Reproduction for the methodology of “Ranking Inferences Based on the Top Choice of Multi-way Comparisons”

1. Real Data Analysis

Data

1.1. Data Description

Raw Data: 4.1 million continuous ratings of 100 jokes from 73421 users: collected between Apirl 1999-May 2003.

1.2. Data Availability

The original Dataset and detailed description can be found in https://goldberg.berkeley.edu/jester-data/ (https://goldberg.berkeley.edu/jester-data/).

In our .zip file, the original Dataset is provided in the folder `/Ranking_Inference_Reproduction/Real_Data/Original_Raw_Data/jester-data-1.csv` and `/Ranking_Inference_Reproduction/Real_Data/Original_Raw_Data/jester-data-2.csv` .

Description: Data files contain anonymous ratings data from 73,421 users. Ratings are real values ranging from -10.00 to +10.00 (the value “99” corresponds to “null” = “not rated”).

One row per user.

The first column gives the number of jokes rated by that user. The next 100 columns give the ratings for jokes 01 - 100.

1.3 Data Pre-Processing.

The pre-processing method is quite straightforward where we pick those users who rate whole 100 jokes in the raw dataset according to the first column and delete the first column (because now all of them rate 100 jokes). Therefore, since every user rates all jokes, there is no missing value “99” in our processed dataset. The code for pre-processing is given in: `/Ranking_Inference_Reproduction/Real_Data/Original_Raw_Data/Pre-Processing.R`

The cleaned datasets are put in `/Ranking_Inference_Reproduction/Real_Data/jester_1.csv` and `/Ranking_Inference_Reproduction/Real_Data/jester_2.csv` . The following implementations are all based on the processed datasets.

Codes Description and Implementation Details:

There are two files under the folder: `Ranking_Inference_Reproduction/Real_Data/`

1. Table4.R,
2. Table5-6.R

Folder /Ranking_Inference_Reproduction/Real_Data/:

It contains the codes for reproduce results in Table 4 and Table 5-6. The documents `jester_1.csv` and `jester_2.csv` in that folder are processed data by choosing all users that rate all 100 jokes.

The code file **Table4.R** contains the codes for reproducing Table 4 of the real data analysis. The setting is given in lines 11-14 (corresponds to the setting of Table 4), one only needs to run this code file directly. The outputs are given in line 113 and lines 188-194. We also mark their corresponding reproduced columns in Table 4.

Run time of **Table4.R** is less than 10 min.

The code file **Table5-6.R** contains the codes for reproducing Tables 5 and 6 of the real data analysis. The setting of is given in lines 14-17, one only need to set up these parameters according to the different settings of Table 5 and 6 in our paper and run this code file directly. The outputs are given in lines 96-103. We also mark their corresponding reproduced columns in Table 5-6.

Run time of **Table5-6.R** is less than 10 min.

2. Simulation Studies

Codes Description and Implementation Details:

1. File /Ranking_Inference_Reproduction/Simulation/Figure1-2.R:

This is the code file `Figure1-2.R` contains codes for reproducing the methodology of Figures 1 and 2. The setting of the code is given in lines 2-5. Here we let L be fixed and p vary. (Following the caption of Figure 2, we can also let p fixed and let L vary).

The output of this simulation is given in lines 34-37, where the l_2 and l_∞ errors are recorded for 500 times. The mean value and standard errors can be computed via these outputs directly.

Run time of **Figure1-2.R** for any given setting described above is around 8 hrs.

2. File /Ranking_Inference_Reproduction/Simulation/Figure_3.R:

This is the code file `Figure_3.R` contains codes for reproducing the methodology of right panel of Figure 3.

The setting of the code is given in lines 2-4. Here we follow the setting of figure 3 in the codes (n=60,p=0.05,L=80). The output of this simulation is given in line 38, where the empirical p-values are recorded. With these data, one is able to reproduce the right_panel of figure 3.

Run time of **Figure_3.R** for any given setting described above is greater than 24 hrs.

3. File /Ranking_Inference_Reproduction/Simulation/Figure_7.R:

This is the code file `Figure_7.R` contains codes for reproducing the methodology of Figure 7 in appendix (also left panel of Figure 3).

The setting of the code is given in lines 2-5. Here we follow the setting of figure 7 by letting L=10 and p=(0.008,0.015,0.03) respectively. One is also able to set L by different numbers (e.g. 5, 10, 20 suggested in figure 7). The output of this simulation is given in line 33, where we record the distribution of the MLE. With these data, one is able to reproduce the right_panel of figure 7 and left panel of Figure 3.

Run time of **Figure_7.R** for any given setting described above is around 2 hrs.

4. File /Ranking_Inference_Reproduction/Simulation/Table1-3.R:

This is the code file `Table1-3.R` contains codes for reproducing the methodology of Tables 1-3.

The setting of the code is given in lines 3. The output of this simulation is given in lines 255-297, where we specify the meanings of all outputs and their corresponding position in Tables 1-3.

Run time of **Table1-3.R** for any given setting described above is around 6 hrs.