure 5.2. For the considered example on Figure 5.1, $m = 6$, $n = 4$, $[K_m \times K_n]^* = \{(3,9), (5,8)\}$, $s = 2$, $a_0^{(m)} = 1 =\mid \{4\} \mid$, $a_1^{(m)} = 1 =\mid \{6\} \mid$, $b_1^{(m)} = 2 =\mid \{1,2\} \mid$, $a_0^{(n)} = 0$, $a_1^{(n)} = 1 =\mid \{7\} \mid$ and $b_1^{(n)} = 1 =\mid \{10\} \mid$.

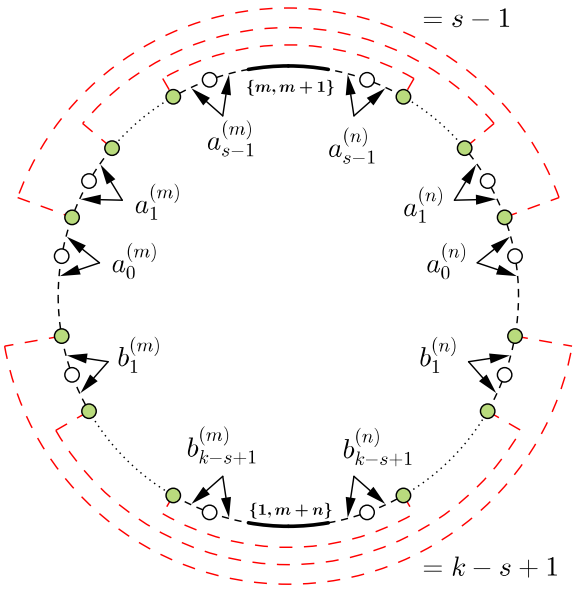


FIGURE 5.2: An illustration of the notations

The following proposition shows that $L_{m,n}$ can be determined by counting all possible values of s, $\{a_i^{(m)}\}_{i=0}^{s-1}$, $\{a_i^{(n)}\}_{i=0}^{s-1}$, $\{b_j^{(m)}\}_{j=1}^{k-s+1}$ and $\{b_j^{(n)}\}_{j=1}^{k-s+1}$.

Proposition 5.1. *Let $L_{m,n}$ be defined by* (5.7) *and $H_{m,n} =\mid K_m \mid =\mid K_n \mid$ be the number of elements of the set K_m, defined by* (5.6). *Then the joint distribution of $L_{m,n}$ and $H_{m,n}$ under the null hypothesis is given by*

$$P(L_{m,n} = l, H_{m,n} = k) = \begin{cases} \dfrac{m!n!}{(m+n)!} & , for\ \ l = 0\ \ and\ \ k = 0 \\ \displaystyle\sum_{s}\sum_{a,b} \frac{m!n!}{(m+n)!} & , for\ \ 1 \le k \le \min(m,n)\ \ and \end{cases} \tag{5.8}$$

$\left[\dfrac{k^2+1}{2}\right] \le l \le \left[\dfrac{(m+n-k)k+1}{2}\right]$, *where $[x]$ is the integer part of x. The first summation in* (5.8) *is taken over all s such that $1 \le s \le k+1$ and $(s-1)^2 + (k-s+1)^2 \le l$. The second summation is over all nonnegative integers $\{a_i^{(m)}\}_{i=0}^{s-1}$, $\{a_i^{(n)}\}_{i=0}^{s-1}$, $\{b_j^{(m)}\}_{j=1}^{k-s+1}$ and $\{b_j^{(n)}\}_{j=1}^{k-s+1}$ that satisfy:*

(i) $\displaystyle\sum_{i=0}^{s-1} a_i^{(m)} + \sum_{j=0}^{k-s+1} b_j^{(m)} = m-k$ ***(ii)*** $\displaystyle\sum_{i=0}^{s-1} a_i^{(n)} + \sum_{j=0}^{k-s+1} b_j^{(n)} = n-k$

(iii) $\displaystyle l = (s-1)^2 + (k-s+1)^2 + \sum_{i=0}^{s-1} i\left(a_i^{(m)} + a_i^{(n)}\right) + \sum_{j=0}^{k-s+1} j\left(b_j^{(m)} + b_j^{(n)}\right)$

(iv) $2(s-1)+\sum_{i=0}^{s-1}\left(a_i^{(m)}+a_i^{(n)}\right)\geq 2(k-s)+\sum_{j=0}^{k-s+1}\left(b_j^{(m)}+b_j^{(n)}\right)$, *if* $s\in\{1,2,\ldots,k\}$

(v) $2(s-2)+\sum_{i=1}^{s-1}\left(a_i^{(m)}+a_i^{(n)}\right)<2(k-s+1)+a_0^{(m)}+a_0^{(n)}+\sum_{j=0}^{k-s+1}\left(b_j^{(m)}+b_j^{(n)}\right)$,

if $s\in\{2,3,\ldots,k+1\}$. *The integers* $b_0^{(m)}$ *and* $b_0^{(n)}$ *are defined to be zeros,* $b_0^{(m)}:=0$, $b_0^{(n)}:=0$, *for completeness in conditions* ***(i)-(v)***.

Although the joint distribution of $L_{m,n}$ and $H_{m,n}$ is given in Proposition 5.1, for large values of m and n the computational process of checking conditions **(i)-(v)** for all possible non-negative integers $\{a_i^{(m)}\}_{i=0}^{s-1}$, $\{a_i^{(n)}\}_{i=0}^{s-1}$, $\{b_j^{(m)}\}_{j=1}^{k-s+1}$ and $\{b_j^{(n)}\}_{j=1}^{k-s+1}$ is time-consuming and requires a large amount of computational resources. Next proposition gives recursive relation of the number of terms in (5.8) which significantly decreases the computational complexity of formula (5.8).

Proposition 5.2. *Let* $N(m,n,k,l):=\frac{(m+n)!}{m!n!}P(L_{m,n}=l,H_{m,n}=k)$, *i.e.* $N(m,n,k,l)$ *is the number of terms of the summations in* (5.8). *Then for* $m,n\geq 2$

$$N(m,n,k,l)=\begin{cases}N(m-1,n,k,l)+N(m,n-1,k,l)\\ \quad+N\left(m-1,n-1,k-1,l-\frac{m+n-1}{2}\right)-N(m-1,n-1,k,l) & \text{, if } m+n \text{ is odd}\\ N(m-1,n,k,l)+N(m,n-1,k,l)\\ \quad+N(m-1,n-1,k,l-1)-N(m-1,n-1,k,l)\\ \quad-N(m-2,n-1,k,l-1)-N(m-1,n-2,k,l-1)\\ \quad+N(m-2,n-2,k,l-1)-N\left(m-2,n-2,k-1,l-\frac{m+n-2}{2}\right)\\ \quad+N\left(m-2,n-1,k-1,l-\frac{m+n}{2}\right)+N\left(m-1,n-2,k-1,l-\frac{m+n}{2}\right)\\ \quad+N(m-2,n-2,k-2,l-m-n+2)+N\left(m-2,n-2,k-1,l-\frac{m+n}{2}\right) & \text{, if } m+n \text{ is even.}\end{cases}\tag{5.9}$$

Proof. It is not hard to check that if $a_0^{(m)}\neq 0$ or $a_0^{(n)}\neq 0$ then the number of terms in (5.8) is $N(m-1,n,k,l)+N(m,n-1,k,l)-N(m-1,n-1,k,l)$. In the case when $m+n$ is odd, the number of terms in (5.8) for which $a_0^{(m)}=0$ and $a_0^{(n)}=0$ is $N\left(m-1,n-1,k-1,l-\frac{m+n-1}{2}\right)$. Thus, the first case of relation (5.9) is proved. Let us consider the case when $m+n$ is even and let $R(m,n,k,l)$ be the number of terms in (5.8) for which $a_0^{(m)}=0$, $a_0^{(n)}=0$ and

$$2(s-1)+\sum_{i=0}^{s-1}\left(a_i^{(m)}+a_i^{(n)}\right)=2(k-s)+\sum_{j=0}^{k-s+1}\left(b_j^{(m)}+b_j^{(n)}\right),$$

i.e. $R(m,n,k,l)$ is the number of all possible combinations for which $L_{m,n}=l$, $H_{m,n}=k$ and there is a pair $(x,y)\in[K_m\times K_n]^*$ such that the distance between x and y on the cycle graph C is exactly $\frac{m+n-2}{2}$. The other possible combinations with $a_0^{(m)}=0$ and $a_0^{(n)}=0$ are obtained when there are two pairs $(x_1,y_1)\in[K_m\times K_n]^*$ and $(x_2,y_2)\in[K_m\times K_n]^*$ such that the distance between x_i and y_i on C is exactly $\frac{m+n-4}{2}$ for $i=1,2$. Thus, by using simple

combinatorial reasoning it follows that when $m+n$ is even

$$N(m,n,k,l) = N(m-1,n,k,l) + N(m,n-1,k,l) - N(m-1,n-1,k,l) + R(m,n,k,l) + R\left(m-1,n-1,k-1,l-\frac{m+n-2}{2}\right) - R(m-1,n-1,k,l) \tag{5.10}$$

and

$$N\left(m-1,n-1,k-1,l-\frac{m+n}{2}\right) - R\left(m-1,n-1,k-1,l-\frac{m+n}{2}\right) = R(m,n,k,l) - R(m,n,k,l-1). \tag{5.11}$$

From (5.10) and (5.11) it follows that

$$R(m,n,k,l) - R(m,n,k,l+1) = N(m,n,k,l) - N(m-1,n,k,l) - N(m,n-1,k,l) + N(m-1,n-1,k,l) - N\left(m-1,n-1,k-1,l-\frac{m+n-2}{2}\right) \tag{5.12}$$

Substituting $R(m,n,k,l)$ from (5.11) to (5.10) gives

$$N(m,n,k,l) = N(m-1,n,k,l) + N(m,n-1,k,l) - N(m-1,n-1,k,l) + R\left(m-1,n-1,k-1,l-\frac{m+n-2}{2}\right) - R\left(m-1,n-1,k-1,l-\frac{m+n}{2}\right) + R(m,n,k,l-1) - R(m-1,n-1,k,l) + N\left(m-1,n-1,k-1,l-\frac{m+n}{2}\right) \tag{5.13}$$

The second case of formula (5.9) is obtained by combining (5.12) and (5.13), which completes the proof. □

From Proposition 5.2 it follows that formula (5.9) combined with the initial condition $N(m,n,0,0) = 1$ can be used to calculate the joint distribution of $L_{m,n}$ and $H_{m,n}$ for large values of m and n. The statistic $H_{m,n}$ is equivalent to Mood's statistic derived by Critchlow's method and based on Hamming distance. Since $H_{m,n}$ is the number of elements of K_m, the marginal distribution $H_{m,n}$ under the null hypothesis H_0 is hypergeometric with parameters $m+n$, m and n, i.e. $H_{m,n} \sim HG(m+n,m,n)$ and

$$P(H_{m,n} = k) = \frac{\binom{m}{k}\binom{n}{n-k}}{\binom{m+n}{n}}, \qquad \text{for } k = 0,1,\ldots,min(m,n).$$

Given the joint distribution of $L_{m,n}$ and $H_{m,n}$ the distribution of $L_{m,n}$ under the null hypothesis is presented by

$$P(L_{m,n}=l)=\begin{cases}\dfrac{m!n!}{(m+n)!} & \text{, for } l=0\\ \displaystyle\sum_k P(L=l,K=k) & \text{, for } l=1,2,\ldots,\left[\dfrac{mn+1}{2}\right],\end{cases} \tag{5.14}$$

where the sum is over all $k\in\{1,2,\ldots,min(m,n)\}$ for which

$$\left[\frac{k^2+1}{2}\right]\le l\le\left[\frac{(m+n-k)\,k+1}{2}\right].$$

The following proposition can easily be proved by combining (5.14) and Proposition 5.2.

Proposition 5.3. *The probability mass function of $L_{m,n}$ under the null hypothesis H_0 is given by*

$$P(L_{m,n}=l)=\begin{cases}\frac{m}{m+n}P(L_{m-1,n}=l)+\frac{n}{m+n}P(L_{m,n-1}=l) & \\ \quad+\frac{mn}{(m+n)(m+n-1)}\left[P\left(L_{m-1,n-1}=l-\frac{m+n-1}{2}\right)-P(L_{m-1,n-1}=l)\right] & \text{, if } m+n \text{ is odd}\\ \frac{m}{m+n}P(L_{m-1,n}=l)+\frac{n}{m+n}P(L_{m,n-1}=l) & \\ \quad+\frac{mn}{(m+n)(m+n-1)}\left[P(L_{m-1,n-1}=l-1)-P(L_{m-1,n-1}=l)\right] & \\ \quad+\frac{mn(m-1)}{(m+n)(m+n-1)(m+n-2)}\left[P\left(L_{m-2,n-1}=l-\frac{m+n}{2}\right)-P(L_{m-2,n-1}=l-1)\right] & \\ \quad+\frac{mn(n-1)}{(m+n)(m+n-1)(m+n-2)}\left[P\left(L_{m-1,n-2}=l-\frac{m+n}{2}\right)-P(L_{m-1,n-2}=l-1)\right] & \\ \quad+\frac{mn(m-1)(n-1)}{(m+n)(m+n-1)(m+n-2)(m+n-3)}P(L_{m-2,n-2}=l-1) & \text{, if } m+n \text{ is even.}\\ \quad-\frac{mn(m-1)(n-1)}{(m+n)(m+n-1)(m+n-2)(m+n-3)}P\left(L_{m-2,n-2}=l-\frac{m+n-2}{2}\right) & \\ \quad-\frac{mn(m-1)(n-1)}{(m+n)(m+n-1)(m+n-2)(m+n-3)}P\left(L_{m-2,n-2}=l-\frac{m+n}{2}\right) & \\ \quad+\frac{mn(m-1)(n-1)}{(m+n)(m+n-1)(m+n-2)(m+n-3)}P(L_{m-2,n-2}=l-m-n+2) & \end{cases} \tag{5.15}$$

for $l=0,1,\ldots,\left[\frac{mn+1}{2}\right]$ and $m,n\ge 2$.

The mean and variance of $L_{m,n}$ can be derived by using recursive relation (5.15). The results are presented in Theorem 5.1, which is proved in Appendix E.

Theorem 5.1. *Let $L_{m,n}$ be defined by* (5.7)*. Then the mean and variance of $L_{m,n}$ under the null hypothesis H_0 are*

$$\mathbf{E}(L_{m,n})=\begin{cases}\frac{mn(m+n+1)}{4(m+n)} & \text{, if } m+n \text{ is odd}\\ \frac{mn(m+n)}{4(m+n-1)} & \text{, if } m+n \text{ is even}\end{cases} \tag{5.16}$$

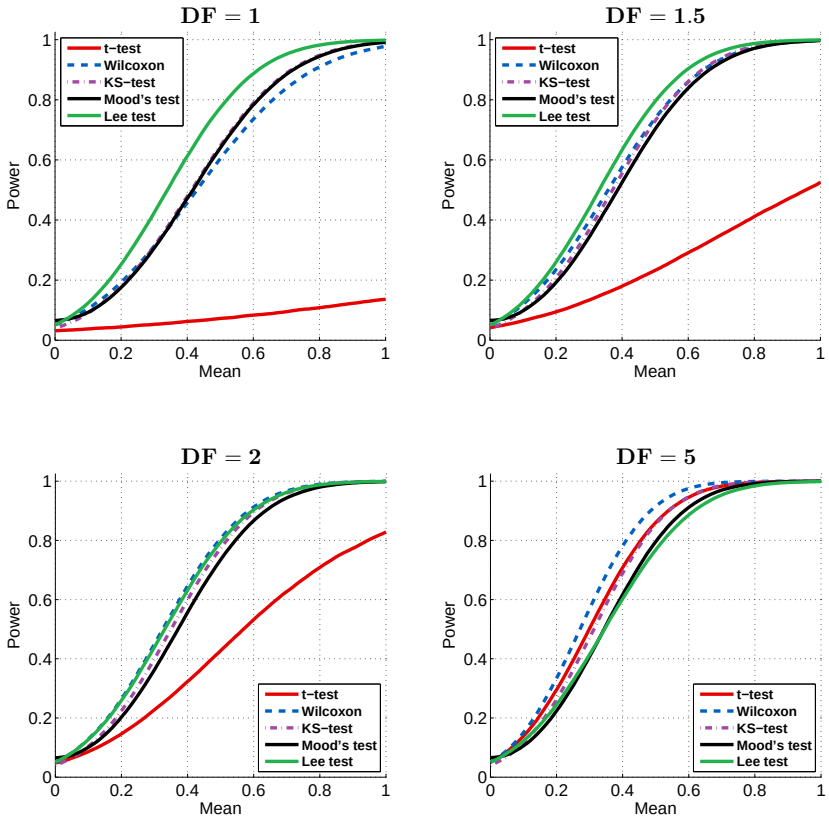
$$\mathbf{Var}(L_{m,n})=\begin{cases}\frac{mn\left\{(m+n)^4-(m+n)^3+7(m+n)^2-15(m+n)-6mn(m+n-1)\right\}}{48(m+n)^2(m+n-2)} & \text{, if } m+n \text{ is odd}\\ \frac{mn\left\{(m+n-1)^4+11(m+n-1)^2-24(m+n-1)-6mn(m+n-2)\right\}}{48(m+n-1)^2(m+n-3)} & \text{, if } m+n \text{ is even}\end{cases}, \tag{5.17}$$

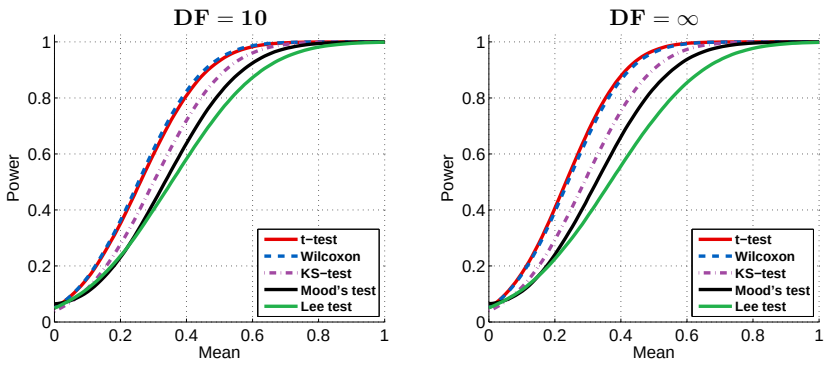
Furthermore, under the null hypothesis H_0 the standardized statistic $\frac{L_{m,n}-\mathbf{E}(L_{m,n})}{\sqrt{\mathbf{Var}(L_{m,n})}}$ has asymptotically normal distribution as $m,n\to\infty$.

5.4 Simulation study

In this section, we perform an illustrative simulation study in order to compare the power of the test statistic $L_{m,n}$ in (5.7) with other statistics for two-samples. The rank test procedure presented in Section 5.2 is compared to Wilcoxon, Kolmogorov-Smirnov and Mood's tests. More details and an elaborate description of these nonparametric rank procedures can be found in Gibbons and Chakraborti [27]. We also included the t-test in this analysis since it is the uniformly most powerful unbiased test when the populations are normally distributed. Many researchers, see e.g. Tiku et al. [75] and Marozzi [48], showed that the t-test maintains its type-I error rate close to the nominal level α, even for non-normally distributed samples with moderate sizes, except for samples from very heavy-tailed distributions like the Cauchy, where the sample mean is not useful to assess the location aspect.

In order to study these tests under diverse hypothesis in terms of the corresponding underlying distributions, we used the t-distributions since they have heavy-tails when the degrees of freedom (DF) are small and converge to the standard normal distribution when $DF \to \infty$ (see *Chapter 28* in Johnson et al. [37]). The comparison analysis is constructed from Monte Carlo simulations of two samples with sizes $m = 100$ and $n = 100$ from t-distributions with different degrees of freedom ($DF = 1, 1.5, 2, 5, 10, \infty$). For a given value of DF, we fixed the distribution of the first sample to have mean 0 and let the mean varies from 0 to 1 (with step 0.01) for the second sample. For each settings, 100 000 trails of two samples are generated and the power of a given testing procedure is estimated by the ratio of the correctly rejected null hypothesis H_0 in (5.1) to the total number of trails (=100 000). The results obtained for nominal level $\alpha = 0.05$ are shown on the figures below.





Not surprisingly, the power of all tests is close to the nominal level $\alpha = 0.05$ when the two samples have equal means and approaches 1 when the difference between the mean values increases. Notice that when the degrees of freedom (DF) are small the power of the t-test increases remarkably slower compared to the nonparametric tests and to the cases of large values of DF. This could be explained by the assumptions of the t-test which is designed to be most powerful for two normal samples ($DF = \infty$) and faces problems with heavy-tailed distributions, for example the Cauchy distribution ($DF = 1$). In this sense, the power of the t-test is not very robust for non-normal distributions, especially for very heavy-tailed ones.

Since the nonparametric test methods are prized for their lack of assumptions concerning the underlying distributions, they are appropriate for situations where outliers or "broad-tails" could be potentially observed. It is worth mentioning that the test statistics in Wilcoxon, Kolmogorov-Smirnov and Mood's tests and $L_{m,n}$ in (5.7) are based on ranks and are not computationally or time demanding. The result of Theorem 5.1 gives a normal approximation of the distribution of $L_{m,n}$ under the null hypothesis and thus it is not hard to find the critical region when m and n are large. From our empirical experience we recommend to apply the normal approximation when $m, n \geq 20$. Otherwise, formula (5.15) can be used for computing the critical region.

From the results presented on the figures above, we can notice that the test based on Lee distance is more powerful than the others when the generating distributions have heavy-tails. When the degrees of freedom increase to infinity (corresponding to standard normal distribution) we see that the test procedures are ordered with respect to their power in a reverse manner, i.e. the t-test is the most powerful followed by Wilcoxon test, Komogorov-Smirnov test (KS-test), Mood's test and Lee test. Hence, from the simulation study we can conclude that in the testing procedures for the two-sample location problem there is a trade-off between the testing power and the robustness with respect to the underlying distributions. In this regard, the nonparametric rank test based on Lee distance gives a new testing procedure which inherits the robust properties of Lee distance and is more powerful in the heavy-tailed case.

Appendix A

Main contributions

The main accomplishments in the thesis due to the author are listed below.

1. The random variable induced by Lee distance under uniformity of the rankings is studied in details and some of its characteristics such as mean, variance, range and symmetry are given for an arbitrary size of the rank vectors (see Section 1.2). An asymptotic normality for the corresponding distribution is proved and used to approximate the normalizing constant in the Distance-based probability model for rank data. This result could be also applied to other models for rankings which are based on Lee distance (see Theorem 1.2 and Section 2.1).

2. The Expectation-Maximization (EM) algorithm for computing the maximum likelihood estimates of the parameters in the Latent-class Distance-based model is generalized for the case where the *modal* rankings of the latent classes are unknown (see Section 2.4). The convergence of the proposed algorithm to a stationary point is proved and the method is applied to the to the well-studied APA election dataset (see Proposition 2.1 and Subsection 2.5.3). By using the EM algorithm we can fit the model to the data, make statistical inference and compare models based on different distances on permutations (see Subsection 2.4.3).

3. An asymptotic approximation of the normalizing constant used in the measure of "tightness" is given for the "K-means" rank clustering based on Lee distance (see Corollary 3.1). The obtained result reduces the computational time and resources for calculating the "tightness" coefficient when there are two clustering groups ($K = 2$) and the size of the rank vectors is relatively large ($N \geq 7$).

4. The Mallows' model is proposed as an alternative model for imperfect ranking in the framework of balanced ranked set sampling (RSS). An EM algorithm for estimating the unknown parameter in the model is described and its convergence to a stationary point is shown (see Sections 4.3 and 4.4). Asymptotic results for the corresponding probability error matrices based on Spearman's footrule, Spearman's rho and Lee distance are derived for the case when the size of each cycle of the RSS is too large (see Theorems 4.1, 4.2 and 2.1). The proposed alternative model can be used to study the effect of imperfect ranking on the performance of some statistical procedures based on RSS and to compare the ranking abilities of two judges or ranking methods.

5. The nonparametric rank statistic based on Critchlow's method and Lee distance is derived for the two-sample location problem. Asymptotic normality of the obtained test statistic under the null hypothesis is proved and can be used for finding the critical regions when the samples sizes are too large (see Section 5.2 and Theorem 5.1). The Lee

test statistic is shown to be more powerful for heavy-tailed underlying distributions via a simulation study (see Section 5.4).

Appendix B

Publications related to the thesis

[57] N. I. Nikolov (2016) Lee distance in two-sample rank tests. In: *Computer Data Analysis and Modeling: Theoretical and Applied Stochastics: Proceedings of the Eleventh International Conference*, Minsk: Publishing Center of BSU, pp. 100–103.

[58] N. I. Nikolov and E. Stoimenova (2017) Mallows' model based on Lee distance. In: *Proceedings of the 20-th European Young Statisticians Meetings*, pp. 59–66.

[59] N. I. Nikolov and E. Stoimenova (2019a) Asymptotic properties of Lee distance.*Metrika*, Vol. 82(3), 385–408.

[60] N. I. Nikolov and E. Stoimenova (2019b) EM estimation of the parameters in latent Mallows' models. In: *Studies in Computational Intelligence*, Springer Series, Vol. 793, pp. 317–325.

[61] N. I. Nikolov and E. Stoimenova (2019c) Mallows' models for imperfect ranking in ranked set sampling.*AStA Advances in Statistical Analysis.* `https://doi.org/10.1007/s10182-019-00354-4`, 1–26.

[62] N. I. Nikolov and E. Stoimenova (2020) Rank data clustering based on Lee distance. In: *Proceedings of 13-th Annual Meeting of the Bulgarian Section of SIAM*, pp. 1–11, (accepted).

Appendix C

Approbation of the thesis

The results from the thesis have been presented in the following talks:

1. *"Lee distance in two-sample rank tests"*, 11-th International Conference: Computer Data Analysis and Modeling, Minsk, Belarus (September 7, 2016).

2. *"Mallows' models based on Lee distance"*, 20-th European Young Statisticians Meetings, Uppsala, Sweden (August 17, 2017).

3. *"Mallows' models for imperfect rankings in ranked set sampling"*, 13-th International Conference on Ordered Statistical Data Cadiz, Spain (May 22, 2018).

4. *"Some properties of Lee distance in two-sample location problem"*, 18-th International Summer Conference on Probability and Statistics, Pomorie, Bulgaria (June 27, 2018).

5. *"Rank data models based on Lee distance"*, International Conference on Trends and Perspectives in Linear Statistical Inference, Bedlewo, Poland (August 21, 2018).

6. *"Two-sample rank test based on Lee distance"*, 15-th Applied Statistics International Conference, Ribno, Slovenia (September 24, 2018).

7. *"Distance-based models for imperfect ranking in ranked set sampling"*, XLIV Mathematical Statistics Conference, Bedlewo, Poland (December 3, 2018).

8. *"Rank data clustering based on Lee distance"*, 13-th Annual Meeting of the Bulgarian Section of SIAM, Sofia, Bulgaria (December 19, 2018).

Appendix D

Declaration of originality

The author declares that the thesis and the related publications contain original results obtained by him in cooperation with his supervisor. The usage of results of other scientists is accompanied by suitable citations.

Appendix E

Proofs

E.1 Proofs – Chapter 2

In order to prove Theorem 2.1, let us consider the random variables $D_{N,k} = d_L(\pi, e_N)$, where $k = 1,2,\ldots,N$ and π is randomly selected from $\mathbf{S}_{N,k} = \{\sigma \in \mathbf{S}_N : \sigma(N) = k\}$, i.e. $\pi \sim Uniform(\mathbf{S}_{N,k})$. Then, for fixed k,

$$D_{N,k}(\pi) = \sum_{i=1}^{N} c_N(\pi(i), i) = \sum_{i=1}^{N-1} c_N(\pi(i), i) + c_N(k,N) = \sum_{i=1}^{N-1} \tilde{c}_N(\sigma(i), i) + c_N(k,N)\,,$$

where $\sigma \in \mathbf{S}_{N-1}$ and for $i, j = 1,2,\ldots,N-1$,

$$\sigma(i) = \begin{cases} \pi(i), & \text{if } \pi(i) < k \\ \pi(i) - 1, & \text{if } \pi(i) > k, \end{cases} \qquad \tilde{c}_N(j,i) = \begin{cases} c_N(j,i), & \text{if } j < k \\ c_N(j+1,i), & \text{if } j \geq k. \end{cases} \tag{E.1}$$

Lemma E.1. *Let* $\tilde{D}_{N-1}(\sigma) = \sum_{i=1}^{N-1} \tilde{c}_N(\sigma(i), i)$, *where* $\sigma \sim Uniform(\mathbf{S}_{N-1})$ *and* $\tilde{c}_N(\cdot,\cdot)$ *is as in* (E.1). *Then the distribution of* $\tilde{D}_{N-1}$ *is asymptotically normal and the mean and variance of* $\tilde{D}_{N-1}$ *are*

$$\mathbf{E}(\tilde{D}_{N-1}) = \frac{c_N(k,N)}{N-1} + \frac{N-2}{N-1}\left\lceil \frac{N+1}{2} \right\rceil \left\lceil \frac{N}{2} \right\rceil,$$

$$\mathbf{Var}(\tilde{D}_{N-1}) = \frac{N^2 \left(c_N(k,N)\right)^2 - 2N \left\lceil \frac{N+1}{2} \right\rceil \left\lceil \frac{N}{2} \right\rceil c_N(k,N)}{(N-2)(N-1)^2} + \beta_{N-1},$$

where

$$\beta_{N-1} = \begin{cases} \dfrac{N^2\left(N^3 - 2N^2 + 10N - 12\right)}{48(N-1)^2}, & \textit{for } N \textit{ even} \\ \dfrac{(N+1)\left(N^3 - 3N^2 + 6N - 6\right)}{48(N-2)}, & \textit{for } N \textit{ odd.} \end{cases} \tag{E.2}$$

Proof. From (1.7) of Theorem 1.1 and formulas (E.1) and (1.11), it follows that

$$\begin{aligned}
\mathbf{E}\left(\tilde{D}_{N-1}\right) &\overset{(1.7)}{=} \frac{1}{N-1}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\tilde{c}_N(i,j) \overset{(E.1)}{=} \frac{1}{N-1}\sum_{\substack{i=1\\ i\neq k}}^{N}\sum_{j=1}^{N-1}c_N(i,j)\\
&= \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j=1}^{N}c_N(i,j) - \frac{1}{N-1}\sum_{i=1}^{N}c_N(i,N) - \frac{1}{N-1}\sum_{j=1}^{N}c_N(k,j) + \frac{c_N(k,N)}{N-1}\\
&\overset{(1.11)}{=} \frac{N}{N-1}\left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right] - \frac{1}{N-1}\left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right] - \frac{1}{N-1}\left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right] + \frac{c_N(k,N)}{N-1}\\
&= \frac{c_N(k,N)}{N-1} + \frac{N-2}{N-1}\left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right].
\end{aligned}$$

Using (1.8) of Theorem 1.1,

$$\mathbf{Var}\left(\tilde{D}_{N-1}\right) = \frac{1}{N-2}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\tilde{b}_N^2(i,j) = \frac{1}{N-2}\sum_{\substack{i=1\\ i\neq k}}^{N}\sum_{j=1}^{N-1}b_N^2(i,j), \text{ where}$$

$$b_N(i,j) = c_N(i,j) - \sum_{\substack{g=1\\ g\neq k}}^{N}\frac{c_N(g,j)}{N-1} - \sum_{h=1}^{N-1}\frac{c_N(i,h)}{N-1} + \frac{1}{(N-1)^2}\sum_{\substack{g=1\\ g\neq k}}^{N}\sum_{h=1}^{N-1}c_N(g,h), \tag{E.3}$$

for $i,j = 1,2,\ldots,N$. Simplifying expression (E.3) gives

$$b_N(i,j) = c_N(i,j) + \frac{c_N(i,N)+c_N(k,j)}{N-1} + \frac{c_N(k,N)}{(N-1)^2} - \frac{N}{(N-1)^2}\left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right]. \tag{E.4}$$

When N is even, the variance of $\tilde{D}_{N-1}$ can be calculated by

$$\begin{aligned}
\mathbf{Var}\left(\tilde{D}_{N-1}\right) = \frac{1}{N-2}\sum_{\substack{i=1\\ i\neq k}}^{N}\Bigg\{&\sum_{j=1}^{k-\frac{N}{2}}b_N^2(i,j) + \sum_{j=k-\frac{N}{2}+1}^{\frac{N}{2}}b_N^2(i,j) + \sum_{j=\frac{N}{2}+1}^{k}b_N^2(i,j)\\
&+ \sum_{j=k+1}^{N-1}b_N^2(i,j)\Bigg\} = \frac{1}{N-2}\left(Q_1+Q_2+Q_3+Q_4\right),
\end{aligned}$$

where the summation $\sum_{j=l_1}^{l_2} = 0$, if $l_1 > l_2$. Since the computations for Q_1, Q_2, Q_3 and Q_4 are similar, only the steps for Q_1 are presented herein.

$$\begin{aligned}
Q_1 = \sum_{\substack{i=1\\ i\neq k}}^{N}\sum_{j=1}^{k-\frac{N}{2}}b_N^2(i,j) = \sum_{j=1}^{k-\frac{N}{2}}\sum_{\substack{i=1\\ i\neq k}}^{N}b_N^2(i,j) = \sum_{j=1}^{k-\frac{N}{2}}\Bigg\{&\sum_{i=1}^{j-1}b_N^2(i,j) + \sum_{i=j}^{\frac{N}{2}}b_N^2(i,j)\\
&+ \sum_{i=\frac{N}{2}+1}^{\frac{N}{2}+j-1}b_N^2(i,j) + \sum_{i=\frac{N}{2}+j}^{N}b_N^2(i,j) - b_N^2(k,j)\Bigg\} = Q_1^{(1)}+Q_1^{(2)}+Q_1^{(3)}+Q_1^{(4)}-Q_1^{(5)},
\end{aligned}$$

where

$$Q_1^{(1)} = \sum_{j=1}^{k-\frac{N}{2}} \sum_{i=1}^{j-1} b_N^2(i,j) = \sum_{j=1}^{k-\frac{N}{2}} \sum_{i=1}^{j-1} \left(j-i+\frac{i+(N-k+j)}{N-1} + B_N(k) \right)^2,$$

$$Q_1^{(2)} = \sum_{j=1}^{k-\frac{N}{2}} \sum_{i=j}^{\frac{N}{2}} b_N^2(i,j) = \sum_{j=1}^{k-\frac{N}{2}} \sum_{i=j}^{\frac{N}{2}} \left(i-j+\frac{i+(N-k+j)}{N-1} + B_N(k) \right)^2,$$

$$Q_1^{(3)} = \sum_{j=1}^{k-\frac{N}{2}} \sum_{i=\frac{N}{2}+1}^{\frac{N}{2}+j-1} b_N^2(i,j) = \sum_{j=1}^{k-\frac{N}{2}} \sum_{i=\frac{N}{2}+1}^{\frac{N}{2}+j-1} \left(i-j+\frac{N-i+(N-k+j)}{N-1} + B_N(k) \right)^2,$$

$$Q_1^{(4)} = \sum_{j=1}^{k-\frac{N}{2}} \sum_{i=\frac{N}{2}+j}^{N} b_N^2(i,j) = \sum_{j=1}^{k-\frac{N}{2}} \sum_{i=\frac{N}{2}+j}^{N} \left(N-i+j+\frac{N-i+(N-k+j)}{N-1} + B_N(k) \right)^2,$$

$$Q_1^{(5)} = \sum_{j=1}^{k-\frac{N}{2}} b_N^2(k,j) = \sum_{j=1}^{k-\frac{N}{2}} \left(N-k+j+\frac{N-k+(N-k+j)}{N-1} + B_N(k) \right)^2,$$

for $B_N(k) = \dfrac{c_N(k,N)}{(N-1)^2} - \dfrac{N}{(N-1)^2}\left[\dfrac{N+1}{2}\right]\left[\dfrac{N}{2}\right] = \dfrac{4(N-k)-N^3}{4(N-1)^2}$ and $\sum_{i=l_1}^{l_2} = 0$, if $l_1 > l_2$. The calculation of Q_1 is completed by repeatedly using the formula

$$\sum_{i=1}^{n} (i-a)^2 = na^2 + \frac{n(n+1)(2n+1-6a)}{6} \tag{E.5}$$

for appropriate values of a and n.

The quantities Q_2, Q_3 and Q_4 can be decomposed and calculated in a similar fashion as shown for Q_1. The final result for the variance of $\tilde{D}_{N-1}$, when N is even, is

$$\mathbf{Var}\left(\tilde{D}_{N-1}\right) = \frac{2N^2\left(c_N(k,N)\right)^2 - N^3 c_N(k,N)}{2(N-2)(N-1)^2} + \frac{N^2\left(N^3-2N^2+10N-12\right)}{48(N-1)^2}.$$

The variance $\mathbf{Var}\left(\tilde{D}_{N-1}\right)$, when N is odd, can be obtained by decomposing it to four decomposable double sums and applying formula (E.5), as in the case when N is even.

From (E.4) and (1.3), it follows that

$$\max_{1\le i,j\le N} b_N^2(i,j) \le \left(\left[\frac{N}{2}\right] + \frac{\left[\frac{N}{2}\right]+\left[\frac{N}{2}\right]}{N-1} + \frac{\left[\frac{N}{2}\right]}{(N-1)^2} - \frac{N}{(N-1)^2}\left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right] \right)^2.$$

By using (E.2),

$$\frac{1}{N-1}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1} \tilde{b}_N^2(i,j) = \frac{N-2}{N-1}\mathbf{Var}\left(\tilde{D}_{N-1}\right) \ge \frac{N-2}{N-1}\beta_{N-1} = N^3\left(\frac{1}{48} + O\left(\frac{1}{N}\right)\right),$$

where $\lim_{N\to\infty} O\left(\frac{1}{N}\right) = 0$. Therefore,

$$\lim_{N\to\infty} \frac{\max_{1\le i,j\le N-1} \tilde{b}_N^2(i,j)}{\frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1} \tilde{b}_N^2(i,j)} \le \lim_{N\to\infty} \frac{N^2\left(\frac{1}{16}+O\left(\frac{1}{N}\right)\right)}{N^3\left(\frac{1}{48}+O\left(\frac{1}{N}\right)\right)} = 0,$$

i.e. the condition (1.9) of Theorem 1.1 is fulfilled and the distribution of $\tilde{D}_{N-1}$ is asymptotically normal. □

Proof of Theorem 2.1. From (2.2), (2.7) and (2.3), it follows that

$$m_{ij}(\theta,N) = \sum_{\pi(i)=j} \exp\left(\theta d(\pi,e_N) - \psi_N(\theta)\right) = \frac{(N-1)!\tilde{g}_{N-1}(\theta)}{N!g_N(\theta)} = \frac{1}{N}\frac{\tilde{g}_{N-1}(\theta)}{g_N(\theta)},$$

where $g_N(\cdot)$ and $\tilde{g}_{N-1}(\cdot)$ are the moment generating functions of $D_L(\pi)$ and $D_{i,j}(\sigma)$, for $\pi \sim Uniform(\mathbf{S}_N)$ and $\sigma \sim Uniform(\mathbf{S}_{i,j})$. Since $D_{i,j}$ depends on i and j only through $c_N(i,j)$, the random variables $D_{i,j}$ and $D_{N,k}$ are identically distributed for $k = N - c_N(i,j)$. From Theorem 1.2 and Lemma E.1, $g_N(\cdot)$ and $\tilde{g}_{N-1}(\cdot)$ can be approximated, so

$$m_{ij}(\theta,N)\frac{N}{\exp\left(\theta\mu + \frac{\theta^2\nu^2}{2}\right)} \xrightarrow[N\to\infty]{} 1,$$

where $\mu = \mathbf{E}(D_{i,j}) - \mathbf{E}(D_L)$ and $\nu^2 = \mathbf{Var}(D_{i,j}) - \mathbf{Var}(D_L)$.

According to Lemma E.1,

$$\mathbf{E}(D_{i,j}) = \frac{c_N(i,j)}{N-1} + \frac{N-2}{N-1}\left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right] + c_N(i,j),$$

$$\mathbf{Var}(D_{i,j}) = \frac{N^2\left(c_N(i,j)\right)^2 - 2N\left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right]c_N(i,j)}{(N-2)(N-1)^2} + \beta_{N-1}.$$

The values of μ and σ are obtained by combining the results above with formulas (1.11) and (1.13). □

E.2 Proofs – Chapter 4

Consider the random variables based on Spearman's footrule and Spearman's rho:

$$D_F(\pi) = \sum_{s=1}^{k} |\,\pi(s) - s\,| \quad \text{and} \quad D_R(\pi) = \sum_{s=1}^{k} (\pi(s) - s)^2,$$

where $\pi \sim Uniform(\mathbf{S_k})$. By applying Theorem 1.1 to D_F and D_R (see, e.g., Marden [47, p.83]) it can be shown that D_F and D_R are asymptotically normal with means and variances:

$$\mathbf{E}(D_F) = \frac{1}{k}\sum_{r=1}^{k}\sum_{s=1}^{k} \mid r-s \mid = \frac{k^2-1}{3}, \qquad \mathbf{Var}(D_F) = \frac{(k+1)\left(2k^2+7\right)}{45}, \tag{E.6}$$

$$\mathbf{E}(D_R) = \frac{1}{k}\sum_{r=1}^{k}\sum_{s=1}^{k} (r-s)^2 = \frac{k\left(k^2-1\right)}{6}, \qquad \mathbf{Var}(D_R) = \frac{k^2(k-1)(k+1)^2}{36}. \tag{E.7}$$

In order to prove Theorem 4.1, let us define the random variables $D_F^{(i,j)} = d_F(\pi, e_k)$ for $i, j = 1,2,\ldots,k$, where $d_F(\cdot,\cdot)$ is Spearman's footrule, and π is uniformly and randomly selected from $\mathbf{S_k^{(i,j)}} = \{\sigma \in \mathbf{S_k} : \sigma(j) = i\}$, i.e. $\pi \sim Uniform\left(\mathbf{S_k^{(i,j)}}\right)$. Then, for a fixed pair (i,j),

$$D_F^{(i,j)} = \sum_{s=1}^{k} \mid \pi(s)-s \mid = \sum_{\substack{s=1 \\ s\neq j}}^{k} \mid \pi(s)-s \mid + \mid i-j \mid = \sum_{s=1}^{k-1} \tilde{a}_k(\sigma(s),s) + \mid i-j \mid \ ,$$

where

$$\sigma(s) = \begin{cases} \pi(s), & \text{if } s < j \text{ and } \pi(s) < i, \\ \pi(s)-1, & \text{if } s < j \text{ and } \pi(s) > i, \\ \pi(s+1), & \text{if } s \geq j \text{ and } \pi(s+1) < i, \\ \pi(s+1)-1, & \text{if } s \geq j \text{ and } \pi(s+1) > i, \end{cases} \tag{E.8}$$

and

$$\tilde{a}_k(r,s) = \begin{cases} \mid r-s \mid, & \text{if } s < j \text{ and } r < i, \\ \mid r+1-s \mid, & \text{if } s < j \text{ and } r \geq i, \\ \mid r-s-1 \mid, & \text{if } s \geq j \text{ and } r < i, \\ \mid r+1-s-1 \mid, & \text{if } s \geq j \text{ and } r \geq i, \end{cases} \tag{E.9}$$

for $r,s = 1,2,\ldots,k-1$ and $\pi \sim Uniform\left(\mathbf{S_k^{(i,j)}}\right)$.

Lemma E.2. *Let* $\tilde{D}_F(\sigma) = \sum_{s=1}^{k-1} \tilde{a}_k(\sigma(s),s)$, *where* $\sigma(\cdot)$ *and* $\tilde{a}_k(\cdot,\cdot)$ *are given in* (E.8) *and* (E.9)*, respectively. Then the distribution of* $\tilde{D}_F$ *is asymptotically normal with mean and variance*

$$\mathbf{E}(\tilde{D}_F) = \frac{k(k+1)}{3} - \frac{f(i)+f(j)-\mid i-j \mid}{k-1}, \tag{E.10}$$

$$\mathbf{Var}(\tilde{D}_F) = \frac{1}{k-2}\left\{\sum_{\substack{r=1 \\ r\neq i}}^{k}\sum_{\substack{s=1 \\ s\neq j}}^{k}\left[\mid r-s \mid + \frac{k(k+1)}{3(k-1)} - \frac{f(r)+f(s)-\mid i-s \mid - \mid r-j \mid}{k-1} - \frac{f(i)+f(j)-\mid i-j \mid}{(k-1)^2}\right]^2\right\}, \tag{E.11}$$

where

$$f(x) = \frac{x(x-1)+(k-x)(k-x+1)}{2}.$$

Proof. From the definition of σ in (E.8) it is easy to check that $\sigma \sim Uniform(\mathbf{S_{k-1}})$ for $\pi \sim Uniform\left(\mathbf{S_k^{(i,j)}}\right)$. Therefore, Theorem 1.1 can be applied to the random variable $\tilde{D}_F$. By using (1.7), (E.9) and the expectation in (E.6), it follows that

$$\begin{aligned}\mathbf{E}\left(\tilde{D}_F\right) &\overset{(1.7)}{=} \frac{1}{k-1}\sum_{r=1}^{k-1}\sum_{s=1}^{k-1}\tilde{a}_k(r,s) \overset{(E.9)}{=} \frac{1}{k-1}\sum_{\substack{r=1\\r\neq i}}^{k}\sum_{\substack{s=1\\s\neq j}}^{k} \mid r-s \mid \\ &= \frac{1}{k-1}\sum_{r=1}^{k}\sum_{s=1}^{k} \mid r-s \mid -\frac{1}{k-1}\sum_{r=1}^{k} \mid r-j \mid -\frac{1}{k-1}\sum_{s=1}^{k} \mid i-s \mid +\frac{\mid i-j \mid}{k-1} \\ &\overset{(E.6)}{=} \frac{k(k+1)}{3} - \frac{f(i)+f(j)-\mid i-j \mid}{k-1},\end{aligned}$$

where

$$f(x) = \sum_{r=1}^{k} \mid r-x \mid = \frac{x(x-1)+(k-x)(k-x+1)}{2}, \tag{E.12}$$

for $x = 1,2,\ldots,k$. Using (1.8) of Theorem 1.1,

$$\mathbf{Var}\left(\tilde{D}_F\right) = \frac{1}{k-2}\sum_{\substack{r=1\\r\neq i}}^{k}\sum_{\substack{s=1\\s\neq j}}^{k}\tilde{b}_k^2(r,s), \tag{E.13}$$

where

$$\tilde{b}_k(r,s) = \mid r-s \mid -\sum_{\substack{l=1\\l\neq i}}^{k}\frac{\mid l-s \mid}{k-1} - \sum_{\substack{m=1\\m\neq j}}^{k}\frac{\mid r-m \mid}{k-1} + \frac{1}{(k-1)^2}\sum_{\substack{l=1\\l\neq i}}^{k}\sum_{\substack{m=1\\m\neq j}}^{k} \mid l-m \mid,$$

for $r,s = 1,2,\ldots,k$. Simplifying this expression gives

$$\tilde{b}_k(r,s) = \mid r-s \mid -\frac{f(r)+f(s)-\mid i-s \mid -\mid r-j \mid}{k-1} + \frac{k(k+1)}{3(k-1)} - \frac{f(i)+f(j)-\mid i-j \mid}{(k-1)^2}. \tag{E.14}$$

The variance of $\tilde{D}_F$ given in (E.11) is obtained by substituting (E.14) in formula (E.13).

From (E.12) it is easy to check that

$$\frac{k^2-1}{4} \leq f(x) \leq \frac{k(k-1)}{2} \quad \text{for } 1 \leq x \leq k. \tag{E.15}$$

Combining

$$1 \leq \mid x-y \mid \leq k-1 \quad \text{for } 1 \leq x,y \leq k,$$

with (E.14) and (E.15), it follows that

$$\mid r-s \mid -\frac{2k}{3} + \varepsilon_1 \leq \tilde{b}_k(r,s) \leq \mid r-s \mid -\frac{k}{6} + \varepsilon_2,$$

where $r,s = 1,2,\ldots,k$, $\lim\limits_{k\to\infty}\dfrac{\varepsilon_1}{k} = 0$ and $\lim\limits_{k\to\infty}\dfrac{\varepsilon_2}{k} = 0$. Therefore, there exists a constant $c_1 > 0$ such that

$$\max_{1\leq r,s\leq k}\tilde{b}_k^2(r,s) \leq c_1 k^2, \tag{E.16}$$

and a number $N > 0$ such that for $k \geq N$

$$| r-s | -k \leq \tilde{b}_k(r,s) \leq | r-s | -\frac{k}{7} .$$

Suppose that r is a fixed index from the set $\{1,2,\ldots,k\}$. Then for $k \geq N$

$$\sum_{\substack{s=1 \\ s\neq j}}^{k} \tilde{b}_k^2(r,s) = \sum_{s=1}^{k} \tilde{b}_k^2(r,s) - \tilde{b}_k^2(r,j) \geq \sum_{|r-s|=0}^{k/7} \tilde{b}_k^2(r,s) - \tilde{b}_k^2(r,j) \geq \sum_{v=0}^{k/7} \left(v - \frac{k}{7}\right)^2 - \tilde{b}_k^2(r,j),$$

where $\sum_{|r-s|=0}^{k/7}$ is a summation over all values of s such that $0 \leq | r-s | \leq \frac{k}{7}$. Thus, for $k \geq N$ there exists a constant $c_2 > 0$ such that

$$\sum_{\substack{s=1 \\ s\neq j}}^{k} \tilde{b}_k^2(r,s) \geq c_2 k^3 . \tag{E.17}$$

By using (E.16) and (E.17),

$$\lim_{k\to\infty} \frac{\max\limits_{1\leq r,s\leq k} \tilde{b}_k^2(r,s)}{\frac{1}{k-1} \sum\limits_{\substack{r=1 \\ r\neq i}}^{k} \sum\limits_{\substack{s=1 \\ s\neq j}}^{k} \tilde{b}_k^2(r,s)} \leq \lim_{k\to\infty} \frac{c_1 k^2}{\frac{1}{k-1} \sum\limits_{\substack{r=1 \\ r\neq i}}^{k} c_2 k^3} = 0,$$

i.e. the condition (1.9) of Theorem 1.1 is fulfilled and the distribution of $\tilde{D}_F$ is asymptotically normal. □

Similarly to $\left\{D_F^{(i,j)}\right\}_{i,j=1}^{k}$, consider the random variables $D_R^{(i,j)} = d_R\big(\pi, e_k\big)$ based on Spearman's rho. For a fixed pair (i,j),

$$D_R^{(i,j)} = \sum_{s=1}^{k} \big(\pi(s) - s\big)^2 = \sum_{\substack{s=1 \\ s\neq j}}^{k} \big(\pi(s) - s\big)^2 + (i-j)^2 = \sum_{s=1}^{k-1} \bar{a}_k(\sigma(s), s) + (i-j)^2 ,$$

where

$$\bar{a}_k(r,s) = \begin{cases} (r-s)^2, & \text{if } s < j \text{ and } r < i, \\ (r+1-s)^2, & \text{if } s < j \text{ and } r \geq i, \\ (r-s-1)^2, & \text{if } s \geq j \text{ and } r < i, \\ (r+1-s-1)^2, & \text{if } s \geq j \text{ and } r \geq i, \end{cases} \tag{E.18}$$

for $r,s = 1,2,\ldots,k-1$, $\pi \sim Uniform\left(\mathbf{S}_{\mathbf{k}}^{(i,j)}\right)$ and $\sigma(s)$ is defined as in (E.8).

Lemma E.3. *Let* $\bar{D}_R(\sigma) = \sum_{s=1}^{k-1} \bar{a}_k(\sigma(s), s)$, *where* $\sigma(\cdot)$ *and* $\bar{a}_k(\cdot,\cdot)$ *are given in* (E.8) *and* (E.18)*, respectively. Then the distribution of* $\bar{D}_R$ *is asymptotically normal with mean and*

variance

$$\mathbf{E}\,(\bar{D}_R) = \frac{k^2(k+1)}{6} - \frac{h(i)+h(j)-(i-j)^2}{k-1},$$

$$\mathbf{Var}\,(\bar{D}_R) = \frac{1}{k-2}\left\{\sum_{\substack{r=1\\ r\neq i}}^{k}\sum_{\substack{s=1\\ s\neq j}}^{k}\left[(r-s)^2 + \frac{k^2(k+1)}{6(k-1)} - \frac{h(r)+h(s)-(i-s)^2-(r-j)^2}{k-1} - \frac{h(i)+h(j)-(i-j)^2}{(k-1)^2}\right]^2\right\},$$

where

$$h(x) = \frac{x(x-1)(2x-1)+(k-x)(k-x+1)(2k-2x+1)}{6}.$$

Proof. By using (E.7) and the fact that for $x \in \{1,2,\ldots,k\}$

$$h(x) = \sum_{r=1}^{k}(r-k)^2 = \frac{x(x-1)(2x-1)+(k-x)(k-x+1)(2k-2x+1)}{6}, \tag{E.19}$$

$\mathbf{E}\,(\bar{D}_R)$ and $\mathbf{Var}\,(\bar{D}_R)$ can be evaluated in a similar way as in the proof of Lemma E.2.

Now, consider the quantities

$$\bar{b}_k(r,s) = (r-s)^2 + \frac{k^2(k+1)}{6(k-1)} - \frac{h(r)+h(s)-(i-s)^2-(r-j)^2}{k-1} - \frac{h(i)+h(j)-(i-j)^2}{(k-1)^2}$$

for $r,s = 1,2,\ldots,k$. Since (E.19)

$$\frac{k(k^2+2)}{12} \le h(x) \le \frac{k(k-1)(2k-1)}{6} \quad \text{for } 1 \le x \le k,$$

it follows that

$$(r-s)^2 - \frac{k^2}{2} + \varepsilon_1 \le \bar{b}_k(r,s) \le (r-s)^2 + \varepsilon_2,$$

where $r,s = 1,2,\ldots,k$, $\lim\limits_{k\to\infty}\dfrac{\varepsilon_1}{k^2} = 0$ and $\lim\limits_{k\to\infty}\dfrac{\varepsilon_2}{k^2} = 0$. Hence, there exists a constant $c_1 > 0$ such that

$$\max_{1\le r,s\le k}\bar{b}_k^2(r,s) \le c_1 k^4. \tag{E.20}$$

Further, fix the indexes $1 \le r,s \le \dfrac{k}{4}$. Then, since

$$\frac{k(7k^2+12k+8)}{48} \le h(x) \le \frac{k(k-1)(2k-1)}{6} \quad \text{for } 1 \le x \le \frac{k}{4},$$

it follows that

$$(r-s)^2 - \frac{k^2}{2} + \varepsilon_3 \le \bar{b}_k(r,s) \le (r-s)^2 - \frac{k^2}{8} + \varepsilon_4,$$

where $\lim\limits_{k\to\infty}\dfrac{\varepsilon_3}{k^2} = 0$ and $\lim\limits_{k\to\infty}\dfrac{\varepsilon_4}{k^2} = 0$. Thus, there exists a number $N > 0$ such that for $k \ge N$

$$(r-s)^2 - k^2 \le \bar{b}_k(r,s) \le (r-s)^2 - \frac{k^2}{9}.$$

Hence, for $k \geq N$

$$\sum_{\substack{r=1\\r\neq i}}^{k}\sum_{\substack{s=1\\s\neq j}}^{k}\bar{b}_k^2(r,s) \geq \sum_{\substack{r=1\\r\neq i}}^{k/4}\sum_{\substack{s=1\\s\neq j}}^{k/4}\bar{b}_k^2(r,s) = \sum_{\substack{r=1\\r\neq i}}^{k/4}\left\{\sum_{s=1}^{k/4}\bar{b}_k^2(r,s) - \bar{b}_k^2(r,j)\right\} \geq$$

$$\geq \sum_{\substack{r=1\\r\neq i}}^{k/4}\left\{\sum_{(r-s)^2=0}^{k^2/9}\bar{b}_k^2(r,s) - \bar{b}_k^2(r,j)\right\} \geq \sum_{\substack{r=1\\r\neq i}}^{k/4}\left\{\sum_{v=0}^{k/3}\left(v^2-\frac{k^2}{9}\right)^2 - \bar{b}_k^2(r,j)\right\},$$

where $\sum_{(r-s)^2=0}^{k^2/9}$ is a summation over all values of s, such that $0 \leq (r-s)^2 \leq \frac{k^2}{9}$. Thus, for $k \geq N$ there exists a constant $c_2 > 0$, such that

$$\sum_{\substack{r=1\\r\neq i}}^{k}\sum_{\substack{s=1\\s\neq j}}^{k}\bar{b}_k^2(r,s) \geq c_2 k^6\,. \tag{E.21}$$

From (E.20) and (E.21), it is easy to check that the condition (1.9) of Theorem 1.1 is fulfilled and the distribution of $\bar{D}_R$ is asymptotically normal. □

of Theorem 4.1. From (4.5), (2.3) and (4.6), it follows that

$$q(i,j,k,\theta) = \sum_{\pi(j)=i}\exp\left(\theta d(\pi,e_k) - \psi_k(\theta)\right) = \frac{(k-1)!\tilde{m}_{k-1}(\theta)}{k!m_k(\theta)} = \frac{1}{k}\frac{\tilde{m}_{k-1}(\theta)}{m_k(\theta)},$$

where $m_k(\cdot)$ and $\tilde{m}_{k-1}(\cdot)$ are the moment generating functions of $D_F(\pi)$ and $D_F^{(i,j)}(\sigma)$ for $\pi \sim Uniform(\mathbf{S_k})$ and $\sigma \sim Uniform\left(\mathbf{S_k^{(i,j)}}\right)$. Since $D_F^{(i,j)} = \tilde{D}_F + \mid i-j \mid$ and according to Lemma E.2 $\tilde{D}_F$ is asymptotically normal, it follows that $D_F^{(i,j)}$ is asymptotically normal. Therefore, $m_k(\cdot)$ and $\tilde{m}_{k-1}(\cdot)$ can be approximated with the moment generating function of the normal distribution and

$$q(i,j,k,\theta)\frac{k}{\exp\left(\theta\mu + \frac{\theta^2\nu^2}{2}\right)} \xrightarrow[k\to\infty]{} 1,$$

where $\mu = \mathbf{E}\left(D_F^{(i,j)}\right) - \mathbf{E}\left(D_F\right)$ and $\nu^2 = \mathbf{Var}\left(D_F^{(i,j)}\right) - \mathbf{Var}\left(D_F\right)$.

The values of μ and ν^2 given in Theorem 4.1 are obtained by combining formulas (E.6), (E.10) and (E.11) with

$$\mathbf{E}\left(D_F^{(i,j)}\right) = \mathbf{E}\left(\tilde{D}_F\right) + \mid i-j \mid \quad \text{and} \quad \mathbf{Var}\left(D_F^{(i,j)}\right) = \mathbf{Var}\left(\tilde{D}_F\right).$$

□

Proof of Theorem 4.2. The proof is similar to the proof of Theorem 4.1 using Lemma E.3. □

E.3 Proofs – Chapter 5

Proof of Theorem 5.1. Let us first notice that $P(L_{m,n}=l)=P(L_{n,m}=l)$ for all values of m, n and l. Furthermore, from Proposition 5.1 it follows that $P(L_{0,n}=0)=1$, i.e. $L_{0,n}$ is the constant 0, and

$$P(L_{1,n}=l)=\begin{cases}\dfrac{1}{n+1}, & \text{for } l=0,\\ \dfrac{2}{n+1}, & \text{for } l=1,2,\ldots,\dfrac{n}{2},\end{cases} \tag{E.22}$$

when n is even and

$$P(L_{1,n}=l)=\begin{cases}\dfrac{1}{n+1}, & \text{for } l=0,\dfrac{n+1}{2}\\ \dfrac{2}{n+1}, & \text{for } l=1,2,\ldots,\dfrac{n-1}{2},\end{cases} \tag{E.23}$$

when n is odd.

By multiplying (5.15) by l and summing for $l=0,1,\ldots,\left[\frac{mn+1}{2}\right]$, we have that

$$\begin{aligned}\sum_{l=0}^{\left[\frac{mn+1}{2}\right]} lP(L_{m,n}=l) =&\mathbf{E}(L_{m,n})\\ =&\frac{m}{m+n}\mathbf{E}(L_{m-1,n})+\frac{n}{m+n}\mathbf{E}(L_{m,n-1})\\ &+\frac{mn}{(m+n)(m+n-1)}\left\{\mathbf{E}\left(L_{m-1,n-1}+\frac{m+n-1}{2}\right)-\mathbf{E}(L_{m-1,n-1})\right\}\\ =&\frac{m}{m+n}\mathbf{E}(L_{m-1,n})+\frac{n}{m+n}\mathbf{E}(L_{m,n-1})+\frac{mn}{2(m+n)},\end{aligned} \tag{E.24}$$

when $m+n$ is odd and

$$\mathbf{E}(L_{m,n})=\frac{m}{m+n}\mathbf{E}(L_{m-1,n})+\frac{n}{m+n}\mathbf{E}(L_{m,n-1})+\frac{mn}{2(m+n-1)}, \tag{E.25}$$

when $m+n$ is even. Substituting (E.25) for $\mathbf{E}(L_{m-1,n})$ and $\mathbf{E}(L_{m,n-1})$ in (E.24) gives

$$\begin{aligned}\mathbf{E}(L_{m,n}) =&\frac{m(m-1)}{(m+n)(m+n-1)}\mathbf{E}(L_{m-2,n})+\frac{n(n-1)}{(m+n)(m+n-1)}\mathbf{E}(L_{m,n-2})\\ &+\frac{2mn}{(m+n)(m+n-1)}\mathbf{E}(L_{m-1,n-1})+\frac{mn}{(m+n)},\end{aligned} \tag{E.26}$$

when $m+n$ is odd. From Proposition 5.1, (E.22) and (E.23) it follows that

$$\mathbf{E}(L_{0,1})=\mathbf{E}(L_{1,0})=0,\quad \mathbf{E}(L_{3,2})=\mathbf{E}(L_{2,3})=\frac{9}{5},\quad \mathbf{E}(L_{1,n})=\begin{cases}\dfrac{n(n+2)}{4(n+1)}, & \text{if } n \text{ is even},\\ \dfrac{n+1}{4}, & \text{if } n \text{ is odd}.\end{cases}$$

By using the mean values above and (E.26) it is not hard to prove by induction that

$$\mathbf{E}(L_{m,n})=\frac{mn(m+n+1)}{4(m+n)}, \tag{E.27}$$

when $m+n$ is odd. The mean of $L_{m,n}$, given in (5.16), is derived from (E.27) and (E.25).

Formula (5.17) can be proved in a similar way. By multiplying (5.15) by l^2, summing for $l = 0, 1, \ldots, \left[\frac{mn+1}{2}\right]$ and using (5.16), we get

$$\begin{aligned}\mathbf{E}\left(L_{m,n}^2\right) =& \frac{m}{m+n}\mathbf{E}\left(L_{m-1,n}^2\right) + \frac{n}{m+n}\mathbf{E}\left(L_{m,n-1}^2\right) \\ &+ \frac{mn}{(m+n)(m+n-1)}\left\{\mathbf{E}\left[\left(L_{m-1,n-1} + \frac{m+n-1}{2}\right)^2\right] - \mathbf{E}\left(L_{m-1,n-1}^2\right)\right\} \\ =& \frac{m}{m+n}\mathbf{E}\left(L_{m-1,n}^2\right) + \frac{n}{m+n}\mathbf{E}\left(L_{m,n-1}^2\right) + \frac{mn(m+n-1)(mn-1)}{4(m+n)(m+n-2)}, \end{aligned} \tag{E.28}$$

when $m+n$ is odd and

$$\begin{aligned}\mathbf{E}\left(L_{m,n}^2\right) =& \frac{m}{m+n}\mathbf{E}\left(L_{m-1,n}^2\right) + \frac{n}{m+n}\mathbf{E}\left(L_{m,n-1}^2\right) \\ &+ \frac{mn(m-1)(n-1)}{4(m+n)(m+n-1)}\left\{4 + \frac{(m+n-2)(m+n-4)}{(m+n-3)} + \frac{(m+n)^2}{(m-1)(n-1)}\right\}, \end{aligned} \tag{E.29}$$

when $m+n$ is even. If we substitute $\mathbf{E}\left(L_{m-1,n}^2\right)$ and $\mathbf{E}\left(L_{m,n-1}^2\right)$ from (E.28) to (E.29), we get

$$\begin{aligned}\mathbf{E}\left(L_{m,n}^2\right) =& \frac{m(m-1)}{(m+n)(m+n-1)}\mathbf{E}\left(L_{m-2,n}^2\right) + \frac{n(n-1)}{(m+n)(m+n-1)}\mathbf{E}\left(L_{m,n-2}^2\right) \\ &+ \frac{2mn}{(m+n)(m+n-1)}\mathbf{E}\left(L_{m-1,n-1}^2\right) + \frac{(mn)^2}{2(m+n)} - \frac{mn\{mn-2(m+n-2)\}}{2(m+n)(m+n-1)(m+n-3)}, \end{aligned} \tag{E.30}$$

when $m+n$ is even. From Proposition 5.1, (E.22) and (E.23) it follows that

$$\mathbf{E}\left(L_{2,0}^2\right) = \mathbf{E}\left(L_{0,2}^2\right) = 0, \quad \mathbf{E}\left(L_{2,2}^2\right) = \frac{7}{3}, \quad \mathbf{E}\left(L_{1,n}^2\right) = \begin{cases} \dfrac{n(n+2)}{12}, & \text{if } n \text{ is even,} \\ \dfrac{n^2+2n+3}{12}, & \text{if } n \text{ is odd.} \end{cases}$$

By using the expected values above and (E.30) it can be proved by induction that

$$\mathbf{E}\left(L_{m,n}^2\right) = \frac{mn\left\{(m+n-1)^3 + 3mn(m+n-1)^2 - 15mn + 11m + 11n - 35\right\}}{48(m+n-1)(m+n-3)}, \tag{E.31}$$

when $m+n$ is even. From (E.28) and (E.31) it follows that

$$\mathbf{E}\left(L_{m,n}^2\right) = \frac{mn\left\{(m+n)^3 + (3mn-1)(m+n)^2 - 15mn + 7m + 7n - 15\right\}}{48(m+n)(m+n-2)}, \tag{E.32}$$

when $m+n$ is odd. Formula (5.17) is obtained by combining (5.16), (E.31) and (E.32).

Let $\phi_{m,n}(t)$ be the moment generating function of $L_{m,n}$, i.e. $\phi_{m,n}(t) = \mathbf{E}\{\exp(tL_{m,n})\}$. We will prove by induction that

$$\phi_{m,n}\left(\frac{t}{\sigma_{m,n}}\right)\exp\left(-\mu_{m,n}\frac{t}{\sigma_{m,n}}\right) = \exp\left(\frac{t^2}{2}\right)\left\{1 + O\left(\frac{t}{\sqrt{\min(m,n)}}\right)\right\}, \tag{E.33}$$

where $\mu_{m,n} = \mathbf{E}(L_{m,n})$, $\sigma^2_{m,n} = \mathbf{Var}(L_{m,n})$ and $O\left(\frac{t}{\sqrt{\min(m,n)}}\right)$ is the class of functions such that if $f \in O\left(\frac{t}{\sqrt{\min(m,n)}}\right)$, then $\mid f(t,m,n) \mid < A\frac{t}{\sqrt{\min(m,n)}}$ for some constant A and for $t < 1$.

First, by multiplying (5.15) by $\exp(lt)$ and summing for $l = 0,1,\ldots,\left[\frac{mn+1}{2}\right]$, we obtain that

$$\phi_{m,n}(t) = \frac{m}{m+n}\phi_{m-1,n}(t) + \frac{n}{m+n}\phi_{m,n-1}(t) + \frac{mn}{(m+n)(m+n-1)}\phi_{m-1,n-1}(t)\left\{\exp\left(\frac{m+n-1}{2}t\right) - 1\right\}, \quad \text{(E.34)}$$

if $m+n$ is odd and

$$\begin{aligned}\phi_{m,n}(t) =& \frac{m}{m+n}\phi_{m-1,n}(t) + \frac{n}{m+n}\phi_{m,n-1}(t) + \frac{mn}{(m+n)(m+n-1)}\phi_{m-1,n-1}(t)\{\exp(t) - 1\} \\ &+ \frac{mn(m-1)}{(m+n)(m+n-1)(m+n-2)}\phi_{m-2,n-1}(t)\left\{\exp\left(\frac{m+n}{2}t\right) - \exp(t)\right\} \\ &+ \frac{mn(n-1)}{(m+n)(m+n-1)(m+n-2)}\phi_{m-1,n-2}(t)\left\{\exp\left(\frac{m+n}{2}t\right) - \exp(t)\right\} \\ &+ \frac{mn(m-1)(n-1)}{(m+n)(m+n-1)(m+n-2)(m+n-3)}\phi_{m-2,n-2}(t)\times \\ &\times\left\{\exp(t) - \exp\left(\frac{m+n}{2}t\right) - \exp\left(\frac{m+n-2}{2}t\right) + \exp((m+n-2)t)\right\},\end{aligned} \quad \text{(E.35)}$$

if $m+n$ is even.

From (E.22) and (E.23) we have that

$$\phi_{1,n}(t) = \begin{cases} \dfrac{2\exp\left(\frac{n+2}{2}t\right) - \exp(t) - 1}{(n+1)(\exp(t)-1)}, & \text{if } n \text{ is even,} \\ \dfrac{\exp\left(\frac{n+3}{2}t\right) + \exp\left(\frac{n+1}{2}t\right) - \exp(t) - 1}{(n+1)(\exp(t)-1)}, & \text{if } n \text{ is odd,} \end{cases}$$

and it is easy to check that (E.33) holds for $m = 1$. By substituting t with $\frac{t}{\sigma_{m,n}}$ in (E.34) and using the induction assumption, we get

$$\begin{aligned}
&\phi_{m,n}\left(\frac{t}{\sigma_{m,n}}\right)\exp\left(-\mu_{m,n}\frac{t}{\sigma_{m,n}}\right) = \\
&= \frac{m}{m+n}\exp\left(\frac{t^2}{2}+\frac{(\mu_{m-1,n}-\mu_{m,n})\,t}{\sigma_{m,n}}\right)\left\{1+O\left(\frac{t\sigma_{m-1,n}}{\sigma_{m,n}\sqrt{\min(m-1,n)}}\right)\right\} \\
&\quad+\frac{n}{m+n}\exp\left(\frac{t^2}{2}+\frac{(\mu_{m,n-1}-\mu_{m,n})\,t}{\sigma_{m,n}}\right)\left\{1+O\left(\frac{t\sigma_{m,n-1}}{\sigma_{m,n}\sqrt{\min(m,n-1)}}\right)\right\} \\
&\quad+\frac{mn}{(m+n)(m+n-1)}\exp\left(\frac{t^2}{2}+\frac{(\mu_{m-1,n-1}-\mu_{m,n})\,t}{\sigma_{m,n}}\right)\left\{\exp\left(\frac{(m+n-1)t}{2\sigma_{m,n}}\right)-1\right\}\times \\
&\qquad\qquad\times\left\{1+O\left(\frac{t\sigma_{m-1,n-1}}{\sigma_{m,n}\sqrt{\min(m-1,n-1)}}\right)\right\} \\
&= \exp\left(\frac{t^2}{2}\right)\left\{1+\frac{m\mu_{m-1,n}+n\mu_{m,n-1}-(m+n)\mu_{m,n}+\frac{mn}{2}}{(m+n)\sigma_{m,n}}\,t+O\left(\frac{t\sigma_{m-1,n-1}}{\sigma_{m,n}\sqrt{\min(m,n)-1}}\right)\right\}.
\end{aligned}$$

Since

$$\begin{gathered}
m\mu_{m-1,n}+n\mu_{m,n-1}-(m+n)\mu_{m,n}+\frac{mn}{2}=0, \\
\frac{\mu_{m-1,n-1}-\mu_{m,n}}{-\frac{m+n}{4}}\to 1 \quad\text{and}\quad \frac{\sigma_{m,n}}{\sqrt{\frac{mn(m+n)}{48}}}\to 1 \quad\text{as}\quad m,n\to\infty,
\end{gathered}$$

it follows that

$$\phi_{m,n}\left(\frac{t}{\sigma_{m,n}}\right)\exp\left(-\mu_{m,n}\frac{t}{\sigma_{m,n}}\right)=\exp\left(\frac{t^2}{2}\right)\left\{1+O\left(\frac{t}{\sqrt{\min(m,n)}}\right)\right\},$$

which proves the induction step when $m+n$ is odd. The induction step for even $m+n$ follows from (E.35) in a similar way.

The asymptotic normality of $\frac{L_{m,n}-\mu_{m,n}}{\sigma_{m,n}}$ as $m,n\to\infty$ is obtained by combining (E.33) and the fact that the moment generating function of the standard normal distribution is $\exp\left(\frac{t^2}{2}\right)$.

□

Bibliography

[1] M. Alvo and P. L. H. Yu (2014). *Statistical Methods for Ranking Data*. Springer, NY.

[2] M. Aragon, G. Patil, and C. Taillie (1999). A performance indicator for ranked set sampling using ranking error probability matrix. *Environmental and Ecological Statistics*. Vol. 6 (1), 75–80.

[3] N. Balakrishnan and T. Li (2008). Ordered ranked set samples and applications to inference. *Journal of Statistical Planning and Inference*. Vol. 138 (11), 3512–3524.

[4] S. Beggs, S. Cardell, and J. Hausman (1981). Assessing the potential demand for electric cars. *Journal of Econometrics*. Vol. 17 (1), 1–19.

[5] R. A. Bradley and M. E. Terry (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*. Vol. 39 (3), 324–345.

[6] L. M. Busse, P. Orbanz, and J. M. Buhmann (2007). Cluster analysis of heterogeneous rank data. In: *Proceedings of the 24-th International Conference on Machine Learning*, pp. 113–120.

[7] C. H. Chan, F. Yan, J. Kittler, and K. Mikolajczyk (2015). Full ranking as local descriptor for visual recognition: A comparison of distance metrics on S_n. *Pattern Recognition*. Vol. 48 (4), 1328–1336.

[8] R. G. Chapman and R. Staelin (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*. Vol. 19 (3), 288–301.

[9] Z. Chen, Z. Bai, and B. Sinha (2003). *Ranked Set Sampling: Theory and Applications.* Lecture Notes in Statistics. Vol. 176. Springer, NY.

[10] D. E. Critchlow (2012). *Metric Methods for Analyzing Partially Ranked Data.* Lecture Notes in Statistics. Vol. 34. Springer, NY.

[11] D. E. Critchlow (1986). *A Unified Approach to Constructing Nonparametric Rank Tests*. Tech. rep. Stanford University Press, Redwood City.

[12] D. E. Critchlow (1992). On rank statistics: an approach via metrics on the permutation group. *Journal of Statistical Planning and Inference*. Vol. 32 (3), 325–346.

[13] M. A. Croon and R. Luijkx (1993). Latent structure models for ranking data. In: *Fligner M. A., Verducci J. S. (eds) Probability Models and Statistical Analyses for Ranking Data.* Lecture Notes in Statistics. Vol. 80. Springer, pp. 53–74.

[14] T. Dell and J. Clutter (1972). Ranked set sampling theory with order statistics background. *Biometrics*. Vol. 28 (2), 545–555.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*. Vol. 39 (1), 1–22.

[16] M. Deza and T. Huang (1998). Metrics on permutations, a survey. *Journal of Combinatorics, Information and System Sciences*. Vol. 23, 173–185.

[17] P. Diaconis (1988). *Group representations in probability and statistics*. Institute of Mathematical Statistics.

[18] P. Diaconis (1989). A generalization of spectral analysis with application to ranked data. *The Annals of Statistics*. Vol. 17 (3), 949–979.

[19] P. Diaconis and R. L. Graham (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society: Series B*. Vol. 39 (2), 262–268.

[20] R. A. Fisher (1936). The coefficient of racial likeness and the future of craniometry. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*. Vol. 66, 57–63.

[21] M. A. Fligner and J. S. Verducci (1986). Distance based ranking models. *Journal of the Royal Statistical Society: Series B*. Vol. 48 (3), 359–369.

[22] D. Freedman and D. Lane (1980). The empirical distribution of fourier coefficients. *The Annals of Statistics*. Vol. 8 (6), 1244–1251.

[23] J. Frey (2012). Nonparametric mean estimation using partially ordered sets. *Environmental and Ecological Statistics*. Vol. 19 (3), 309–326.

[24] J. Frey, O. Ozturk, and J. V. Deshpande (2007). Nonparametric tests for perfect judgment rankings. *Journal of the American Statistical Association*. Vol. 102 (478), 708–717.

[25] J. Frey and L. Wang (2013). Most powerful rank tests for perfect rankings. *Computational Statistics & Data Analysis*. Vol. 60, 157–168.

[26] J. C. Frey (2007). New imperfect rankings models for ranked set sampling. *Journal of Statistical planning and Inference*. Vol. 137 (4), 1433–1445.

[27] J. D. Gibbons and S. Chakraborti (2010). *Nonparametric Statistical Inference*. Taylor & Francis, Boca Raton.

[28] P. Good (2000). *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, NY.

[29] I. C. Gormley and T. B. Murphy (2008). Exploring voting blocs within the Irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*. Vol. 103 (483), 1014–1027.

[30] J. Hájek and Z. Šidák (1967). *Theory of Rank Tests*. Academic Press, NY.

[31] J. A. Hartigan (1975). *Clustering Algorithms*. Wiley, NY.

[32] W. Hoeffding (1951). A combinatorial central limit theorem. *The Annals of Mathematical Statistics*. Vol. 22 (4), 558–566.

[33] M. Hollander and D. A. Wolfe (1999). *Nonparametric Statistical Methods*. Wiley, NY.

[34] L. J. Hubert and J. R. Levin (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*. Vol. 83 (6), 1072.

[35] R. Inglehart (1977). *The Silent Revolution: Changing Values and Political Styles Among Western Publics*. Princeton University Press, New Jersey.

[36] E. Irurozki, B. Calvo, and J. A. Lozano (2014). *Sampling and Learning the Mallows and Weighted Mallows Models Under the Hamming Distance*. Tech. rep. University of the Basque Country, Bilbao.

[37] N. L. Johnson, S. Kotz, and N. Balakrishnan (1994). *Continuous univariate distributions*. Vol. 1. Wiley, NY.

[38] T. Kamishima and S. Akaho (2009). Efficient clustering for orders. In: *Studies in Computational Intelligence*. Vol. 165. Springer, pp. 261–279.

[39] A. Klementiev, D. Roth, and K. Small (2007). An unsupervised learning algorithm for rank aggregation. In: *European Conference on Machine Learning*. Springer, pp. 616–623.

[40] G. Koop and D. Poirier (1994). Rank-ordered logit models: An empirical analysis of Ontario voter preferences. *Journal of Applied Econometrics*. Vol. 9 (4), 369–388.

[41] W. H. Kruskal (1958). Ordinal measures of association. *Journal of the American Statistical Association*. Vol. 53 (284), 814–861.

[42] C. Y. Lee (1961). An algorithm for path connections and its applications. *IRE Transactions on Electronic Computers*. Vol. 10 (3), 346–365.

[43] T. Li and N. Balakrishnan (2008). Some simple nonparametric methods to test for perfect ranking in ranked set sampling. *Journal of Statistical Planning and Inference*. Vol. 138 (5), 1325–1338.

[44] R. D. Luce (1959). *Individual Choice Behavior*. Wiley, NY.

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

[46] A. Mao, A. D. Procaccia, and Y. Chen (2013). Better human computation through principled voting. In: *Twenty-Seventh AAAI Conference on Artificial Intelligence*, pp. 1142–1148.

[47] J. I. Marden (1995). *Analyzing and Modeling Rank Data.* Monographs on Statistics and Applied Probability. Vol. 64. Chapman & Hall, London.

[48] M. Marozzi (2004). A bi-aspect nonparametric test for the two-sample location problem. *Computational Statistics & Data Analysis*. Vol. 44 (4), 639–648.

[49] N. Mattei and T. Walsh (2013). Preflib: A library for preferences http://preflib.org. In: *International Conference on Algorithmic Decision Theory*. Springer, pp. 259–270.

[50] A. Maydeu-Olivares and U. Böckenholt (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*. Vol. 10 (3), 285.

[51] G. McIntyre (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*. Vol. 3 (4), 385–390.

[52] G. McLachlan and T. Krishnan (2007). *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics. Vol. 382. Wiley, NY.

[53] G. Moors and J. Vermunt (2007). Heterogeneity in post-materialist value priorities. Evidence from a latent class discrete choice approach. *European Sociological Review*. Vol. 23 (5), 631–648.

[54] S. Mukherjee (2016). Estimation in exponential families on permutations. *The Annals of Statistics*. Vol. 44 (2), 853–875.

[55] T. B. Murphy and D. Martin (2003). Mixtures of distance-based models for ranking data. *Computational Statistics & Data Analysis*. Vol. 41 (3), 645–655.

[56] R. Murray, M. Ridout, J. Cross, et al. (2000). The use of ranked set sampling in spray deposit assessment. *Aspects of Applied Biology*. Vol. 57, 141–146.

[57] N. I. Nikolov (2016). Lee distance in two-sample rank tests. In: *Computer Data Analysis and Modeling: Theoretical and Applied Stochastics: Proceedings of the Eleventh International Conference*. Minsk: Publishing Center of BSU, pp. 100–103.

[58] N. I. Nikolov and E. Stoimenova (2017). Mallows' model based on Lee distance. In: *Proceedings of 20-th European Young Statisticians Meeting*, pp. 59–66.

[59] N. I. Nikolov and E. Stoimenova (2019a). Asymptotic properties of Lee distance. *Metrika*. Vol. 82 (3), 385–408.

[60] N. I. Nikolov and E. Stoimenova (2019b). EM estimation of the parameters in latent Mallows' models. In: *Studies in Computational Intelligence,* Springer Series. Vol. 793, pp. 317–325.

[61] N. I. Nikolov and E. Stoimenova (2019c). Mallows' models for imperfect ranking in ranked set sampling. *AStA Advances in Statistical Analysis*. `https://doi.org/10.1007/s10182-019-00354-4`, 1–26.

[62] N. I. Nikolov and E. Stoimenova (2020). Rank data clustering based on Lee distance. In: *Proceedings of 13-th Annual Meeting of the Bulgarian Section of SIAM*, pp. 1–11, (accepted).

[63] S. Nombekela, M. Murphy, H. Gonyou, and J. Marden (1994). Dietary preferences in early lactation cows as affected by primary tastes and some common feed flavors. *Journal of Dairy Science*. Vol. 77 (8), 2393–2399.

[64] O. Ozturk (2007). Statistical inference under a stochastic ordering constraint in ranked set sampling. *Journal of Nonparametric Statistics*. Vol. 19 (3), 131–144.

[65] O. Ozturk (2010). Nonparametric maximum-likelihood estimation of within-set ranking errors in ranked set sampling. *Journal of Nonparametric Statistics*. Vol. 22 (7), 823–840.

[66] O. Ozturk (2011). Sampling from partially rank-ordered sets. *Environmental and Ecological statistics*. Vol. 18 (4), 757–779.

[67] F. Pesarin and L. Salmaso (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*. Wiley, NY.

[68] R. L. Plackett (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C*. Vol. 24 (2), 193–202.

[69] A. Plumb, F. Grieve, and S. Khan (2009). Survey of hospital clinicians' preferences regarding the format of radiology reports. *Clinical Radiology*. Vol. 64 (4), 386–394.

[70] M. Regenwetter, M.-H. R. Ho, and I. Tsetlin (2007). Sophisticated approval voting, ignorance priors, and plurality heuristics: A behavioral social choice analysis in a Thurstonian framework. *Psychological Review*. Vol. 114 (4), 994.

[71] J. A. Salomon (2003). Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Population Health Metrics*. Vol. 1 (1), 12.

[72] P. Skowron, P. Faliszewski, and A. Slinko (2013). Achieving fully proportional representation is easy in practice. In: *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 399–406.

[73] E. Stoimenova (2000). Rank tests based on exceeding observations. *Annals of the Institute of Statistical Mathematics*. Vol. 52 (2), 255–266.

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

[75] M. Tiku, W. Tan, and N. Balakrishnan (1986). Robust Inference. *Marcel Dekker, NY*.

[76] J. Tukey, D. Brillinger, and L. Jones (1978). *The Management of Weather Resources II: The Role of Statistics in Weather Resources Management*. US Government Printing Office, Washington.

[77] J. S. Verducci (1982). *Discriminating Between Two Probabilities on the Basis of Ranked Preferences*. Tech. rep.

[78] J. S. Verducci (1989). Minimum majorization decomposition. In: *Contributions to Probability and Statistics*. Springer, pp. 160–173.

[79] M. Vock and N. Balakrishnan (2011). A Jonckheere–Terpstra-type test for perfect ranking in balanced ranked set sampling. *Journal of Statistical Planning and Inference*. Vol. 141 (2), 624–630.

[80] D. A. Wolfe (2012). Ranked set sampling: its relevance and impact on statistical inference. *ISRN Probability and Statistics*. Vol. 2012. `https://doi.org/10.5402/2012/568385`, 1–32.

[81] P. L. H. Yu and H. Xu (2019). Rank aggregation using latent-scale distance-based models. *Statistics and Computing*. Vol. 29 (2), 335–349.

[82] E. Zamanzade, N. R. Arghami, and M. Vock (2012). Permutation-based tests of perfect ranking. *Statistics & Probability Letters*. Vol. 82 (12), 2213–2220.

[83] E. Zamanzade and M. Vock (2018). Some nonparametric tests of perfect judgment ranking for judgment post stratification. *Statistical Papers*. Vol. 59 (3), 1085–1100.