Mallows' Model Based on Lee Distance

Nikolay I. Nikolov*[1] and Eugenia Stoimenova[1,2]

[1] *Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G.Bontchev str., block 8, 1113 Sofia, Bulgaria*
[2] *Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G.Bontchev str., block 25A, 1113 Sofia, Bulgaria*

In this paper the Mallows' model based on Lee distance is considered and compared to models induced by other metrics on the permutation group. As an illustration, the complete rankings from the American Psychological Association election data are analyzed.

Keywords: Rank data analysis, Mallows' models, Lee distance, Metrics on permutations

1 Mallows' models

A full ranking of N items simply assigns a complete ordering to the items. Any such ranking vector can be viewed as an element π of the permutation group $\mathcal{S}_N$ generated by the first N natural integers. Thus the notation

$$\pi =< \pi^{-1}(1), \pi^{-1}(2), \ldots, \pi^{-1}(N) >$$

is used for the permutation π which corresponds to listing the various items in their ranked order. There are various nonparametric methods for modelling rank data. Some models have larger probabilities for rankings that are "close" to a "modal" ranking π_0. An example of such probability model is given by

$$P_{\theta,\pi_0}(\pi) = e^{\theta d(\pi,\pi_0) - \psi(\theta)} \quad \text{for } \pi \in \mathcal{S}_N, \tag{1}$$

where θ is a real parameter ($\theta \in \mathbb{R}$), $d(\cdot,\cdot)$ is a metric on $\mathcal{S}_N$, π_0 is a fixed ranking and $\psi(\theta)$ is a normalizing constant. When $\theta > 0$, π_0 is the modal ranking, for $\theta < 0$, π_0 is an antimode, and for $\theta = 0$, P_{θ,π_0} is the uniform distribution. More general model, with $d(\cdot,\cdot)$ being a discrepancy function, is suggested by Diaconis [4], but since all distances used in this paper are metrics, $d(\cdot,\cdot)$ could be regarded as a metric. Deza and Huang [3] considered some metrics on $\mathcal{S}_N$ which are widely used in applied scientific and statistical problems.

*Corresponding author: n.nikolov@math.bas.bg

$$F(\pi,\sigma) = \sum_{i=1}^{N} \mid \pi(i) - \sigma(i) \mid \qquad \text{Spearman's footrule}$$

$$R(\pi,\sigma) = \left(\sum_{i=1}^{N} (\pi(i) - \sigma(i))^2 \right)^{1/2} \qquad \text{Spearman's rho}$$

$$M(\pi,\sigma) = \max_{1\le i\le N} \mid \pi(i) - \sigma(i) \mid \qquad \text{Chebyshev metric}$$

$$K(\pi,\sigma) = \#\{(i,j) : 1 \le i,j \le N, \pi(i) < \pi(j), \sigma(i) > \sigma(j)\} \qquad \text{Kendall's tau}$$

$$C(\pi,\sigma) = N \text{ minus number of cycles in } \sigma\pi^{-1} \qquad \text{Cayley's distance}$$

$$U(\pi,\sigma) = N \text{ minus length of the longest increasing subsequence in } \sigma\pi^{-1} \qquad \text{Ulam's distance}$$

$$H(\pi,\sigma) = \#\{i \in \{1,2,\ldots,N\} : \pi(i) \ne \sigma(i)\} \qquad \text{Hamming distance}$$

$$L(\pi,\sigma) = \sum_{i=1}^{N} \min\left(\mid \pi(i) - \sigma(i) \mid, N - \mid \pi(i) - \sigma(i) \mid\right) \qquad \text{Lee distance}$$

Easily can be shown that all of the presented metrics possess the following important property.

Definition 1. The metric d on $\mathcal{S}_N$ is called right-invariant, if and only if $d(\pi,\sigma) = d(\pi\circ\tau, \sigma\circ\tau)$ for all $\pi,\sigma,\tau \in \mathcal{S}_N$.

Critchlow [1] pointed that the right-invariance of metric is necessary requirement since it means that the distance between rankings does not depend on the labelling of the items. More properties for these metrics can be found in Critchlow [1, 2], Diaconis [4] and Marden [7].

If $d(\cdot,\cdot)$ is right-invariant, then (1) can be defined by the random variable $D(\pi) = d(\pi,\pi_0) = d(\pi\pi_0^{-1}, e_N)$, where $\pi \sim Uniform(\mathcal{S}_N)$ and e_N is the identity permutation ($e_N =< 1,2,\ldots,N >$). Notice that the distribution of D does not depend on π_0 and it could be assumed that $D(\pi) = d(\pi, e_N)$. Let's use the notation $D_{[*]}$ for the random variable D induced by some distance $[*]$ from the listed above. The special cases of (1) with $D = D_K$ and $D = D_{R^2}$ are first investigated by Mallows [6]. Models based on D_C and D_H can be found in Fligner and Verducci [5].

Model (1) could be significantly simplified if the distribution of D is known and can be written explicitly. Let $m(t)$ be the moment generating function of D. Then, as shown in [5],

$$e^{\psi(\theta)} = \sum_{\pi \in S_N} e^{\theta D(\pi)} = N! \sum_{d_i} P(D = d_i) e^{\theta d_i} = N!\, m(\theta)$$
$$\Rightarrow \quad \psi(\theta) = \log(N!\, m(\theta)) \,. \tag{2}$$

For D_F, D_R, D_M, D_K, D_C, D_U and D_H numerical characteristics, exact distributions, asymptotic approximations and statistical applications can be found in Diaconis [4] and Marden [7]. The goal of this paper is to study the Mallows' model based on D_L and compare it to the models induced by the other given metrics. The rest of the paper is organized as follows. In Section 2 some properties of the distribution of Lee distance are derived under uniformity assumption. Maximum likelihood estimations and testing procedure for deviation from the Uniform distribution are proposed in Section 3. In Section 4 a comparison between the models based on the eight distances is made.

2 Lee distance

Let's first notice that $D_L(\pi) = L(\pi, e_N)$ can be decomposed in N terms:

$$D_L(\pi) = \sum_{i=1}^{N} \min\left(\mid \pi(i) - i \mid, N - \mid \pi(i) - i \mid\right) = \sum_{i=1}^{N} c_N(\pi(i), i) \,. \tag{3}$$

There is an interpretation of $c_N(i,j) := \min\left(\mid i - j \mid, N - \mid i - j \mid\right)$ 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 distances over G between the nodes i and j. 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.

$$0 \le c_N(i,j) \le \left[\frac{N}{2}\right], \quad \text{for all } i, j = 1, 2, \ldots, N \,, \tag{4}$$

where $[x]$ is the greatest integer less than or equal to x. From (3) and (4) it follows that

$$0 \le D_L(\pi) \le N \left[\frac{N}{2}\right], \quad \text{for all } \pi \in \mathcal{S}_N \,. \tag{5}$$

The lower limit in (5) is reached only for $\pi = e_N$. When N is even the upper limit is reached only for π equals to

$$e_N^* := < \frac{N}{2} + 1, \frac{N}{2} + 2, \ldots, N - 1, N, 1, 2, \ldots, \frac{N}{2} - 1, \frac{N}{2} > \,,$$

and in the case of odd integers N the maximum value of D_L is reached when π is equal to e'_N or e''_N, where

$$e'_N :=< \frac{N+1}{2}, \frac{N+1}{2}+1, \dots, N-1, N, 1, \dots, \frac{N+1}{2}-2, \frac{N+1}{2}-1 >$$
$$e''_N :=< \frac{N+1}{2}+1, \frac{N+1}{2}+2, \dots, N-1, N, 1, \dots, \frac{N+1}{2}-1, \frac{N+1}{2} > .$$

Since

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

and

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

the relation

$$L(\pi, e_N) + L(\pi, e^*_N) = \sum_{i=1}^{N} c_N(\pi(i), e_N(i)) + c_N(\pi(i), e^*_N(i)) = \frac{N^2}{2} \quad , \qquad (6)$$

is true for all $\pi \in \mathcal{S}_N$. The right-invariant property of L implies that $L(\pi, e_N)$ and $L(\pi, e^*_N)$ have the same distribution when $\pi \sim Uniform(\mathcal{S}_N)$. From that fact and (6) it follows that

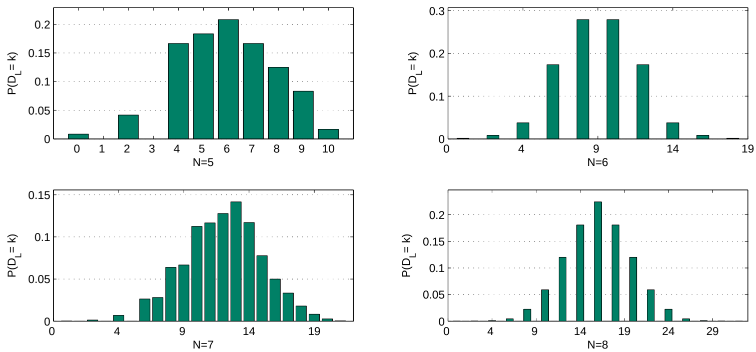
$$P(D_L = k) = P\left(D_L = \frac{N^2}{2} - k\right) \quad , \text{ for } k = 0, 1, \dots, \frac{N^2}{2} \quad , \text{ i.e.}$$

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

$$D_L(\pi) \equiv \sum_{i=1}^{N} \min\left(\mid \pi(i) - i \mid, N - \mid \pi(i) - i \mid\right) \pmod 2$$
$$\Rightarrow D_L(\pi) \equiv \sum_{i=1}^{N} \mid \pi(i) - i \mid \equiv 0 \pmod 2$$

for even integers N.

The probability mass function of D_L for $N = 5, 6, 7, 8$ is shown on the figure below.



3 Parameters estimation and tests for uniformity

Formula (2) can be used to find estimations for the unknown parameters in (1). Suppose that there are n observed complete rankings $\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(n)}$ and the mode π_0 in (1) is unknown. Then the loglikelihood function is given by

$$l(\theta, \pi_0, n) = \theta S(\pi_0) - n\psi(\theta),$$

where $S(\pi_0) = \sum_{i=1}^{n} d(\pi^{(i)}, \pi_0)$. In order to find the maximum likelihood estimations (MLE's), first it is necessary to calculate

$$\hat{\pi}_{min} = \operatorname*{argmin}_{\pi \in \mathcal{S}_N} S(\pi) \quad \text{and} \quad \hat{\pi}_{max} = \operatorname*{argmax}_{\pi \in \mathcal{S}_N} S(\pi)\,.$$

For $\theta < 0$, let $\hat{\theta}_{min}$ be the value for which $l(\theta, \hat{\pi}_{min}, n)$ is maximized. For $\theta > 0$, let the maximum of $l(\theta, \hat{\pi}_{max}, n)$ occurs for $\theta = \hat{\theta}_{max}$. Finally, the MLE's

$$\left(\hat{\theta}, \hat{\pi}_0\right) = \begin{cases} \left(\hat{\theta}_{min}, \hat{\pi}_{min}\right), & \text{if } l(\hat{\theta}_{min}, \hat{\pi}_{min}, n) \geq l(\hat{\theta}_{max}, \hat{\pi}_{max}, n) \\ \left(\hat{\theta}_{max}, \hat{\pi}_{max}\right), & \text{otherwise}\,. \end{cases}$$

If $\hat{\theta} = 0$ then (1) is the uniform model and $\hat{\pi}_0$ is not unique since for all $\pi \in \mathcal{S}_N$ the loglikelihood, $l(0, \pi, n)$, is the same. For Spearman's rho $R(\cdot, \cdot)$ and Kendall's tau $K(\cdot, \cdot)$ it can be shown that $\hat{\theta}_{min} = -\hat{\theta}_{max}$. From (6) it follows that $\hat{\theta}_{min} = -\hat{\theta}_{max}$ is also valid for Lee distance $L(\cdot, \cdot)$, when N is even. In these cases $l(\hat{\theta}_{min}, \hat{\pi}_{min}, n) = l(\hat{\theta}_{max}, \hat{\pi}_{max}, n)$ and it is enough to find just $\hat{\pi}_{min}$ and $\hat{\theta}_{min}$. The described MLE's and other methods for estimating θ and π_0 can be found in [7].

For testing the null hypothesis $H_0 : \theta = 0$ (*uniform model*) against the alternative $H_A : \theta \neq 0$, Marden [7] considered the likelihood ratio statistic (LRS) given by

$$LRS = 2\left[l_A(\hat{\theta}, \hat{\pi}_0, n) - l_0(0, \pi, n)\right] = 2\left[\hat{\theta} S(\hat{\pi}_0) - n\psi(\hat{\theta}) + n\log(N!)\right],$$

where l_0 and l_A are the loglikelihood functions under H_0 and H_A, respectively, and $\left(\hat{\theta}, \hat{\pi}_0\right)$ are the MLE's. Let $k(\pi)$ be the number of observations that are equal to $\pi \in \mathcal{S}_N$. Then the empirical probability for π is $\dfrac{k(\pi)}{n}$ and a quantity, which measures the total nonuniformity of the data, could be defined by

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

Similarly to the multiple correlation coefficient in the linear regression, Marden [7] considered the coefficient

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

which can be used to measure the percentage of nonuniformity in the data that is explained by the fitted model. Thus $R^2 = 1$ when the model exactly fits the data, and $R^2 = 0$ if it performs no better than the uniform model.

4 Comparison between the distance based models

In 1980, the American Psychological Association (APA) conducted an election in which five candidates were running for president and voters were asked to rank order all of the candidates. The complete rankings of 5738 voters are given in [4, p. 96]. The average ranks received by 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 fitted Mallows' models based on the eight distances considered are given in Table 1.

Since the theoretical distribution of LRS is unknown, it is approximated via simulations with 1000 trials for each distance, and the results for the mean and the 95% critical values of LRS's are presented in the last two columns. Notice that for all models the hypothesis of uniform distribution is rejected since LRS's are much larger than the simulated critical values. In fact all LRS's are larger than the maximum simulated values. However, all models explain less than a third of the nonuniformity, where the model based on D_L

Distance	$\hat{\theta}$	$\hat{\pi}_0$	**Ordering**	LRS	R^2	LRS_{sim} **mean**	LRS_{sim} **95% c.v.**
D_F	0.0828	51324	**BDCEA**	282.26	0.1643	4.63	9.62
D_{R^2}	−0.0163	15243	**ACEDB**	150.78	0.0878	3.49	7.95
D_M	−0.2639	15243	**ACEDB**	379.54	0.2210	5.49	10.43
D_K	−0.0722	15243	**ACEDB**	124.28	0.0723	3.83	8.43
D_C	−0.2483	23154	**CABED**	304.21	0.1771	7.22	11.98
D_U	−0.2505	23154	**CABED**	181.52	0.1057	6.80	12.06
D_H	0.2437	51324	**BDCEA**	290.16	0.1689	6.74	11.41
D_L	0.1656	51324	**BDCEA**	524.39	0.3053	5.52	10.73

Table 1: *Fitted Mallows' models for APA data*

has the highest $R^2 = 30.53\%$, and the lowest $R^2 = 7.23\%$ is obtained when using D_K.

The estimated "modal" orderings (antimodes for $\hat{\theta} > 0$) are given in the forth column. The ordering of D_{R^2}, D_M and D_K coincides with the "modal" ordering based on the average ranks. As mentioned in [7], 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 from the orderings of D_{R^2}, D_M, D_K, D_C and D_U. Since the number of candidates $N = 5$ is odd, the maximum value of $L(e_N, \pi)$ is reached for $\pi = e'_N$ and $\pi = e''_N$. Thus the interpretation of the antimode ordering of D_L is more complex.

Candidate B is ranked last in all "modal" rankings, except in models based on D_C and D_U, where B separates the groups {A,C} and {D,E}. The rankings, which are constructed without considering candidate B, could be used to study the influence of B over the complete rankings models. The MLE's of the models' parameters for the new rankings are given in Table 2.

Distance	$\hat{\theta}$	$\hat{\pi}_0$	**Ordering**	LRS_{new}	R^2_{new}	LRS_{diff}	**Simulated** LRS_{diff} **cdf**
D_F	−0.0698	2143	**CAED**	126.22	0.1268	156.05	0.591
D_{R^2}	−0.0177	1243	**ACED**	59.79	0.0601	90.99	0.769
D_M	−0.2239	2143	**CAED**	226.30	0.2273	153.23	0.001
D_K	−0.0663	1243	**ACED**	54.64	0.0549	69.64	0.742
D_C	−0.2532	2143	**CAED**	251.79	0.2529	52.42	0.000
D_U	−0.2319	2143	**CAED**	124.75	0.1253	56.77	0.006
D_H	−0.1832	2143	**CAED**	217.59	0.2186	72.58	0.000
D_L	−0.1311	2143	**CAED**	265.39	0.2666	259.00	0.007

Table 2: *Fitted models without candidate B*

The total nonuniformity of the new data is $TNU_{new} = 995.58$. The "modal" orderings of the remaining four candidates are not changed in the models based on D_{R^2}, D_K, D_C and D_U, whereas there are new "modal" orderings in the other models. For D_C, D_U and D_H the coefficient R^2_{new} increases, while for D_F, D_{R^2}, D_K and D_L it decreases. R^2_{new} is almost the same as R^2 for the model based on D_M. The quantity $LRS_{diff} = LRS - LRS_{new}$ can be used to measure the influence of candidate B over the explanatory power of the models. The value of LRS_{diff} is simulated 1000 times for all complete models with parameters given in Table 1. The observed value of LRS_{diff} and the simulated empirical cumulative distribution function (taken at the observed value of LRS_{diff}) are given in the last two columns. There is a significant decrease in the explanatory power of the models based on D_F, D_{R^2} and D_K, since the values of LRS_{diff} are significant for these models. Thus it can be suggested that the models based on D_M, D_C, D_U, D_H and D_L are more "robust". Similar conclusion is made in [7, p. 30] by analyzing sport related rank data.

Acknowledgements: This work was supported by the grant I02/19 of the Bulgarian National Science Fund.

References

[1] D. E. Critchlow. *Metric Methods for Analyzing Partially Ranked Data.* Lecture Notes in Statistics, 34. Springer, New York, 1985.

[2] D. E. Critchlow. On rank statistics: an approach via metrics on the permutation group. *J. Statist. Plann. Inference*, 32(3):325–346, 1992.

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

[4] P. Diaconis. *Group Representations in Probability and Statistics.* IMS Lecture Notes - Monograph Series, 11. Hayward, Carifornia, 1988.

[5] M. Fligner and T. Verducci. Distance based ranking models, *Journal of the Royal Statistical Society*, 48(3):359–369, 1986.

[6] C. M. Mallows. Non-null ranking models. I. *Biometrika*, 44(1):114–130, 1957.

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