Long Review of Doctoral Thesis

Author: Nikolay Ivanchev Nikolov

Title: Metric Methods for Analyzing and Modeling Rank Data

Reviewer: Professor N. Balakrishnan, PhD

Background: Nikolay Nikolov completed his Bachelor program in Applied Mathematics and then completed his Masters program in Mathematical Modeling in Economic from Sofia University, Sofia, Bulgaria. Since then, he has been a PhD student in Probability Theory and Mathematical Statistics in the Institute of Mathematics and Informatics at the Bulgarian Academy of Science. There, he became an Assistant Professor since November 2018.

General Description: This thesis, of 83 pages, includes five chapters, five Appendices, Conclusion and a list of 83 References in total. The thesis has been presented for acquiring the educational and scientific degree of "Doctor" in the PhD program on Probability Theory and Mathematical Statistics.

The thesis starts by providing in Chapter 1 a brief discussion on some popular distance measures that have been discussed in the literature for measuring distances between rank vectors. It then goes into detail about Lee distance, in particular, and its properties including its moments and asymptotic distribution under uniformity over the set of permutations. This lays the groundwork for all developments in subsequent chapters of the thesis.

Next, Chapter 2 begins with a short description of distance-based models and marginals models for modeling ranked data and establishes some specific properties for the case when Lee distance is used in these models. Detailed discussions are then made on statistical inference for these models and the maximum likelihood estimation of the model parameters with the use of EM algorithm is discussed in length. Finally, a simulation study is carried out to evaluate the models and methods introduced in the preceding sections and three known data sets are also used to illustrate all the results.

Chapter 3 focuses on the problem of clustering when the observed data are on ranks. Here again, Lee distance is utilized to develop methods of clustering and finally a relative performance evaluation is made with the method based on some other distances such as Spearman's, Chebychev's, Kendall's, Cayley's, Ulam's and Hamming's distance measures.

In Chapter 4, the application of Lee distance is taken up in the context of ranked set sampling, and specifically to model the judgement error involved in ranking the objects through easy visual inspection involved in ranked set sampling. Then, two new test statistics are also added to those proposed and studied by Li and Balakrishnan [43], with these two new proposals being based on Chebychev and Lee distances. Then, the discussion shifts to Mallows' model for imperfect ranking and the maximum likelihood estimation of the parameter θ involved in Mallows' model. The detailed steps of the EM-algorithm for the determination of the maximum likelihood estimates are then presented and the use of the error probability matrix based on different distances is then described. Finally, by adopting the simulation study carried out by Li and Balakrishnan [43], a detailed simulation study of all the test statistics is carried out in terms of power under both bivariate normal model and Mallows model. This is a well-crafted work, with elegant mathematical developments with regard to the proposed procedures, even though unfortunately the two newly introduced test statistics do not turn out to be better! An illustrative example, from the ranked set sampling literature, is finally used to illustrate all the results discussed and newly developed in the literaure.

Finally, in Chapter 5, the discussion shifts to the well-known Critchlow's method for two-sample location problem in testing for stochastic ordering between the two underlying distributions. A rank test statistic is developed here based again on Lee distance and its properties, moments and asymptotic distributions are all discussed. The performance of the proposed test is evaluated, based on the family of Student t-distributions, and compared with some other prominent nonparametric tests as well as the classical t-test. It turns out that the proposed test, based on Lee distance, is more powerful than all these tests with the degrees of freedom of t-distribution is either 1 or 2 (cases where the mean and variance, respectively, do not exist) but not so for higher values of degrees of freedom.

Summary: The results established in the thesis are sound, technically correct and true.

Publications: This thesis is based on six publications, two of which are in journals with impact factor and three others are in conference proceedings. Moreover, the results of the thesis have also been presented in at least three conferences.

Concluding Evaluation and Comments: The thesis contains a great deal of original work with substantial results, which are of high scientific level, having been presented in a clear and succinct manner. Therefore, in my opinion, this thesis of Nikolay Nikolov fulfils all the conditions required for obtaining the PhD degree in Probability Theory and Mathematical Statistics at the Bulgarian Academy of Sciences.

Typographical errors to be fixed:

1. Page 2, line 2 ↓, change "*details*" to "*detail*"
2. Page 6, in Eq. (1.6), move "*i.e.*" to the line below
3. Page 10, line 2 ↑, change "*ditance*" to "*distance*"
4. Page 11, line 7 ↓, change "*estimations*" to "*estimates*"
5. Page 12, line 7 ↑, change "*In contrast of*" to "*In contrast to*"
6. Page 15, line 18 ↓, change "*Similarly to the*" to "*Similar to the*"
7. Page 18, line 10 ↑, change "*As it is expected*" to "*As expected*"
8. Page 18, line 9 ↑, change "*Models base on*" to "*Models based on*"
9. Page 21, line 2 ↑, change "*Similarly to the results*" to "*Similar to the results*"
10. Page 22, line 4 in the Second paragraph, change "*forth column*" to "*fourth column*"
11. Page 22, line 1 ↑, change "*two or there*" to "*two or three*"
12. Page 23, line 2 ↓ in the first paragraph, change "*As it was expected*" to "*As expected*"
13. Page 24, line 7 ↓, change "*in details*" to "*in detail*"
14. Page 25, line 12 ↓, change "*and drive some*" to "*and derive some*"
15. Page 31, line 7 ↓, insert "$X_{2[k]}$," before "$X_{n[1]}$,"
16. Page 31, same correction in line 2 ↓ of the next paragraph
17. Page 31, line 5 ↑, change "*Let denote by*" to "*Let us denote by*"
18. Page 32, line 1 ↓, change "*Let assume*" to "*Let us assume*"
19. Page 33, line 8 ↓, change "*values the test*" to "*values of the test*"
20. Page 34, line 5 ↑, change "*similarly to the*" to "*as in the*"
21. Page 35, line 4 ↓, change "*Let consider*" to "*Let us consider*"
22. Page 37, line 11 ↓, change "*Hammning*" to "*Hamming*"
23. Page 39, line 2 ↑, change "*based on an bivariate*" to "*based on a bivariate*"
24. Page 43, line 10 ↑, change "*in more details*" to "*in more detail*"
25. Page 44, line 2 ↓, change "*proved their useful*" to "*proved useful*"
26. Page 41, line 12 ↓, change "*studied in details.*" to "*studied in detail.*"
27. Page 51, five lines above plots, change "*let the mean varies*" to "*let the mean vary*"
28. Page 51, in the next line, change "*trails*" to "*trials*"
29. Page 53, line 5 ↓, change "*in details*" to "*in detail*"

30. Page 65, line 6 ↑, changes "*indexes*" to "*indices*"
31. Page 66, change "*of Theorem 4.1.*" to "*Proof of Theorem 4.1.*"
32. Page 70, line 2 ↑, change "*the*" to "*The*"
33. Page 71, in Reference [31], fix the obvious error
34. Page 74, in Reference [76], change "*Marvel Delker*" to "*Marcel Dekker*"

Minor issues to be addressed:

1. On Page 2, it will be prudent to specify the distance "number of switches" proposed by Li and Balakrishnan, and then show how it is related to Wilcoxon distance measure. This would provide a practical and computational motivation for the Wilcoxon distance measure. For example, this switching measure gives the distance between $(4, 1, 3, 5, 2)$ to $(1, 2, 3, 4, 5)$ to be 5.
2. On Page 8, an interesting monotonicity property of $\mathrm{Var}(D_L)$ in Eq. (1.13) can be established. For example, by taking $N = 2M$ and $N = 2m+1$ and using the corresponding expressions for $\mathrm{Var}(D_L)$ given by $\frac{M^4+2M^2}{3(2M-1)}$ and $\frac{M^3+2M^2+2M+1}{6}$, respectively, it can be shown that the difference between the variances for $N = 2M + 1$ and $N = 2M$ equals 0 if and only if $M = 1$ (meaning when we consider $N = 3$ and $N = 2$), and otherwise positive. Similarly, by considering the variance expressions for the cases $N = 2M+2$ and $N = 2M+1$ and taking their difference, it can be shown that the difference is always positive. This would prove that $\mathrm{Var}(D_L)$ is monotonically increasing from $N = 3$ on (and is equal for $N = 2$ and $N = 3$). This will be worth adding there in that discussion!
3. On Page 8, the statement of Theorem 1.2 as it reads is not proper since the mean and variance of the random variable D_L both go to ∞ as $N \to \infty$. It would be better and proper to state it as $(D_L - \mu(D_L))/\sigma(D_L) \to N(0, 1)$ as $N \to \infty$.
4. In fact, by properly rearranging the terms in the expressions for the mean and variance of D_L presented there, by taking into account the order of various terms, a simplified limiting result can be given as $\sqrt{48N}\left(\frac{D_L}{N^2} - \frac{1}{4}\right) \to N(0, 1)$.
5. On Page 10, in the paragraph where estimation of parameters θ and Π_0 are mentioned, it will be good to mention what is the form of data one would base this estimation on.
6. On Page 12, in Table 2.1, there is some symmetry in the values of $\psi_N(\theta)$, and more so in the values of $\hat{\psi}_N(\theta)$.
   Two questions arise here: Firstly, was this symmetry result made use of during estimation and dealt with as an estimation under constraint? If not done, this could possibly increase the precision of estimation. Symmetry in

$\psi_N(\theta)$ is quite clear for even N (for example, when N is 4, for $\theta = -1, +1$, the values are 0.592 and 8.592, with a difference of $N^2/2 = 4^2/2 = 8$; when N is 6, for $\theta = -1, +1$, the values are 0.991 and 18.991, with a difference of $N^2/2 = 6^2/2 = 18$. Similar patterns are observed for $\theta = -0.75, +0.75$, and $\theta = -0.50, +0.50$).

The second issue is that the symmetry pattern is seen in $\hat{\psi}_N(\theta)$ even for odd values of N, but I do not see this in $\psi_N(\theta$. Why? Can this be explained?

7. On Page 15, R^2 is defined as $\frac{LRS}{TNU}$. When $R^2 = 1$, the model exactly fits the data; but when the model is uniform, will not both LRS and TNU be 0. Why $R^2 = 0$?
8. On Page 27, in the bottom, isn't the distribution simply half normal on the negative side since the upper truncation point is same as the mean of the normal distribution? If so, it should be stated explicitly.
9. On Page 28, at the bottom of Table 3.1, there is no point in presenting that approximation since it does not seem to work for any reasonable choices of N and K, and also seems to become worse as they increase.
10. On Page 29, in the conclusion of Chapter 3, perhaps the following needs to be mentioned: Lee-distance based method is not really useful for clustering purpose since it does not seem to separate different choices of K from $K = 1$. Perhaps, a detailed simulation in this regard would have been useful!
11. On Page 43, it will be good to give standard errors of $\hat{\theta}_F^{\text{low}}$ and $\hat{\theta}_F^{\text{high}}$ to statistically conclude whether their values of -0.494 and -0.696 are indeed significantly different.
12. On Page 51, you constantly refer to the mean of Student's t-distribution. Perhaps, you should change it to location (instead of mean) since the mean does not exist when $DF = 1$ (case of Cauchy).
13. On Page 51, perhaps a comment to this point would be useful: Even when $DF = 5$ (small), the classical t-test turns out to be more powerful than most nonparametric tests.
14. On Page 64, reformulate the asymptotic distribution statement as the distribution of $(\bar{D}_R - E(\bar{D}_R))/\sqrt{\text{Var}(\bar{D}_R)}$ is $N(0, 1)$.

Hamilton, Ontario, Canada
March 23, 2020 **N. Balakrishnan**