Българска академия на науките
Институт по математика и информатика

МЕТРИЧНИ МЕТОДИ ЗА АНАЛИЗИРАНЕ И МОДЕЛИРАНЕ НА НАРЕДЕНИ ДАННИ

Николай Иванчев Николов

АВТОРЕФЕРАТ

на дисертация

за присъждане на образователна и научна степен

Доктор

в професионално направление 4.5 Математика
(Теория на вероятностите и математическа статистика)

Научен ръководител:
проф. дмн Евгения Стоименова

София, 2020

Дисертационният труд е написан на английски език и съдържа 75 страници, от които 70 страници основен текст и 5 страници библиография с 83 заглавия.

Номерацията на теоремите, твърденията, следствията и тъждествата в автореферата съответства точно на номерацията им в дисертацията.

Изследванията са проведени в рамките на докторска програма "Теория на вероятностите и математическа статистика" в секция "Изследване на операциите, вероятности и статистика" на Института по математика и информатика към Българска академия на науките.

Автор: Николай Иванчев Николов

Заглавие: Метрични методи за анализиране и моделиране на наредени данни

Въведение

Наредените данни често възникват в ситуации, при които е необходимо да се подреди множество от индивиди или обекти по отношение на някакъв критерий. Такива данни могат да се наблюдават директно или могат да произлизат от подредба на множество или подмножество от елементи със зададени числови стойности. Най-общо съществуват два типа наредби: пълна и частична, в зависимост от това дали е необходимо да се подредят всички обекти или не. В този дисертационен труд ще се ограничим в случая, при който са подредени всички обекти, т.е. когато се наблюдават пълни наредби.

Пълна наредба на N обекта съответства на цялостно подреждане на обектите. Всеки такъв вектор на наредба може да се разглежда като елемент π от пермутационната група $\mathbf{S}_N$, породена от първите N естествени числа. Една пермутация $\pi \in \mathbf{S}_N$ представлява функция от $\{1, \ldots, N\}$ в себе си, чиито аргументи са обектите и чиито стойности са ранговете в тяхната подредба. Ако обектите са означени с числата $\{1, \ldots, N\}$, тогава $\pi(i)$ е рангът на обекта i, а $\pi^{-1}(i)$ е обектът с ранг i в подредбата. Така

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

е пермутацията в $\mathbf{S}_N$, която съответства на подредбата на обектите според техните рангове. За извадка от n пълни наредби ще използваме означението $\pi^1, \pi^2, \ldots, \pi^n \in \mathbf{S}_N$.

Наредените данни имат естествена структура, която за разлика от типичните многомерни извадки предоставя нови предизвикателства и възможности. Съществува богата литература върху анализа на наредени данни, която включва класическите вероятностни модели на Thurstone [23], Bradley и Terry [2], Luce [16], Plackett [22] и Mallows [17]. Моделите на Mallows често са удобен първоначален инструмент за анализиране на наредени данни. Те улавят главната структура в данните само с един параметър и могат да са основата на бъдещо изследване. Въпреки това обикновено е нереалистично да очакваме еднопараметричен модел да разкрие всички характеристики на данните. Едно възможно обобщение на тези модели може да се направи ако се предположи, че има няколко латентни групи в генералната съвкупност. Проблемът за намиране на "водещи" наредби в различните клъстери (групи) е изследван от много автори, виж Busse и др. [3], Klementiev и др. [13] и Murphy и Martin [19].

По-голямата част от тези методи могат да се опишат чрез разстояния върху пермутации, които са мощно средство за откриването на скрити структури в наредените данни. Числените характеристики, точните разпределения, асимптотичните приближения и статистическите приложения на случайните величини получени от най-често използваните разстояния върху $\mathbf{S}_N$ могат да бъдат намерени в Diaconis [8] и Marden [18]. Пример за по-екзотично разстояние е разстоянието на Lee, което е предложено от Lee [14] като обобщение на разстоянието на Hamming. Статистическите свойства на разстоянието на Lee обаче не са добре изучени. В настоящия дисертационен труд се извеждат някои асимптотични приближения на случайната величина базирана на разстоянието на Lee и се прилагат към няколко вероятностни модела за наредени данни и други статистически проблеми включващи наредби.

Съществуват разнообразни приложения включващи наредби в много приложни научни области. Като пример може да се посочи анализа на несъвършени наредби при извадка от наредени множества (ИНМ). ИНМ може да се използва за създаването на по-ефективни методи в широк клас от статистически проблеми. Ползата от конструирането на ИНМ е най-значима при извършването на съвършена наредба, което обаче не винаги е осъществимо. Поради тази причина е желателно да се разгледат статистически модели, които улавят несигурността в процеса на подредба и проверяват дали наредбата е съвършена или не. Тези модели могат да се дефинират чрез матрицата от вероятностните за грешка при наредба, която може да се използва за изучаването на ефекта от несъвършена наредба върху свойствата на статистическите процедури основани на ИМН, виж Aragon и др. [1] и §3.1.2 в Chen и др. [4]. Непараметрични критерии за проверка на нулевата хипотеза за съвършена наредба срещу общата алтернатива за несъвършена наредба са разработени от Frey и др. [9], Li и Balakrishnan [15], Vock и Balakrishnan [24] и Zamanzade и др. [25]. В случая, когато хипотезата за съвършена наредба е отхвърлена, е необходимо по-детайлно да се анализира процеса на наредба в множествата. Frey и Wang [10] разглеждат четири модела на несъвършена наредна: Двумерен нормален модел предложен от Dell и Clutter [6], Модел на пропорция от случайни наредби от Frey и др. [9], Модел на пропорция от обратни наредби от Frey и др. [9] и Модел на пропорция от наредби в съседство от Vock и Balakrishnan [24]. Като допълнение, тези модели могат да се

използват за сравняване на способностите за наредба на два метода, чрез което да се увеличи ефективността на процедурата при ИНМ за бъдещи наблюдения.

Наредбите и разстоянията върху пермутации намират и приложение в един от най-важните статистически проблеми: сравняването на две извадки. Ако предположим, че разпределенията на двете генерални съвкупности се различават само по местоположение, то разполагаме с разнообразни параметрични и непараметрични критерии. При непараметричния подход са необходими по-малко предположения относно разпределението, от което произлизат данните, и позволяват да се избере най-подходящия статистически критерий за поставената задача. Съществуват многообразни техники за конструиране на непараметрични рангови критерии при две извадки, виж Hájek и Šidák [12] и Good [11]. Critchlow [5] предлага общ подход основан на минимално разстояние между две пермутационни множества съответстващи на нулевата и алтернативната хипотеза. Използвайки различни разстояния върху пермутации, получаваме различни статистически критерии. Някои от най-популярните рангови статистики: на Kolmogorov-Smirnov, Wilcoxon и Mann-Whitney, се извеждат съответно чрез разстоянието на Ulam, правилото на Spearman и τ на Kendall. Едно от предимствата при наличието на няколко статистически критерия е, че те могат да се комбинират с цел подобряване на тяхната мощност. Pesarin [21] разработва интересна теория, Непараметрична комбинация на зависими критерии, която дава добри резултати за редица сложни многомерни проблеми, включително такива, които още не са решени параметрично.

Основната цел на настоящата дисертация е да изучи статистическите свойства на разстоянието на Lee и да изследва приложенията му в няколко модела за наредени данни основани на разстояния. Основните задачи могат да се формулират по следния начин:

- Получаване на асимптотичен резултат за разпределението на случайната величина получена от разстоянието на Lee при равномерно разпределени наредби.
- Сравнение на модела на Mallows получен от разстоянието на Lee с други вероятностни модели за наредени данни.
- Създаване на EM алгоритъм за оценка на неизвестните параметри в моделите базирани на разстояние за наредени данни с няколко латентни групи.

- Приближаване на мярката за "сгъстеност" при клъстеризация с "K-средни" основана на разстояние на Lee.
- Намиране на асимптотично приближение на матрицата от вероятностните за грешка при наредба получена от моделите основани на разстоянието на Lee, правилото на Spearman и ρ на Spearman за извадки от наредени множества.
- Конструиране на процедура за оценка на неизвестните параметри в модела на Mallows за несъвършени наредби като се използва EM метода.
- Извеждане на ранговата статистика получена от разстоянието на Lee и метода на Critchlow за местоположението на две извадки и изучаване на нейното разпределение при вярна нулева хипотеза.

Дисертационният труд е структуриран в пет глави.

Глава 1. Уводни бележки

В тази глава се дефинират разстояния върху пермутации и се разглеждат няколко примера. Изведени са някои важни статистически свойства на разстоянието на Lee, които се използват в следващите глави.

Естествената структура на наредбите изисква по-специален подход при анализа на наредени данни. Разстоянията върху пермутации често са удобно средство за моделиране на наредени данни. Те измерват близостта между две наредби и могат да бъдат много полезни за откриване на най-важните особености в данните.

В **§ 1.1** даваме дефиниция за разстояние върху $\mathbf{S}_N$ и я използваме за да получим мярка за централна наредба в извадка от n пълни наредби. Тъй като емпиричната централна наредба зависи от избора на разстояние върху $\mathbf{S}_N$, представяме осем от най-широко използваните разстояния в научноприложни и статистически задачи. За да илюстрираме някои разлики между изброените разстояния, разглеждаме класическия пример за подреждане на книги в азбучен ред. Подробно са разгледани пространствените характеристики на разстоянието на Lee и правилото на Spearman. Освен това, са описани два богати класа от разстояния върху $\mathbf{S}_N$: p-разстоянията и разстоянията на Hoeffding. Дадени са дефинициите на метрика върху $\mathbf{S}_N$ и свойствата лява-инвариантност и дясна-инвариантност.

В **§ 1.2** изучаваме свойствата на случайната величина $D_L(\pi)$ получена от разстоянието на Lee при равномерно разпределение на $\pi \in \mathbf{S}_N$. Използвайки представянето на $D_L(\pi)$ чрез прост цикличен граф, даваме интерпретация на членовете

$$c_N(i,j) := \min\left(\mid i-j \mid, N- \mid i-j \mid\right)$$

в линейното разлагане

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

където $e_N = \langle 1, 2, \ldots, N\rangle \in \mathbf{S}_N$ е пермутацията идентитет. Допълнително показваме, че "противоположната" наредба на e_N за разстоянието на Lee и четни стойности на N е

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

а за нечетни стойности на N съществуват две "противоположни" наредби

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

Използвайки дясно-инвариантното свойство на разстоянието на Lee, доказваме, че разпределението на D_L е симетрично и D_L приема само четни стойности, когато N е четно.

В **§ 1.2.1** и **§ 1.2.2** прилагаме комбинаторната централна гранична теорема (КЦГТ) на Hoeffding към случайната величина D_L и извеждаме нейните очакване и дисперсия:

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

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

където $[x]$ е най-голямото цяло число по-малко или равно на x. Освен това, чрез КЦГТ доказваме следната теорема.

Теорема 1.2. *Разпределението на случайната величина D_L е асимптотично нормално при $N \to \infty$ с очакване и дисперсия дадени съответно чрез* (1.11) *и* (1.13).

Представените свойства на D_L играят важна роля в изследванията в следващите глави.

Глава 2. Вероятностни модели за наредени данни

Тази глава е посветена на вероятностни модели за наредени данни, които се определят от фамилия от вероятностни разпределения $\mathbf{P}$ върху $\mathbf{S}_N$. В **§ 2.1** and **§ 2.2** изследваме моделите базирани на разстояние и тяхната връзка с маргиналните модели.

Моделите базирани на разстояние са част от експоненциалната фамилия и се дефинират чрез

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

където θ е реален параметър ($\theta \in \mathbf{R}$), $d(\cdot,\cdot)$ е разстояние върху $\mathbf{S}_N$, π_0 е фиксирана *модална* (или *антимодална*) наредба и $\psi_N(\theta)$ е нормираща константа. Намирането на стойността на $\psi_N(\theta)$ за дадено θ и избрано разстояние $d(\cdot,\cdot)$ е трудна задача, тъй като калкулирането на $\psi_N(\theta)$ чрез сумиране по всички възможни $N!$ наредби изисква голяма изчислителна мощ за $N \geq 10$. Поради тази причина, предлагаме следното приближение на $\psi_N(\theta)$ за моделите базирани на разстоянието на Lee

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

където $\mu = \mathbf{E}\,(D_L)$ и $\nu^2 = \mathbf{Var}\,(D_L)$ са дадени съответно в (1.11) и (1.13). В **§ 2.1** описваме още приложения на приближението (2.4) в други задачи за наредени данни и сравняваме $\psi_N(\theta)$ и $\hat{\psi}_N(\theta)$ за различни стойности на θ и N. Обобщаваме модел (2.2) като разглеждаме модела с латентни класове и базиран на разстояние, при който се предполага, че в генералната съвкупност има K латентни групи (класове), а разпределенията на наредбите във всяка група се моделират чрез (2.2).

В **§ 2.2** дефинираме маргиналния модел чрез вероятностното разпределение

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

където $\vec{\lambda} = \left\{\lambda_i^{(j)}\right\}_{i,j=1}^{N}$ са N^2 реални параметъра, $\mathbf{I}[\cdot]$ е индикаторната функция и $\psi(\vec{\lambda})$ е нормираща константа. Отбелязваме, че (2.2)базиран на разстоянието на Lee съвпада с (2.6) за

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

Тъй като намирането на стойностите на маргиналната матрица за голямо N изисква много време и изчислителни ресурси, предлагаме следното нейно асимптотично приближение за модела базиран на разстоянието на Lee.

Теорема 2.2. *Нека* $M(\theta,N) = \{m_{ij}(\theta,N)\}_{i,j=1}^{N}$ *е маргиналната матрица базирана на разстоянието на Lee. Тогава*

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

където

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

и

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

В **§ 2.3** описваме някои статистически инструменти за оценка на неизвестните параметри и проверката за съгласуваност на моделите (2.2) и (2.6). За модела с латентни класове и базиран на разстояние обаче е невъзможно неизвестните параметри да се оценят директно. Поради тази причина, използвайки EM алгоритъма даден в Dempster et al. [7] и предлагаме процедура за оценка в случая,

когато не са известни модалните наредби за K-те групи в генералната съвкупност. Целта на алгоритъма е да намери максимума на очакваната стойност на логаритмичната функция на правдоподобие $\ell(\vec{\theta}, \vec{p}, \vec{\pi_0}, \vec{\pi^*}, G)$ при зададени начални стойности на $\vec{\theta}$, $\vec{p}$ и $\vec{\pi_0}$, където $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$, $\vec{\pi_0} = (\pi_{0,1}, \pi_{0,2}, \ldots, \pi_{0,K})$ и $\vec{p} = (p_1, p_2, \ldots, p_K)$ са вектори от неизвестни параметри, $\pi^* = \left(\pi^1, \pi^2, \ldots, \pi^n\right)$ е извадка от n пълни наредби, а с G е означен индекса на групата, т.е. $G \in \{G_1, G_2, \ldots, G_K\}$. Прилагайки обобщения вариант на EM алгоритъма за

$$Q^{(t)}\left(\vec{\theta}, \vec{p}, \vec{\pi_0}, \vec{\pi^*}\right) = \mathbf{E}_{G|\vec{\theta}^{(t)}, \vec{p}^{(t)}, \vec{\pi_0}^{(t)}, \vec{\pi^*}}\left[\ell(\vec{\theta}, \vec{p}, \vec{\pi_0}, \vec{\pi^*}, G)\right]$$

с начални стойности $\vec{\theta}^{(t)}, \vec{p}^{(t)}$ и $\vec{\pi_0}^{(t)}$, показваме, че достатъчното условие за сходимост към локален максимум

$$Q^{(t)}\left(\vec{\theta}^{(t+1)}, \vec{p}^{(t+1)}, \vec{\pi_0}^{(t+1)}, \vec{\pi^*}\right) \geq Q^{(t)}\left(\vec{\theta}^{(t)}, \vec{p}^{(t)}, \vec{\pi_0}^{(t)}, \vec{\pi^*}\right) \tag{2.12}$$

е изпълнено за

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

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

и $\left\{\theta_j^{(t+1)}\right\}_{j=1}^{K}$, които са решения на

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

Този резултат е формулиран и доказан в **§ 2.4** като:

Твърдение 2.1. *Нека* $\vec{p}^{(t+1)}$, $\vec{\pi_0}^{(t+1)}$ *и* $\vec{\theta}^{(t+1)}$ *са дадени съответно чрез (2.10), (2.13) и (2.14). Тогава условието (2.12) е изпълнено.*

Предложеният алгоритъм е използван за извършването на симулационно изследване на обяснителните способности на модела с латентни класове за различни стойности на K.

В **§ 2.5** сравняваме моделите (2.2) базирани на разстоянието на Lee, разстоянието на Hamming и τ на Kendall с маргиналния модел (2.6) като разглеждаме три илюстративни примера. Показано е, че разликата между модела (2.2) базирани на разстоянието на Lee и модела (2.6) може да се изследва само чрез анализиране на извадъчната маргинална матрица и маргиналната матрица получена от разстоянието на Lee. Правим заключението, че изборът на разстоянието на Lee в модел (2.2) е подходящ, когато има няколко групи в наблюдаваните наредби и модалната подредба в една група не е обратната подредба в друга. Освен това, модели базирани на разстоянието на Lee са полезни за проверка дали има няколко групи или клъстери в данните.

Глава 3. Клъстеризация на наредени данни

В тази глава представяме клъстеризация с "K-средни" без учител и базирана на разстоянието на Lee.

За n наблюдения от пълни наредби $\pi^1, \pi^2, \ldots, \pi^n \in \mathbf{S}_N$ и фиксиран брой групи K, определяме центровете на клъстерите

$$\hat{\sigma}^{(1)}, \hat{\sigma}^{(2)}, \ldots, \hat{\sigma}^{(K)} \in \mathbf{S}_N$$

като решение на

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

където

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

за някое разстояние $d(\cdot, \cdot)$ върху $\mathbf{S}_N$. За да сравним резултатите получени от няколко клъстерни анализа с различни стойности на

K, разглеждаме мярката за "сгъстеност" T_K дефинирана чрез

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

където $K = 1, 2, \ldots$ и

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

е стойността на C_K при равномерно разпределение над всички възможни $N!$ наредби. Понеже има $\binom{N!}{K}$ избора за центрове на клъстерите и проверката на всички възможности в (3.4) изисква мощни изчислителни средства за $N \ge 10$, в **§ 3.2** извеждаме следното асимптотично приближение получено от **Теорема 1.2**.

Следствие 3.1. *Нека константата C_2^0 е дефинирана в* (3.4) *за $K = 2$ и чрез разстоянието на Lee $d_L(\cdot, \cdot)$. Тогава за големи четни стойности на N, константата C_2^0 може да се приближи чрез*

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

Показваме, че този резултат е полезен за приближаване на стойностите на C_K^0 за $K = 2$ дори и за нечетно N, както и че не е лесно да се обобщи за $K > 2$.

В **§ 3.3** представената клъстерна процедура е приложена към добре изследваните изборни данни на Американската психологична асоциация (АСА). Получените резултати илюстрират, че разстоянието на Lee се справя добре, когато се изисква конструирания модел да е с малък брой групи K и съответно с по-малко неизвестни параметри.

Глава 4. Несъвършенни наредби при извадки от наредени множества

В тази глава разглеждаме някои статистически мерки за отклонение от съвършената наредба при извадка от наредени множества (ИНМ).

Процедурата за получаване на балансирана ИНМ от n цикъла с размер k и понятието съвършена наредба са описани в **§ 4.1**. В

§ 4.2 представяме непараметричния подход за проверка на хипотезата за съвършена наредба предложен от Li и Balakrishnan [15]. Прилагайки техния метод за конструиране на критерий за извадка от един цикъл чрез измерване на разстоянието между подредбата в ИНМ $\langle i_1, i_2, \ldots, i_k \rangle$ и идентитета $e_k = \langle 1, 2, \ldots, k \rangle$, дефинираме две нови статистики

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

получени съответно от разстоянието на Chebyshev и разстоянието на Lee. Комбинирайки статистиките от всеки цикъл, дефинираме следните статистики за извадка от n цикъла

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

където $M_k^{(i)}$ и $L_k^{(i)}$ са стойностите на статистиките в (4.4) в i-тия цикъл от ИНМ за $i = 1, 2, \ldots, n$.

За да сравним статистиките описани в **§ 4.2** е необходимо да определим алтернативен модел за несъвършени наредби. В **§ 4.3** разглеждаме моделите базиран на разстояние (2.2) като алтернатива за несъвършени наредби и изучаваме техните свойства в контекста на ИНМ. Показваме, че предложения алтернативен модел се определя еднозначно чрез матрицата от вероятностните за грешка при наредба $\mathbf{Q}(k, \theta)$, която съвпада с маргиналната матрица в **Глава 2**.

Едно от предимствата за разглеждането на модела на Mallows като алтернатива на съвършената наредба е, че неизвестните параметри в модела могат да се оценят. Нека $\mathbf{X}_{RSS} = \big\{ X_{i[j]}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, k \big\}$ е балансирана ИНМ от n цикъла, $R_{i[j]}$ е номерът на множеството в i-тия цикъл съответстващо на j-тата наредена статистика за $i = 1, 2, \ldots, n$ и $j = 1, 2, \ldots, k$, и $R_i = \big\langle R_{i[1]}, R_{i[2]}, \ldots, R_{i[k]} \big\rangle$ за $i = 1, 2, \ldots, n$. Използвайки модела на Mallows за несъвършени наредби, описан чрез матрицата от грешките $\mathbf{Q}(k, \theta)$ и параметъра

$\theta \leq 0$, изразяваме функцията на правдоподобие чрез следния израз:

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

където $\sum_{l}$ е сумата по всички възможни вектори $l = \left(l_{[1]}, l_{[2]}, \ldots, l_{[k]}\right)$ такива, че $l_{[j]} \in \{1, 2, \ldots, k\}$ за $j = 1, 2, \ldots, k$. С $R = (R_1, R_2, \ldots, R_n)$ е означен векторът от наблюдаваните подредби в ИНМ, $q\left(R_{i[j]}, l_{[j]}, k, \theta\right)$ са елементите на $\mathbf{Q}(k, \theta)$, а $p\left(l_{[1]}, l_{[2]}, \ldots, l_{[k]}\right)$ са вероятностите да се наблюдава $\langle l_{[1]}, l_{[2]}, \ldots, l_{[k]} \rangle$ при съвършена наредба. Понеже елементите на матрицата $\mathbf{Q}(k, \theta)$ не могат се изразят явно и θ не може да се оцени директно, използваме EM алгоритъма за да намерим максимално правдоподобни оценки (МПО) за θ. В **§ 4.4** показваме, че стойността $\theta^{(t+1)}$ във всяка итерация, за която се максимизира очакването на логаритмичната функция на правдоподобие за стойността $\theta^{(t)}$ от предишната итерация, е решение на уравнението

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

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

където $q'(r, z, k, \theta)$ е производната на $q(r, z, k, \theta)$ по отношение на θ и може да се изрази като

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

за $r, z = 1, 2, \ldots, k$. Доказваме, че EM алгоритъма е монотонно сходящ към някоя стационарна точка на $L\left(R \mid \theta\right)$ и че ако $L\left(R \mid \theta\right)$ има единствена стационарна точка максимум, то алгоритъмът е сходящ към единствена МПО за θ.

Тъй като във всяка стъпка на представения EM алгоритъм е необходимо да се пресметне матрицата $\mathbf{Q}(k, \theta)$, прилагането на този

метод изисква огромни изчислителни ресурси за големи стойности на k. В **§ 4.5** представяме алтернативен начин за намиране на матрицата $\mathbf{Q}(k,\theta)$ за разстоянията на Cayley и Hammning, който намалява броя на операциите в (4.9). Аналогично на асимптотичния резултат в **Теорема 2.1** за матрицата от вероятностните за грешка получена от разстоянието на Lee, извеждаме две приближения на матриците получени от правилото на Spearman и ρ на Spearman.

Теорема 4.1. *Нека* $\mathbf{Q}(k,\theta)$ *е матрицата от вероятностните за грешка при наредба получена от правилото на Spearman. Тогава*

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

където

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

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

и

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

Теорема 4.2. *Нека* $\mathbf{Q}(k,\theta)$ *е матрицата от вероятностните за грешка при наредба получена от* ρ *на Spearman. Тогава*

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

където

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

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

$$-\frac{h(r)+h(s)-(i-s)^2-(r-j)^2}{k-1}\Bigg]^2\Bigg\}-\frac{k^2(k-1)(k+1)^2}{36}$$

u

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

В **§ 4.6** прилагаме модела на Mallows като алтернатива за несъвършена наредба и сравняваме описаните статистики за ИНМ от един цикъл. В **§ 4.7** илюстрираме ползата от алтернативата на Mallows за ИНМ от n цикъла като анализираме примерни данни дадени в Murrary и др. [20].

Глава 5. Разстояние на Lee в рангови критерии за две извадки

В тази глава изучаваме ранговата статистика получена от разстоянието на Lee и общия подход на Critchlow към задачата за местоположението на две извадки. Нека $X_1, X_2, \ldots, X_m$ и $Y_1, Y_2, \ldots, Y_n$ са две независими извадки с непрекъснати функции на разпределение съответно $F(x)$ и $G(x)$. Разглеждаме задачата за проверка на нулевата хипотеза H_0 срещу алтернативата H_1, където

$$H_0 : F(x) \equiv G(x) \tag{5.1}$$
$$H_1 : F(x) \geq G(x), \tag{5.2}$$

със строго неравенство за някое x. Ако $\alpha(i)$ е рангът на X_i за $i = 1, 2, \ldots, m$ и $\alpha(m+j)$ е рангът на Y_j за $j = 1, 2, \ldots, n$ в обединената извадка $X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n$, тогава ранговият вектор за всички наблюдения е $\alpha = \langle \alpha(1), \alpha(2), \ldots, \alpha(m+n) \rangle$ и $\alpha \in \mathbf{S}_{m+n}$.

В **§ 5.1** представяме метода на Critchlow [5] за конструиране на критерий, който е основан на намирането на минималното разстояние между точките от класа на еквивалентност $[\alpha] = \alpha(\mathbf{S}_m \times \mathbf{S}_n) = \{\alpha \circ \pi : \pi \in \mathbf{S}_m \times \mathbf{S}_n\}$ и екстремалното множество $E = \mathbf{S}_m \times \mathbf{S}_n = \{\pi \in \mathbf{S}_{m+n} : \pi(i) \leq m, \forall i \leq m\}$, т.е.

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

където d е произволно разстояние върху $\mathbf{S}_{m+n}$.

В **§ 5.2** показваме, че ранговата статистика в (5.3), получена от разстоянието на Lee, е решение на задачата

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

където $e = \langle 1, 2, \ldots, m+n\rangle$ е пермутацията идентитет. Изразяваме $d_L([\alpha], E)$ като

$$d_L([\alpha], E) = 2 \sum_{i\in K_m} \min\left(\mid \alpha(i) - \gamma_n^{-1}(k+1-\gamma_m(\alpha(i))) \mid,\right.$$
$$\left. m+n- \mid \alpha(i) - \gamma_n^{-1}(k+1-\gamma_m(\alpha(i))) \mid\right) \quad (5.5)$$

където

$$K_m = \{i \in \{1,2,\ldots,m\} : \alpha(i) > m\} \quad , \quad (5.6)$$
$$K_n = \{i \in \{m+1, m+2, \ldots, m+n\} : \alpha(i) \le m\} \quad ,$$

k е броят на елементите на K_m $(k =\mid K_m \mid = \mid K_n \mid)$, $\gamma_m(\alpha(i))$ е рангът на $\alpha(i)$ сред множеството $\{\alpha(i) : i \in K_m\}$, $\gamma_n(\alpha(i))$ е рангът на $\alpha(i)$ сред $\{\alpha(i) : i \in K_n\}$ и γ^{-1} е обратната на γ, т.е.

$$\gamma^{-1}(\gamma(\alpha(i))) = \alpha(i).$$

Тъй като $d_L([\alpha], E)$ е еквивалентна на

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

използваме $L_{m,n}$ за проверка на H_0 срещу H_1.

В **§ 5.3** доказваме следното твърдение като използваме факта, че $L_{m,n}$ е минималната сума от разстояния върху C между елементите на K_m и K_n, където C е прост цикличен граф с върхове $\{i\}_{i=1}^{m+n}$ и ръбове $\bigcup_{i=1}^{m+n-1}\{i, i+1\}$ и $\{m+n, 1\}$.

Твърдение 5.1. *Нека $L_{m,n}$ е дефинирано чрез* (5.7) *и $H_{m,n} =\mid K_m \mid$ е броят на елементите на множеството K_m, дефинирано в* (5.6).

Тогава съвместното разпределение на $L_{m,n}$ *и* $H_{m,n}$ *при вярна нулева хипотеза се задава чрез*

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

$\left[\dfrac{k^2+1}{2}\right]\le l\le\left[\dfrac{(m+n-k)\,k+1}{2}\right]$, *където* $[x]$ *е цялата част на* x. *Първата сума в (5.8) е по всички положителни цели числа* s *такива, че* $1\le s\le k+1$ *и* $(s-1)^2+(k-s+1)^2\le l$. *Втората сума е по всички неотрицателни цели числа* $\{a_i^{(m)}\}_{i=0}^{s-1}$, $\{a_i^{(n)}\}_{i=0}^{s-1}$, $\{b_j^{(m)}\}_{j=1}^{k-s+1}$ *и* $\{b_j^{(n)}\}_{j=1}^{k-s+1}$, *които удовлетворяват:*

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

където $s\in\{1,2,\ldots,k+1\}$ *в условия* **(i)-(iii)**, $s\in\{1,2,\ldots,k\}$ *в условие* **(iv)** *и* $s\in\{2,3,\ldots,k+1\}$ *в условие* **(v)**. *Целите числа* $b_0^{(m)}$ *и* $b_0^{(n)}$ *се дефинират като нула,* $b_0^{(m)}:=0$, $b_0^{(n)}:=0$, *за пълнота в условия* **(i)-(v)**.

Тъй като проверката на условия на **(i)-(v)** по всички възможни неотрицателни цели числа $\{a_i^{(m)}\}_{i=0}^{s-1}$, $\{a_i^{(n)}\}_{i=0}^{s-1}$, $\{b_j^{(m)}\}_{j=1}^{k-s+1}$ и $\{b_j^{(n)}\}_{j=1}^{k-s+1}$ отнема много време, когато стойностите на m и n са големи, то извеждаме рекурсивни формули за броя на членовете в

сумите (5.8) (виж **Твърдение 5.2**) и за вероятностите на $L_{m,n}$ при вярна H_0 (виж **Твърдение 5.3**).

Очакването и дисперсията на статистиката $L_{m,n}$, както и нейното асимптотично разпределение при H_0, са изведени в следната теорема.

Теорема 5.1. *Нека $L_{m,n}$ е дефинирано чрез* (5.7). *Тогава очакването и дисперсията на $L_{m,n}$ при вярна нулева хипотеза H_0 са*

$$\mathbf{E}\left(L_{m,n}\right) = \begin{cases} \frac{mn(m+n+1)}{4(m+n)} & \text{, ако } m+n \text{ е нечетно} \\ \\ \frac{mn(m+n)}{4(m+n-1)} & \text{, ако } m+n \text{ е четно,} \end{cases} \tag{5.16}$$

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

където първият случай е при $m+n$ нечетно, а вторият е при $m+n$ четно.

Освен това, нормираната статистика $\frac{L_{m,n}-\mathbf{E}(L_{m,n})}{\sqrt{\mathbf{Var}(L_{m,n})}}$ има асимптотично нормално разпределение при $m,n \to \infty$ и вярна H_0.

Този резултат дава нормално приближение на разпределението на $L_{m,n}$ при вярна H_0 и е изключително полезен при определянето на критичната област за големи m и n.

В **§ 5.4** сравняваме ранговата статистика $L_{m,n}$, дефинирана в (5.7), със статистиките на Student, Wilcoxon, Kolmogorov-Smirnov и Mood за две извадки. Чрез илюстративно симулационно изследване показваме, че критерият построен от разстоянието на Lee е помощен от останалите, когато разпределенията на двете генерални съвкупности са с тежки опашки. Заключваме, че за задачата с местоположението на две извадки има компромис между мощност на критерия и неговата устойчивост по отношение на разпределението на генералната съвкупност.

Доказателствата на Теореми 2.1, 4.1, 4.2 и 5.1 са дадени в **Приложение Е**.

Авторска справка

По мнение на автора, основните приноси в дисертацията са:

1. Случайната величина получена от разстоянието на Lee при равномерно разпределени наредби е подробно изучена и са изведени някои нейни характеристики като очакване, дисперсия и симетричност за произволна големина на наредбите. Доказана е асимптотична нормалност на нейното разпределение, което се използва за приближаване на нормиращата константа в модела базиран на разстояние. Този резултат намира приложение и в други модели за наредени данни, в които се използва разстоянието на Lee.

2. EM алгоритъмът за изчисляване на максимално правдоподобни оценки на параметрите в модела с латентни класове и базиран на разстояние е обобщен в случая, когато *модалните* наредби на латентните класове не са известни. Доказана е сходимостта на предложения алгоритъм към стационарна точка и методът е приложен към добре изследваните изборни данни на Американската психологична асоциация (ACA). Използвайки EM алгоритъма, можем да приложим модела към данните, да правим статистически изводи и да сравняваме модели базирани на различни разстояния върху пермутации.

3. За клъстеризацията с "K-средни" основана на разстоянието на Lee е получено асимптотично приближение на нормиращата константа в мярката на "сгъстеност". Полученият резултат значително намалява времето и ресурсите за изчисляване на коефициента на "сгъстеност" при клъстеризация с две групи ($K = 2$) и при относително голям размер на наредбите ($N \geq 7$).

4. Моделът на Mallows е разгледан като алтернативен модел за несъвършени наредби при балансирани извадки от наредени множества (ИНМ). Описан е EM алгоритъм за оценка на неизвестните параметри в модела и е показана неговата сходимост към стационарна точка. Изведени са асимптотични резултати за матриците от вероятностните за грешка получени от правилото на Spearman, ρ на Spearman и разстоянието на Lee при голям брой наблюдения на ИНМ във всеки цикъл. Предложеният алтернативен модел може да се използва за изучаването

на ефекта на несъвършена наредба върху ефективността на някои статистически процедури за ИНМ, както и за сравняване на способностите за наредба на два метода.

5. Изведен е непараметричен рангов критерий получен от разстоянието на Lee и метода на Critchlow за задачата за местоположението на две извадки. Доказана е асимптотична нормалност на получената статистика при вярна нулева хипотеза, което е полезно за намирането на критичните области при извадки с голям обем. Чрез симулационно изследване е показано, че статистиката получена от разстоянието на Lee е по-мощна, когато разпределенията на двете генерални съвкупности са с тежки опашки.

Списък на публикациите по дисертацията

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, (приета).

Апробация на резултатите

Резултатите от дисертацията са докладвани на следните научни форуми:

1. *"Lee distance in two-sample rank tests"*, 11-th International Conference: Computer Data Analysis and Modeling, Minsk, Belarus (07.09.2016).

2. *"Mallows' models based on Lee distance"*, 20-th European Young Statisticians Meetings, Uppsala, Sweden (17.08.2017).

3. *"Mallows' models for imperfect rankings in ranked set sampling"*, 13-th International Conference on Ordered Statistical Data Cadiz, Spain (22.05.2018).

4. *"Some properties of Lee distance in two-sample location problem"*, 18-th International Summer Conference on Probability and Statistics, Pomorie, Bulgaria (27.06.2018).

5. *"Rank data models based on Lee distance"*, International Conference on Trends and Perspectives in Linear Statistical Inference, Bedlewo, Poland (21.08.2018).

6. *"Two-sample rank test based on Lee distance"*, 15-th Applied Statistics International Conference, Ribno, Slovenia (24.09.2018).

7. *"Distance-based models for imperfect ranking in ranked set sampling"*, XLIV Mathematical Statistics Conference, Bedlewo, Poland (03.12.2018).

8. *"Rank data clustering based on Lee distance"*, 13-th Annual Meeting of the Bulgarian Section of SIAM, Sofia, Bulgaria (19.12.2018).

Благодарности

Изказвам сърдечна благодарност и дълбока признателност на научния си ръководител проф. Евгения Стоименова за ценните напътствия, мотивация и помощ. Бих желал да благодаря и на колегите от секция "Изследване на операциите, вероятности и статистика" на ИМИ-БАН за тяхната подкрепа и приятелско отношение.

Изследванията в дисертацията са подкрепени по Национална програма "Млади учени и постдокторанти", ПМС №577/17.08.2018.

Списък на преведените понятия

Понятие на английски	Превод на български
Antimodal ranking	Антимодална наредба
Cluster	Клъстер (група)
Distance-based model	Модел базиран на разстояние
Goodness-of-fit	Съгласуваност
Imperfect ranking	Несъвършена наредба
"K-means" clustering	Клъстеризация с "К-средни"
Kendall's tau	τ на Kendall
Latent group	Латентна група
Latent-class Distance-based model	Модел с латентни класове и базиран на разстояние
Lee distance	Разстояние на Lee
Marginals model	Маргинален модел
Measure of "tightness"	Мярка на "сгъстеност"
Modal ranking	Модална наредба
Ordering	Подредба
Perfect ranking	Съвършена наредба
Population distribution	Разпределение на генерална съвкупност
Rank	Ранг
Rank data	Наредени данни
Ranked set sampling (RSS)	Извадки от наредени множества (ИНМ)
Ranking	Наредба
Ranking error probability matrix	Матрица от вероятностните за грешка при наредба
Spearman's footrule	Правило на Spearman
Spearman's rho	ρ на Spearman
T-test	Критерий на Student

Библиография

[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.