CSIC 5011 Human age ranking from pairwise comparison data via HodgeRank

Lui Go Nam[1] gnlui@connect.ust.hk
[1]: Department of Mechanical and Aerospace Engineering, HKUST

1. Introduction

In this final project, I try to evaluate the capability of HodgeRank on the human age ranking by evaluating pairwise comparison data. When we seek to construct a complete subjective assessment on videos/photos/events by crowdsourcing, pairwise comparison is a relatively comfort way for annotators. Pairwise comparison doesn't required the annotators to present their opinions on every photos, which seems complicated for them. However, as the number of samples increases, the difficulty of direct comparison is exponentially increasing. The volunteers will definitely lose their patience towards the impossible mission by comparing photos one by one. HodgeRank on Random Graphs (HRRG) is proved as a reliable and efficient framework for pairwise comparison data analysis on VQA(video quality assessment), which I tend to incorporate on the paired comparison on age ranking.

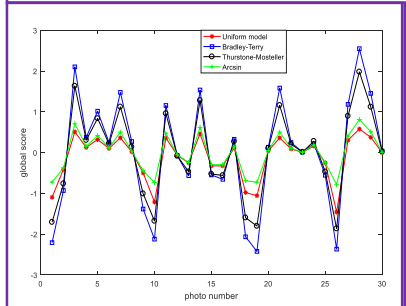
2. Data Description

The data I use in this project is Agedata.mat. This dataset includes 30 images from human age dataset which are annotated by a group of volunteer users on ChinaCrowds platform. The annotators are shown with two portrait photos and given his choice of which one is older (or difficult to judge). Eventually, 14,011 pairwise comparisons from 94 annotators are collected. The last column of the dataset shows the annotator's choice: 1 indicates the second column is older than the third one; -1 indicates the second column is younger than the third one; 0 indicates the second and third are difficult to judge.

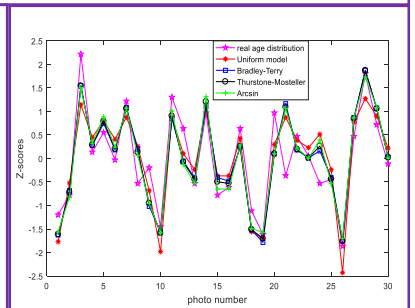
In this project, I didn't consider the impact of "hard to judge" option into the model, thus, after the data rearrangement, there are 12778 sets of pairwise comparison left. Next section, I will illustrate the global score for each photo generated by four generalized linear models (Uniform model, Bradley-Terry model, Thurstone-Mosteller model, and Angular transform model.)

3. Global Score and Z-score

The figure 1 below shows the global scores generated by four different generalized linear models. As the plot illustrates, Bradley-Terry model has a similar trend as Thurstone-Mosteller model while uniform model's trend is likely to angular transform model's.
The second figure shows the comparison between real age and ranking score after standardization. As we can see the trend is basically fits the real age.



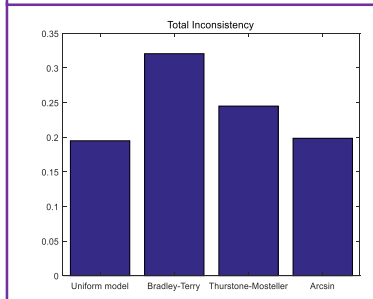
Global scores generated by HodgeRank via four different GLM.

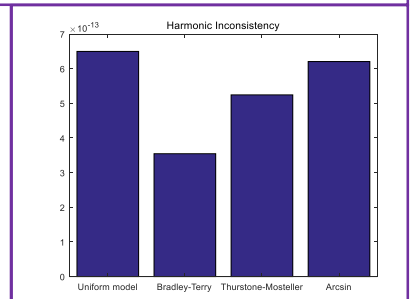


The comparison between real age and ranking score by standardization.

4. Inconsistency Analysis

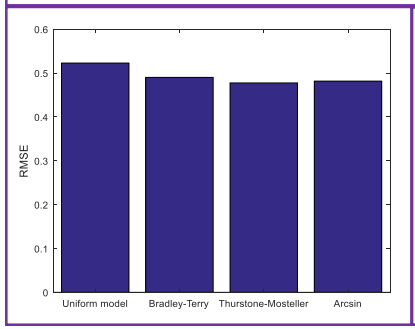
Inconsistency shows the reliability of the annotators. In this dataset, since each annotators evaluate different pairs of photo, we just check the inconsistency for different GLMs. As we can see, Brad-Terry model has the highest total inconsistency and uniform model has the lowest. Harmonic inconsistency are small for each models.

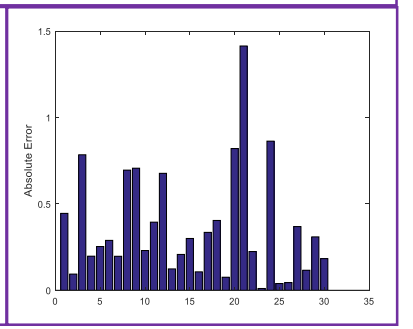




5. Error Analysis

Since I make the comparison between real age distribution and generated ranking, I also want to evaluate the accuracy of people's judgement by investigating the RMSE. The left figure shows that each model has similar RMSE which is around 0.5 and Thurstone-Mosteller model has the lowest which close to 0.48.
To analyze the reason, I try to check the absolute error between real age z score and the mean ranking z score. The right figure shows that the No.21, No.24, No.20 and No.3 is relatively hard to recognize their real age.





6. Conclusion

Through our experiment, four GLMs are tested and their inconsistencies are evaluated. The conclusion for this report is that, HHRG is a suitable framework for managing the incomplete pairwise comparison data on the human age ranking. Based on the Error analysis and Inconsistency analysis, I am convinced that Thurstone-Mosteller model is the most appropriate GLM on human age ranking.

7. References

Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. HodgeRank on Random Graphs for Subjective Video Quality Assessment. IEEE Transactions on Multimedia, 14(3):844-857, 2012, https://github.com/yao-lab/yao-lab.github.io/blob/master/reference/TMM12-final.pdf)
Yao, Yuan. "A Mathematical Introduction to Data Science." (2019)