Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

METRIC METHODS FOR ANALYZING AND MODELING RANK DATA

Nikolay Ivanchev Nikolov

SUMMARY

OF 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

The thesis is written in English and contains 75 pages, from which 70 pages body text and 5 pages Bibliography, that refers to 83 scientific sources.

The numeration of the theorems, propositions, corollaries and equations in the summary follows their numeration in the thesis.

The research was conducted in the framework of doctoral program "Probability Theory and Mathematical Statistics" at the department "Operations Research, Probability and Statistics" in Institute of Mathematics and Informatics of Bulgarian Academy of Sciences.

Author: Nikolay Ivanchev Nikolov
Title: Metric methods for analyzing and modeling rank data

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. 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 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 [23], Bradley and Terry [2], Luce [16], Plackett [22] and Mallows [17]. 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. [3], Klementiev et al. [13] and Murphy and Martin [19]. 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 [8] and Marden [18]. An example of a more exotic distance is the

Lee distance, which has been developed by Lee [14] as a generalization of the Hamming distance. 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. 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. [1] and Section 3.1.2 in Chen et al. [4]. Nonparametric tests for null hypothesis of perfect rankings against a general alternative of imperfect ranking have been developed by Frey et al. [9], Li and Balakrishnan [15], Vock and Balakrishnan [24] and Zamanzade et al. [25]. 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 [10] considered four models for imperfect ranking: Bivariate normal model proposed by Dell and Clutter [6], Fraction of random rankings by Frey et al. [9], Fraction of inverse rankings by Frey et al. [9] and Fraction of neighbor rankings by Vock and Balakrishnan [24]. 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. There are various of techniques for constructing nonparametric rank tests for hypothesis testing of two samples, see Hájek and Šidák [12] and Good [11]. Critchlow [5] 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 [21] 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.

The thesis is structured in five chapters.

Chapter 1. Preliminaries

In this chapter, distances on permutations are defined and several examples are considered. Some important statistical properties of Lee distance, which are used in the next chapters, are derived.

The structural nature of rankings suggests more special approach in rank data analysis. Distances on permutations are often convenient tools for modeling rank data. They measure the closeness between two rankings and can be very useful and informative for revealing the main features of the data.

In **Section 1.1**, we give a definition of a distance on $\mathbf{S}_N$ and use it to obtain a measure of cental location in a sample of n complete rankings. Since the empirical central ranking depends on the choice of the distance on $\mathbf{S}_N$, we present eight of the most widely used distances in applied scientific and statistical problems. In order to illustrate some differences between the listed distances, we consider the canonical example of arranging books on a shelf into alphabetic order. The spatial characteristics of Lee distance and Spearman's footrule are given in more details. Additionally, we describe two rich classes of distances on $\mathbf{S}_N$: p-distances and Hoeffding's distances, and give definitions of the left-invariance, right-invariance and metric properties.

In **Section 1.2**, we study the properties of the random variable $D_L(\pi)$ induced by Lee distance under uniformity of $\pi \in \mathbf{S}_N$. By using a representation of $D_L(\pi)$ with a simple cycle graph, we give an interpretation of the quantities $c_N(i,j) := \min(\mid i-j\mid, N- \mid i-j\mid)$ in the linear decomposition

$$D_L(\pi) = d_L(\pi, e_N) = \sum_{i=1}^{N} \min(\mid \pi(i) - i\mid, N- \mid \pi(i) - i\mid) = \sum_{i=1}^{N} c_N(\pi(i), i)\,,$$

where $e_N = \langle 1, 2, \ldots, N\rangle \in \mathbf{S}_N$ is the identity permutation. We also show that when N is even the "opposite" ranking of e_N for Lee distance is

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

and when N is odd there are two the "opposite" rankings

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

By making use of the right-invariant property of Lee distance, we prove that when N is even the distribution of D_L is symmetric and D_L can take only even values.

In Subsections 1.2.1 and 1.2.2, we apply Hoeffding's combinatorial central limit theorem (CCLT) to the random variable D_L and derive the mean and variance of D_L:

$$\mathbf{E}\left(D_L\right) = \left[\frac{N+1}{2}\right]\left[\frac{N}{2}\right], \tag{1.11}$$

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

where $[x]$ is the greatest integer less than or equal to x. Furthermore, from the CCLT we prove the following theorem.

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)*.*

The presented properties of D_L play an important role for the research in the next chapters.

Chapter 2. Probability models for rank data

This chapter is devoted to probabilistic models for rank data, which are determined by a family of probability distributions $\mathbf{P}$. In **Sections 2.1** and **2.2**, we study the Distance-based models and their relation to the Marginals models.

Distance-based models are part of the exponential family and are defined by

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

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 *modal* (or *antimodal*) ranking and $\psi_N(\theta)$ is a normalizing constant. Finding the value of $\psi_N(\theta)$ for fixed θ and chosen distance $d(\cdot,\cdot)$ is a difficult task, since evaluating $\psi_N(\theta)$ by summing over all possible $N!$ rankings becomes computationally demanding for $N \geq 10$. Thus, for the models based on Lee distance we propose the following approximation of $\psi_N(\theta)$

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

where $\mu = \mathbf{E}\left(D_L\right)$ and $\nu^2 = \mathbf{Var}\left(D_L\right)$ are given in (1.11) and (1.13), respectively. In **Section 2.1**, we describe more applications of the approximation

(2.4) in other problems for rank data and compare $\psi_N(\theta)$ and $\hat{\psi}_N(\theta)$ for various values of θ and N. In order to extend model (2.2), we consider the Latent-class Distance-based model, which assumes that there are K latent groups (classes) in the population and that the distributions of the rankings within each group are modeled by (2.2).

In **Section 2.2**, we present the Marginals model by the probability distribution

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

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. We point out that (2.2) based on Lee distance corresponds to (2.6) with

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

Since finding the values of the Marginals matrix for large values of N requires a lot of time and computational resources, we propose the following asymptotic approximation for the model (2.2) based on Lee distance.

Theorem 2.2. *Let $M(\theta,N) = \left\{m_{ij}(\theta,N)\right\}_{i,j=1}^{N}$ be the Marginals matrix, based on Lee distance. Then*

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

where

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

and

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

In **Section 2.3**, we give a description of some statistical tools for estimating the unknown parameters and testing the goodness-of-fit of models (2.2) and (2.6). However, for the Latent-class Distance-based model it is not possible to estimate the unknown parameters directly. Therefore, we make use of the Expectation-Maximization (EM) algorithm suggested by Dempster et al. [7] and propose an estimation procedure in the case when there are K groups in the population with unknown modal rankings. The aim of the algorithm is to find maximize 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 $\vec{\pi}_0$, 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, $\pi^*=\left(\pi^1,\pi^2,\ldots,\pi^n\right)$ is a sample of n complete rankings and G denotes the group index, i.e. $G\in\{G_1,G_2,\ldots,G_K\}$. By applying the generalized version of the EM algorithm for

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

with some initial values $\vec{\theta}^{(t)},\vec{p}^{(t)}$ and $\vec{\pi}_0{}^{(t)}$, we show that the sufficient condition for convergence to local maximum

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

holds for

$$p_j^{(t+1)}=\frac{1}{n}\sum_{i=1}^{n}\frac{p_j^{(t)}P_{\theta_j^{(t)},\pi_{0,j}^{(t)}}\left(\pi^i\right)}{\sum\limits_{s=1}^{K}p_s^{(t)}P_{\theta_s^{(t)},\pi_{0,s}^{(t)}}\left(\pi^i\right)}, \qquad (2.10)$$

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

and $\left\{\theta_j^{(t+1)}\right\}_{j=1}^{K}$ which are the solutions of

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

This result is formulated and proved in **Section 2.4** as:

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.*

The proposed algorithm is used to perform a simulation study of the explanatory abilities of the Latent-class models for several values of K.

We compare the models (2.2) based on Lee distance, Hamming distance and Kendall's tau to the Marginals model (2.6) by providing three illustrative examples in **Section 2.5**. It is shown that the contrast between the model (2.2) based on Lee distance and the model (2.6) can be explored only through analyzing the sample Marginals matrix and the Marginals matrix induced by Lee distance. We concluded that choosing 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 are useful to detect if there are more than one groups or clusters in the data.

Chapter 3. Rank data clustering

In this chapter, we present an unsupervised "K-means" clustering procedure based on Lee distance.

For n observations of complete rankings $\pi^1, \pi^2, \ldots, \pi^n \in \mathbf{S}_N$ and fixed number of groups K, we determine the cluster centers $\hat{\sigma}^{(1)}, \hat{\sigma}^{(2)}, \ldots, \hat{\sigma}^{(K)} \in \mathbf{S}_N$ as the solution of

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

where

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

for some distance $d(\cdot, \cdot)$ on $\mathbf{S}_N$. In order to compare the results obtained from several clustering analysis based on different values of K, we consider the measure of "tightness" T_K defined by

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

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

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

is the value of C_K under uniform distribution over all possible $N!$ rankings. Since there are $\binom{N!}{K}$ possible choices for cluster centers and the complete search in (3.4) becomes computationally demanding for $N \ge 10$, in **Section 3.2** we propose the following asymptotic approximation which is based on **Theorem 1.2**.

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*

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

We show that this result is useful for approximating the values of C_K^0 for $K = 2$, but it is not easy to be generalized for $K > 2$.

In **Section 3.3**, the presented clustering method is applied to the well-studied American Psychological Association (APA) election dataset. The obtained results illustrate that Lee distance performs well in situations where it is desirable to construct models with less number of groups K and respectively with fewer unknown parameters.

Chapter 4. Imperfect ranking in ranked set sampling

In this chapter, we consider some statistical measures of deviation from the perfect ranking in the framework of ranked set sampling (RSS).

The procedure for obtaining n-cycle balanced RSS of size k and the concept of perfect ranking are described in **Section 4.1**. In **Section 4.2**, we preset the nonparametric approach for testing the null hypothesis for perfect ranking proposed by Li and Balakrishnan [15]. By applying their method for constructing test statistic for one-cycle sample through measuring the distance between the ordered RSS $\langle i_1, i_2, \dots, i_k \rangle$ and the identity $e_k = \langle 1, 2, \dots, k \rangle$, we define two new test statistics

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

based on Chebyshev's distance and Lee's distance, respectively. By combining the statistics in each cycle, we define two test statistics for n-cycle sample

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

where $M_k^{(i)}$ and $L_k^{(i)}$ are the values the test statistics in (4.4) for the i-th cycle of RSS, for $i = 1, 2, \ldots, n$.

In order to compare the test statistics described in **Section 4.2**, it is necessary to fix an alternative model for the imperfect judgment ranking. In **Section 4.3**, we propose the Distance-based models (2.2) as an imperfect ranking alternative and study the their properties in the framework of RSS. We show that the described alternative model is completely specified by the ranking error probability matrix $\mathbf{Q}(k,\theta)$, which is referred as the Marginals matrix in **Chapter 2**.

One of the benefits of considering the Mallows' model as an alternative of the perfect ranking is that the unknown parameter of the model can be estimated. 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, $R_{i[j]}$ is the number of the set in the i-th cycle from which comes the j-th ordered statistic for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$, and $R_i = \left\langle R_{i[1]}, R_{i[2]}, \ldots, R_{i[k]} \right\rangle$ for $i = 1, 2, \ldots, n$. By using Mallows' model for imperfect ranking associated with error matrix $\mathbf{Q}(k,\theta)$ and free parameter $\theta \leq 0$, we expressed the likelihood function 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 ordered RSS, $q\left(R_{i[j]}, l_{[j]}, k, \theta\right)$ are elements of $\mathbf{Q}(k,\theta)$ and $p\left(l_{[1]}, l_{[2]}, \ldots, l_{[k]}\right)$ are the probabilities of observing $\langle l_{[1]}, l_{[2]}, \ldots, l_{[k]} \rangle$ under the assumption of perfect ranking. Since there is no closed expression for the elements of the matrix $\mathbf{Q}(k,\theta)$ and it is not possible to estimate θ directly, we make use of the EM algorithm to find the maximum likelihood estimate (MLE) of θ. In **Section 4.4**, we show that the value $\theta^{(t+1)}$ in each iteration, which maximizes the expected value of the complete loglikelihood function for the value $\theta^{(t)}$ of the previous iteration, is the solution of the equation

$$\sum_{z} \frac{\prod_{i=1}^{n}\left[\left(\prod_{j=1}^{k} q\left(R_{i[j]}, z_{i[j]}, k, \theta^{(t)}\right)\right) p\left(z_{i[1]}, z_{i[2]}, \dots, z_{i[k]}\right)\right]}{\sum_{l}\left\{\prod_{i=1}^{n}\left[\left(\prod_{j=1}^{k} q\left(R_{i[j]}, l_{i[j]}, k, \theta^{(t)}\right)\right) p\left(l_{i[1]}, l_{i[2]}, \dots, l_{i[k]}\right)\right]\right\}} \times$$

$$\times \sum_{i=1}^{n}\sum_{j=1}^{k} \frac{q'\left(R_{i[j]}, z_{i[j]}, k, \theta\right)}{q\left(R_{i[j]}, z_{i[j]}, k, \theta\right)} = 0, \qquad (4.9)$$

where $q'(r,z,k,\theta)$ is the derivative of $q(r,z,k,\theta)$ with respect to θ and is expressed 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 $r,z = 1,2,\dots,k$. We prove that the proposed EM algorithm converges monotonically to some stationary point of $L(R \mid \theta)$ and if $L(R \mid \theta)$ is unimodal, i.e. $L(R \mid \theta)$ has only one stationary point, then the algorithm converges to the unique MLE of θ.

Since in each step of the presented EM algorithm it is required to calculate the error matrix $\mathbf{Q}(k,\theta)$, the application of this method takes a lot of computational resources for large values of k. In **Section 4.5**, we give an alternative way to find the matrix $\mathbf{Q}(k,\theta)$ for Cayley and Hammning distances that decreases the computational operations in (4.9). Similar to the asymptotic result in **Theorem 2.1** for the probability error matrix based on Lee distance, we provide two approximations for the error matrices based on Spearman's footrule and Spearman's rho.

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

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

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

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

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

In **Section 4.6**, we apply the Mallows' model as an imperfect ranking alternative and compare the described test statistics for one-cycle RSS. In **Section 4.7**, we illustrate the use of Mallows' alternative for n-cycle RSS by analyzing an example dataset given in Murrary et al. [20].

Chapter 5. Lee distance in two-sample rank tests

In this chapter, we study a rank test statistic induced by Lee distance and Critchlow's [5] unified approach for the two-sample location problem. For two independent random samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ with continuous distribution functions $F(x)$ and $G(x)$ respectively, we consider the problem of testing the null hypothesis H_0 against 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. If $\alpha(i)$ is the rank of X_i for $i = 1,2,\ldots,m$ and $\alpha(m+j)$ is 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 $\alpha \in \mathbf{S}_{m+n}$.

In **Section 5.1**, we present the Critchlow's [5] method for construction of test statistics that is based on finding the minimum interpoint distance between the class of equivalence $[\alpha] = \alpha(\mathbf{S}_m \times \mathbf{S}_n) = \{\alpha \circ \pi : \pi \in \mathbf{S}_m \times \mathbf{S}_n\}$ and the extremal set $E = \mathbf{S}_m \times \mathbf{S}_n = \{\pi \in \mathbf{S}_{m+n} : \pi(i) \leq m, \forall i \leq m\}$

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

where d is an arbitrary distance on $\mathbf{S}_{m+n}$.

In **Section 5.2**, we show that the rank test statistic in (5.3) induced by the Lee distance is the solution of the problem

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

where $e = \langle 1, 2, \ldots, m+n \rangle$ is the identity permutation. We express $d_L([\alpha], E)$ as

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

where

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

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

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

we use $L_{m,n}$ for testing H_0 against the alternative H_1.

In **Section 5.3**, we prove the following proposition by using the fact that $L_{m,n}$ is the minimum sum of distances over C between the elements of K_m and the elements of K_n, where C is a simple cycle graph with vertices $\{i\}_{i=1}^{m+n}$ and edges $\bigcup_{i=1}^{m+n-1}\{i, i+1\}$ and $\{m+n, 1\}$.

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

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

where $s \in \{1, 2, \ldots, k+1\}$ in conditions **(i)-(iii)***, $s \in \{1, 2, \ldots, k\}$ in condition* **(iv)** *and $s \in \{2, 3, \ldots, k+1\}$ in condition* **(v)***. 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)**.

Since for large values of m and n the computational process of checking conditions **(i)-(v)** for all possible 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 $\{b_j^{(n)}\}_{j=1}^{k-s+1}$ is time-consuming and computationally demanding, we give recursive relations for the number of terms in the summations (5.8) (see **Proposition 5.2**) and the probability mass function of $L_{m,n}$ under H_0 (see **Proposition 5.3**).

The mean and variance of $L_{m,n}$ and its asymptotic distribution under H_0 are derived in the following theorem.

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*

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

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

where $m+n$ is odd in the first case and $m+n$ is even in the second. 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 $m,n \to \infty$.

This result gives a normal approximation of the distribution of $L_{m,n}$ under H_0 and is useful for finding the critical region when m and n are large.

In **Section 5.4**, we compare the rank test statistic $L_{m,n}$, defined by (5.7), to t-test, Wilcoxon, Kolmogorov-Smirnov and Mood's tests statistics for two-samples. We perform an illustrative simulation study and show that the test based on Lee distance is more powerful than the others when the generating distributions have heavy-tails. We 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.

The proofs of Theorems 2.1, 4.1, 4.2 and 5.1 are given in **Appendix E**.

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. 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.

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. 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. 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.

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. 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. 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. 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. The Lee test statistic is shown to be more powerful for heavy-tailed underlying distributions via a simulation study.

Publications related to the thesis

1. 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.

2. 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.

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

4. 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.

5. 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.

6. 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).

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).

Acknowledgements

I wish to express my sincerely gratitude to my advisor 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 post-doctoral students” approved by DCM #577/17.08.2018.

Bibliography

[1] 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.

[2] 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.

[3] 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.

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

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

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

[7] 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.

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

[9] 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.

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

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

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

[13] 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.

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

[15] 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.

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

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

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

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

[20] 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.

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

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

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

[24] 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.

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