Modelling Partial Ranking Data

Tita Vanichbuncha

Modelling Partial Ranking Data

School of Mathematics, Statistics and Actuarial Science

University of Kent

September 2017

Abstract

Ranking is one of the most used methods in not only in statistics but also in other field such as computer science and psychology. This method helps us determine order of objects in a group such as preference of animal species, and has very broad applications. However, when the number of objects to be ranked becomes larger, the uncertainty of the ranking typically increases since it is harder for the ranker to express their preference accurately. This leads to the idea of partial ranking which allows rankers to rank just a subset of objects in the group and then combine their results together to form the global ranking. This thesis focuses on this type of data. The main challenge is how to accurately analyze partially ranked data and decide the global ranking. There are several models that address this kind of problem such as the Bradley-Terry (BT) model and the Plackett-Luce (PL) model.

The BT model is for paired comparisons while the PL model is for any number of ranked objects. The PL model is slow to fit using existing R packages. We implement the algorithms in R and do empirical studies using simulated data. The results show that our algorithms perform faster than the existing packages in R. We also implement R code for computing the observed information matrix. Rank-breaking methods are also considered in order to be able to use the BT model with different weightings instead of using the PL model. We examine the performance of various weightings by experimental studies with the simulated data and with real-world data. Our BTw-Sqrt weighting performs best when the number of rankers is small.

In order to choose subsets of objects to be ranked, we consider three existing criteria which are D-optimality, E-optimality, and Wald and we propose three new methods. Experiments have been done using simulated data and the results compared with random selection. Our result shows that the existing criteria sometimes perform better than random selection. Our proposed methods usually ensure that the PL model can be fitted to data from fewer rankers than random selection.

We describe two extensions of the PL model, the Rank-Ordered Logit (ROL) model and the Benter model. The ROL model extends the PL model by allowing covariates to be incorporated and the Benter model allows preferences for higher-ranked items to be stronger than for lower-ranked items. Both extensions improve the fit of the PL model to an example dataset when using the Likelihood Ratio (LR) test to compare models. We combine these two extensions to give a model that incorporates covariates and allows for a dampening effect. The combined model further improves the fit to our example data when compared with the ROL model by using LR test. We implement R codes for analyzing and computing the observed information matrices of the ROL, Benter, and combined models.

We also explore another type of partial ranking data where individuals are allowed to mention any objects rather than being given a predefined list of objects to rank. We consider the idea of Participatory Risk Mapping (PRM) which provides severity and incidence scores. The severity and incidence scores can be modelled using the PL model and a new proposed model, respectively.

Acknowledgements

My grateful thanks are due to my supervisor, Professor Martin S. Ridout, for his help, expert guidance, and patience throughout my Ph.D.

I would also like to express my appreciation for Dr Fabrizio Leisen and Professor Elizabeth Mansfield for their helpful comments and suggestions.

My grateful thanks due to Chulalongkorn University for funding my Ph.D.

Thanks to all the staffs in the School of Mathematics, Statistics and Actuarial Science for being supportive. My sincere thanks go to all my friends for their support, especially, Anita for always being there to help me.

Finally, my sincere thanks are due to my family for their encouragement and care. For my brother, Gunn and my friend, Pakin for reading and checking my writing.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Ranking is one of the fundamental methods of data collection that is used in many areas such as social choice (Caplin and Nalebuff, 1991; Soufiani and Parkes, 2014), information retrieval (Cohen et al., 1999; Dwork et al., 2001), voting and elections (Diaconis, 1988; Koop and Poirier, 1994; Gormley and Murphy, 2008), market research (Beggs et al., 1981), and psychology (Maydeu Olivares and Bockenholt, 2005). In this thesis we focus on ranking of objects. The rankings of objects are very common in everyday life e.g. horse-racing competitions for gamblers, in business, companies want to know customers' preference on products, election systems, etc. The term *preference ranking* refers to information generated by humans who rank a given set of objects, or rank the set of object that they have in their mind, based on their own preferences according to a specific objective. We are interested to know the overall preferences of a population of individuals, which is sometimes called the *social choice* problem. Generally speaking, social choice addresses the problem of choosing an object or a decision from a set of objects for a group of individuals.

There are many ways of making comparisons between objects, including ranking, top-h ranking, discrete choice, maxDiff, and rating. The general meaning of ranking is that individuals rank objects from most preferred to

least preferred. When individuals rank only the first h rank positions, it is called top-h ranking (Ailon, 2010). If individuals only choose their single most preferred object, this is called discrete choice (Train, 2003). The MaxDiff method asks each individual to pick his/her most preferred and least preferred objects (Marley and Louviere, 2005). Finally, in the rating method, each individual is asked to give a numerical score to each object. In this thesis, we mainly focus on the *ranking* method.

1.1 A Brief History of Preference Ranking

There are a number of historical examples showing that people's options are complex in social choice (McLean et al., 1995). Dating back to eighteenth century, the study of the social choice problem was introduced by Borda (1781), a French engineer, philosopher, mathematician, and political scientist. In the year 1781, Borda was interested in a political voting system in which each voter ranks all candidates from most to least preferred. Borda introduced a method for analyzing the voting which became known as *Borda count*. The candidates are assigned scores according to their rank positions in the election, where one indicates most preferred and the least preferred candidate receives a score that is equal to the number of candidates in that particular ranking. The candidate who gets the minimum total score is the winner. This system leads to a family of voting rules. Not long after that, Condorcet (1785), who was also interested in elections, argued against Borda's rule. He proposed instead the *Condorcet Winner* concept. This is based on pairwise comparisons where the winning candidate is the one who gets a majority of voters support among all pair comparisons. For example, if there are 6, 5, and 4 rankers who give $(A \succ B \succ C)$, $(B \succ C \succ A)$, and $(C \succ B \succ A)$, respectively where $\succ$ means preferred to. The pairs $\{A, B\}$, $\{A, C\}$, and $\{B, C\}$ are considered. For $\{A, B\}$, A is preferred to B 6 times and B is preferred to A 9 times.

We drop the pairs with A out since A cannot be a Condorcet winner. Next, $\{B, C\}$, since $(B \succ C)$ 11 times and $(C \succ B)$ 4 times, B is selected in this pair. Thus, B is the Condorcet winner since B is chosen in all pairs. However, he observed that there can be a paradox in the ranking, which became known as the Condorcet paradox. This states that, when there are at least three candidates, it is possible that the majority of voters prefer A over B, B over C, and C over A. In other words, the majority preference relation turns out to be cyclic. Hence, Condorcet's proposal does not always lead to a clear outcome.

In the early twentieth century, probability models for ranking data were introduced. Thurstone (1927) proposed his *law of comparative judgement*, which models the comparison of perceived intensities of physical stimuli. Experiments showed that the same individual may give different rankings on different occasions. Thurstone introduced randomness and modelled the perceived intensity of each stimulus of taking this issue into account. This model is based on the normal distribution and is called the Thurstonian order statistics model. Luce and Suppes (1965) provided a review of the experimental validation of Thurstonian model. Thurstone also proposed a model for pairwise comparisons.

Later, Arrow (1951) published a paper about what become known as *Arrow's impossibility theorem*. Arrow studied the problem of preference aggregation and his theorem illustrated the impossibility of having an ideal voting system that satisfied reasonable fairness criteria. The theorem states that there is no clear order of preferences to be determined. Arrow illustrated the difficulty in using this kind of information in social choice and economics.

Many probability models were proposed in the twentieth century after the Thurstonian model. The Bradley-Terry (BT) model (Bradley and Terry, 1952) is a widely used model for paired comparisons. The idea is that each object is assigned a latent value and the probability that one object is preferred to an-

other depends on the difference in their latent values, where objects with high latent value likely to be preferred. In 1957, Mallows (1957) proposed a general distance-based model. The Mallows' model assumes that there is a modal ranking and that the probability of a ranking decreases as its distance from the modal ranking increases. Special cases of this model are called Mallows' ϕ model and Mallows' θ model, when using Kendall distance and Spearman distance, respectively.

Luce (1959) introduced an alternative way of analyzing ranking sets which was an axiomatic approach to choice modeling. Luce proposed what is now called the *Luce choice axiom* (LCA) and this led to the development of multistage models e.g. models from Luce (1959), Plackett (1975), Henery (1981), and Fligner and Verducci (1986). The Plackett-Luce (PL) model belongs to the class of multistage models (Plackett, 1975). The PL model can be viewed as an extension of the BT model. The relationship between Luce's (1959) model and the Thurstonian model was established by Yellott (1977). Yellott (1977) showed that the Luce model satisfies LCA and Thurstonian's comparative law with independent random variables.

Many other models have been proposed based on different approaches. Critchlow et al. (1991) divided these models into four classes, which are order statistics models, paired comparisons models, distance-based models, and multistage models. A brief review is provided in Chapter 2.

In this thesis, we are interested mainly in the BT and the PL models. Our main objective is to learn and exploit these models for finding the global preference.

1.2 Why ranking?

We believe that a hidden true preference underlies the ranker's choices such as ratings, ranking list. A ranking refers to a rank-ordered list of objects while

a rating refers to a list of scores e.g. assign between 1 to 10 points to a movie on http://www.imdb.com.

A common method of collecting rating data is the use of Likert scales in which respondents record their level of preference on a predefined number of scale points. Such scales are often unreliable since the interpretation varies from ranker to ranker, for instance a rating of "4" may not have the same meaning for every ranker. Therefore, the rating method often provides unstable and inconsistent preference information (Peng et al., 1997).

Conversely, the ranking method does not have this problem. The information from ranking method is more absolute e.g. if two rankers ranked A higher than B, this means they both prefer A to B; however, this does not give an information about how much they like A more than B.

In term of achieving a global ranking, the rating method can simply use average scores, while the ranking method has some difficulty in this stage and ranking method may require more complex computation.

1.3 Thesis Outline

This thesis is organized as follows.

Chapter 2 provides an overview of background information related to ranking, including types of ranking data and parametric models. There are three different real world datasets involved in this thesis. All of them involve partial ranking data. Descriptive statistics for these datasets are provided in this chapter. The first dataset is the Animal dataset, which has four groups of data. This data was collected by giving the same number of images to all individuals to rank them, where the images are randomly selected for each individual. Information about individuals and images is presented. The second dataset, the Sushi dataset, is similar to the Animal dataset. However, sushi flavours are not randomly assigned to the individual. The final dataset is the

Sundarbans dataset. The Sundarbans dataset is different from the previous datasets. An individual was asked an open-ended question and the individual mentioned all choices that they thought were important and then ranked these choices.

Chapter 3 describes two models for analyzing ranking data, the BT model and the PL model. We present a necessary axiom, the Luce Choice Axiom. Algorithms for fitting the BT and the PL models are implemented and their performances are compared with existing algorithms. The PL model is applied to the Group I data from the Animal dataset for illustration. A bootstrap goodness-of-fit test is performed in order to test whether this data can be fitted by the PL model. We then explore rank-breaking methods. Three rank-breaking methods are compared using simulated data. The results show that the full rank-breaking method is the best among them. We study further the full rank-breaking method. Different weights for the BT model are introduced in order to compare and improve the performance of equal weighting. We apply non-weighting and weighting to full breaking pairs from simulated data, the Sushi dataset, and the Group I data from the Animal dataset.

Chapter 4 shows that we can roughly estimate the logarithm of the observed information matrix of the estimates from the PL model by using a regression model. This gives us insight into the loss of information that occurs when individuals rank only a subset of the items of interest. In this chapter, we choose a subset of objects that maximize the gain in expected information. Two criteria, D-optimality and E-optimality, which are adopted from the experimental design framework, are considered. Another criterion is the Wald criterion. We compare these criteria with random selection. Empirical results are presented using synthetic data with a small number of items and a large number of items. We propose three systematic methods. We apply these three methods to simulated data with large numbers of items. We compare the three statistical criteria with the three proposed methods and discuss results.

Chapter 5 extends the PL model from the third chapter in two ways. The rank-order logit (ROL) model can incorporate covariates and the Benter model allows dampening parameters. Furthermore, we provide a model that combines the ROL model and the Benter model. We apply these models to the Animal dataset. The results are presented and tests are run to determine whether these models are better than the PL model. We perform bootstrap goodness-of-fit tests to investigate whether the models fit the Animal dataset well.

Chapter 6 explores the analysis of open-ended rankings. In the previous chapters, we use datasets in which researchers give a specified subset of items to an individual to rank. In this chapter, we introduce another type and ranking data where an individual has to identify his/her own list of items. Participatory Risk Mapping (PRM) is an established tool for analyzing this kind of data. We explain the PRM at the beginning of this chapter. The Sundarbans dataset comes from the open-ended questionnaire and we use this dataset here. Moreover, tied rankings are allowed in this dataset. We consider two approximation methods, which are Breslow and random, to handle the ties. After that we explore the number of mentioned objects. Logistic regression is considered in order to estimate these numbers. We propose a new model to analyze the open-ended rankings. The PL, ROL, and the proposed models are applied to the Sundarbans dataset. Results are discussed.

Chapter 7, the final chapter, offers some contributions and ideas for future work.

Chapter 2

Preliminaries

In this chapter, we describe several things that we are going to use later on in this thesis. Notations are defined in Section 2.1. Three types of ranking data are described in Section 2.2. To evaluate performance of models, Kendall tau correlation is used widely in the literature on ranking. As this is less familiar than the Pearson and Spearman correlation coefficients, it is explained briefly in Section 2.3. A goodness-of-fit test is introduced in Section 2.4 that can be used to assess whether a model fits well with the data. Section 2.5 provides a brief review of some existing probability models for ranking data. Sections 2.6 to 2.8 give details of real-world datasets which are used in this thesis, the Animal dataset, the Sushi dataset, and the Sundarbans dataset.

2.1 Notation

Suppose that n rankers participate in a survey in which there is a total of K items to be ranked denoted by $\mathcal{O} = \{1, 2, \ldots, K\}$. If K is large, it is usually impractical for all items to be ranked by all rankers. Let p_i denote the number of items ranked by ranker i, where $p_i \leq K$. Additionally, p is used instead of p_i when all rankers rank the same number of items. Let ρ_{ij} denote the item which was ranked in j^{th} position by ranker i, $j = 1, \ldots, p_i$.

2.2 Types of Ranking Data

There are many different types of ranking data. Here, we define ranking types where ties are not allowed. When ties are allowed that means more than one item can be ranked at the same preference. Mainly, we do not consider data with ties in this thesis, except in Chapter 6 where tied rankings are involved. Three types of ranking data are considered in this thesis, full ranking, partial ranking, and top-h ranking data.

2.2.1 Full Ranking

A full ranking has all K items ranked. A ranker assigns a complete ordering to the items and the observed ordering is denoted by $(\rho_{i1} \succ \rho_{i2} \succ \cdots \succ \rho_{iK})$, where $\succ$ denotes 'is preferred to'. Thus, there is the most preferred item, second most preferred item, ..., and the least preferred item for all items.

2.2.2 Partial Ranking

A partial ranking provides a full ranking of a subset $\mathcal{O}' \subsetneq \mathcal{O}$ of items, where $\mathcal{O}'$ contains at least two items. The ordering is $(\rho_{i1} \succ \rho_{i2} \succ \cdots \succ \rho_{iK'})$ where K' is the number of items in the set $\mathcal{O}'$ and $K' < K$.

This kind of data occurs when the total number of items is too large and/or it is too costly for rankers to undertake a full ranking. This commonly occurs in sports tournaments such as car and horse racing data where only a subset of the racers is compared in each race. In the area of item preference, it may be unreasonable to ask rankers to rank the full set of K items. When the ranker ranks too many things, the quality of judgements will decline. Thus, we may obtain more reliable rankings if each ranker is asked to rank only a subset of items. Miller (1955) suggested that number of objects to be judged/ranked should be no more than seven, because we get more inconsistency in the ranking list when the number exceeds seven.

2.2.3 Top Ranking

A top ranking provides a full ranking of a subset $\mathcal{O}' \subsetneq \mathcal{O}$ of items and the additional information that all items in $\mathcal{O}'$ are preferred over the items in $\bar{\mathcal{O}}'$ where $\bar{\mathcal{O}}' = \mathcal{O} \backslash \mathcal{O}'$. There is no preference information for the items in the set $\bar{\mathcal{O}}'$. An extreme case is when an individual chooses only the single most preferred item.

The observed ordering is $(\rho_{i1} \succ \rho_{i2} \succ \cdots \succ \rho_{iK'})$ where K' is the number of items in the set $\mathcal{O}'$ and $K' < K$. All items in the set $\mathcal{O}'$ are fully ranked. Moreover, $\mathcal{O}'$ is preferred to the remaining items in $\bar{\mathcal{O}}'$. Irish elections provide one example of this type of data. The voters rank their preferred candidates in order of preference but may leave some candidates unranked.

2.3 Kendall Tau Correlation

Two popular methods to measure correlation between two variables are the Spearman rho and Kendall tau correlation coefficients. Another well-known correlation is the Pearson correlation; however, this correlation measures the strength of linear relationship between two continuous variables (Khamis, 2008). The strength of relationship is the strength of tendency of the two variables to move in the same or opposite direction. Thus, Pearson correlation is not suitable to measure the association of two rankings. The Kendall tau is preferred to the Spearman rho in term of robustness and efficiency (Croux and Dehon, 2010). Morever, the Kendall tau correlation is preferred due to simplicity and direct interpretation (Kendall and Gibbons, 1990).

The Kendall tau correlation was developed by Kendall (1938) and it determines the correlation between two rankings of equal size based on the number of pairwise swaps of adjacent items needed to transform one ranking into another ranking. This is termed the Kendall distance and is denoted by d'_τ. The maximum number of swaps is $\frac{1}{2}K(K-1)$. Then the Kendall tau correlation

is given by normalizing the distance, multiplying by 2 and subtracting from 1,

$$\rho_\tau = 1 - \frac{2d'_\tau}{\frac{1}{2}K(K-1)},$$

where the denominator is the total number of pairs of K items in the ranking; thus, $-1 \leq \rho_\tau \leq 1$. If $\rho_\tau = 1$, the two rankings are the same and if $\rho_\tau = -1$, one ranking is the reverse of the other ranking. For example, suppose there are four items ($K = 4$) and let $\mathcal{O} = \{A, B, C, D\}$, and consider the two rankings $\mathcal{S}_1 = \{A \succ D \succ C \succ B\}$, and $\mathcal{S}_2 = \{A \succ C \succ D \succ B\}$. Thus, $d'_\tau = 1$ and $\rho_\tau = \frac{2}{3}$. The normalized Kendall tau distance is denoted by d_τ and can be given as follows

$$\begin{aligned} d_\tau &= \frac{d'_\tau}{\frac{1}{2}K(K-1)} \\ &= \frac{(1-\rho_\tau)}{2}, \end{aligned}$$

where d_τ lies in the interval $[0, 1]$. A value of 0 means the two rankings are exactly the same and a value of 1 indicates maximum disagreement. This makes it easier to compare the rankings.

An equivalent and more common expression for the Kendall tau correlation is given by introducing concordant and discordant pairs. A pair is concordant if the relative ranking of the two items is the same in both ranked lists. From the previous example, A is ranked above B in both rankings above and the pair (A, B) is therefore concordant. A discordant pair is when an item is ranked above another item in one list, but below it in the other list. In the example above the pair (C, D) is discordant. Let n_c denote the number of concordant pairs and n_d denote the number of discordant pairs. The Kendall tau correlation is

$$\rho_\tau = \frac{n_c - n_d}{\frac{1}{2}K(K-1)}.$$

Using the previous example, then we can calculate n_c and n_d as shown in Table 2.1. The n_c and n_d are 5 and 1 pairs, respectively. The Kendall tau correlation between $\mathcal{S}_1$ and $\mathcal{S}_2$ is $\frac{4}{\frac{1}{2}(4)(3)} = \frac{2}{3}$, as before.

Table 2.1: Example for calculating the Kendall tau correlation

Item	Ranker 1	Ranker 2	n_c	n_d
A	1	1	3	0
D	2	3	1	1
C	3	2	1	0
B	4	4		
		Total	5	1

2.4 Goodness-of-Fit Test

It is problematic to test the goodness of fit of ranking models to data because not all possible patterns are observed. The classical approaches such as Pearson χ^2 and likelihood-ratio are not suitable. The Bootstrap is an alternative approach to assess statistical accuracy. The bootstrap is suggested for use with sparse categorical data (von Davier, 1997).

The idea is to simulate data according to the model using the estimated parameters from fitting the model to the original data. The model is re-fitted to the simulated data and this process is repeated B times, where B is number of bootstrap samples. We examine the behaviour of the fits over the B bootstrap samples.

Let T denote a goodness-of-fit statistic and let t be the value of this statistic when it is calculated from the original data. We can approximate the distribution of T by generating a sample of independent outcomes t_b^* for $b = 1, \ldots, B$ and constructing the empirical distribution $\hat{F}_{t^*}$. In our case, we use two test statistics, the mean Kendall tau distance and the IOS statistic as t value. The abbreviation IOS comes from "in-and-out-of-sample" (Presnell and Boos,

2004). If t is not significantly different from the bootstrap sample t^*, it means the model is an appropriate model for fitting the original data.

The mean Kendall tau distance is calculated as the mean over all rankers of the distance between the ranking produced by the ranker and the ranking of the items expected on the basis of the estimated parameters.

Let $Y_1, \ldots, Y_n$ be independent and identically distributed and $\boldsymbol{\lambda}$ be a parameter vector, and let

$$\mathbf{I}(\boldsymbol{\lambda}) = E[-\ell''(Y_1; \boldsymbol{\lambda})]$$
$$\mathbf{B}(\boldsymbol{\lambda}) = E[\ell'(Y_1; \boldsymbol{\lambda})\ \ell'(Y_1; \boldsymbol{\lambda})^\intercal],$$

where $\ell'(Y_1; \boldsymbol{\lambda})$ is the gradient vector with respect to the elements of $\boldsymbol{\lambda}$ and $\ell''(Y_1; \boldsymbol{\lambda})$ is the Hessian matrix. The $\mathbf{I}(\boldsymbol{\lambda})$ is the information matrix and the $\mathbf{B}(\boldsymbol{\lambda})$ is another way of defining the information matrix. The IOS statistic is

$$\begin{aligned} IOS &= E[\ell'(Y_1; \hat{\boldsymbol{\lambda}})^\intercal\ \mathbf{I}(\hat{\boldsymbol{\lambda}})^{-1}\ \ell'(Y_1; \hat{\boldsymbol{\lambda}})] \\ &= \mathrm{tr}[\mathbf{I}(\hat{\boldsymbol{\lambda}})^{-1}\ \mathbf{B}(\hat{\boldsymbol{\lambda}})], \end{aligned}$$

where $\mathrm{tr}(A)$ denotes the trace of a matrix A. The IOS is the ratio of $\mathbf{B}(\hat{\boldsymbol{\lambda}})$ and $\mathbf{I}(\hat{\boldsymbol{\lambda}})$. If $\mathbf{B}(\hat{\boldsymbol{\lambda}})$ and $\mathbf{I}(\hat{\boldsymbol{\lambda}})$ are equivalent, the trace of $\mathbf{I}(\hat{\boldsymbol{\lambda}})^{-1}\mathbf{B}(\hat{\boldsymbol{\lambda}})$ is the number of parameters. Thus, the IOS statistic tends to the number of parameters (K) as the numbers of items under the null hypothesis of correct model specification.

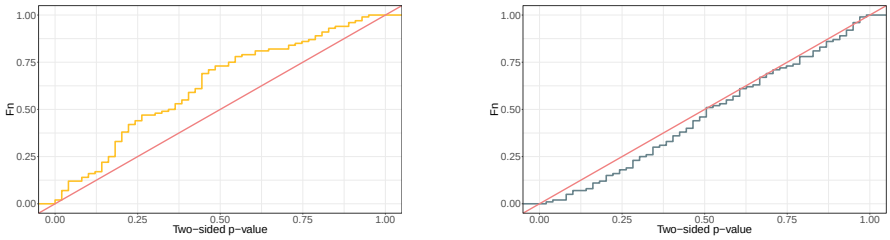
We compute a two-sided p-value based on how far the value of the mean Kendall tau distance lies in the tails of the bootstrap distribution where the null hypothesis is that the model is suitable for fitting the data. For the IOS statistic, if the IOS value approaches K then an one-sided p-value is calculated. The one-sided test looks only the upper tail. However, a two-sided test is suggested instead of the one-sided test (Capanu and Presnell, 2008)

when the IOS value approaches zero. The two-sided p-value is

$$\text{p-value}_{\text{2-sided}} = 2 \times \min(\text{p-value}_{\text{1-sided}}, 1 - \text{p-value}_{\text{1-sided}}).$$

von Davier (1997) suggested that in order to estimate the distribution of T, B has to be very large. However, if the bootstrap is aimed for testing, B can be relatively small.

To illustrate the procedure, we generate data under the Plackett-Luce model with $K = 100$ and $p = 10$ where the true parameter values are generated from a uniform distribution. The bootstrap is performed for 99 times ($B = 99$). We repeat this process 100 times. The IOS statistics from the bootstrap approach zero. This suggests that we should compute the two-sided test instead of the one-sided test.

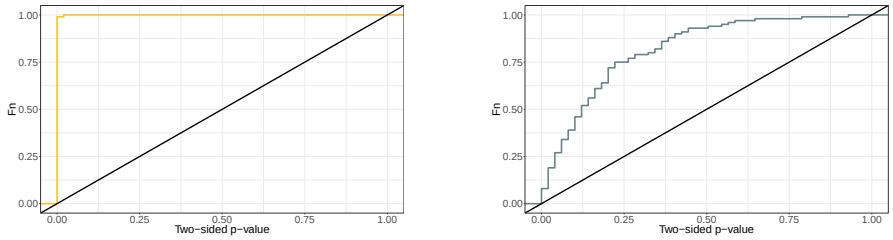


(a) Kendall tau distance (b) IOS

Figure 2.1: Cumulative distribution of the two-sided p-values of the Kendall tau distance and IOS tests from the bootstrap goodness-of-fit for the PL model

The cumulative distributions of the two-sided p-values from the Kendall tau distance and the IOS test are shown in Figure 2.1. If the p-values have an approximate uniform distribution, it means the null hypothesis is true (Murdoch et al., 2008). We conclude that there is no evidence against the PL model. Figure 2.1a shows that the p-values from the Kendall tau distance do not have a uniform distribution. Figure 2.1b presents that the p-values from the IOS test have an approximate uniform distribution. We conclude that the IOS test is more suitable than the Kendall tau distance for assessing the goodness-of-fit of the ranking data.

Next, we would like to confirm that our procedure can detect model failure. We use the same setting as previously. However, instead of generating data under the PL model, we randomly generate data. The two-sided p-values from the Kendall tau distance and the IOS test are shown in Figure 2.2.



(a) Kendall tau distance (b) IOS

Figure 2.2: Cumulative distribution of the two-sided p-values of the Kendall tau distance and IOS tests from the bootstrap goodness-of-fit for the PL model when the data is randomly generated

Both Figure 2.2a and Figure 2.2b show that the p-values of the Kendall tau distance and the IOS test, respectively, do not have uniform distribution. This means the PL model is not a suitable model. In conclusion, the Kendall tau distance and the IOS statistc can be used to detect the model failure.

2.5 A Brief Survey of Probability Models for Ranking Data

Critchlow et al. (1991) broadly categorized probability models on rankings into four classes: (1) Thurstonian order statistics models, (2) paired comparison models, (3) distance-based models, and (4) multistage models. Marden (1995) also categorized the models in the same way. In this section, we briefly introduce these four classes of models.

2.5.1 Thurstonian Order Statistics Models

The class of order statistics models has the longest history in the statistical and psychological literature among the four classes of probability models. The Thurstonian model is one of the oldest and best-known order statistics models. Thurstone (1927) proposed the Law of Comparative Judgement to model paired comparisons data and later on the Thurstonian model was proposed as a scaling method for ranking data (Thurstone, 1931). The latter model assumes that an observer ranks items by ranking unobserved continuous response variables representing the observer's psychological perception of each item. A Thurstonian model ranking derives the probability of a given ranking on the basis of the distribution of these K latent response variables $Z_{i1}, Z_{i2}, \ldots, Z_{iK}$ that depend on the ranker i (Marden, 1995). For example, the items A, B, and C are ranked in the order $A \succ B \succ C$ if and only if $Z_A > Z_B > Z_C$.

Thurstone assumed that these latent response variables have a K-dimensional multivariate normal distribution, $N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with K means, K variances, and $\binom{K}{2}$ correlations. Simpler forms arise by setting the correlations to be equal, the variances to be equal, and/or setting the correlations to zero so that the Z_i are independent. The most popular simplification is the so-called Case V model which assumes that the latent variables are uncorrelated with equal variance.

As alternatives to the normal distribution, Luce (1959) used the Gumbel distribution, which leads to the Plackett-Luce model that is discussed extensively in this thesis, and Henery (1983) and Stern (1990) used Gamma distributions.

2.5.2 Paired Comparison Models

The paired comparison models aim to combine models for paired comparisons to generate a probabilistic model for ranking data. For K items, there

are $\frac{K(K-1)}{2}$ possible comparisons. Babington-Smith (1950) introduced the Babington-Smith model. This model assumed that the ranking has come from a set of $\frac{K(K-1)}{2}$ arbitrary paired comparison probabilities, p_{ab}. The p_{ab} is the probability that item a is preferred to item b where $a < b$. Moreover, $p_{ab} = 1 - p_{ba}$ if ties are not allowed. Let $\pi_i(a)$ be the rank assigned to item a by ranker i then the probability of a ranking π_i is

$$P(\pi_i) = C \prod_{(a,b):\pi_i(a)<\pi_i(b)} p_{ab},$$

where C is a constant to make the probabilities sum to 1. This model assumes that the pairwise comparisons are independent.

Later, Mallows (1957) introduced four simple subclasses of the Babington-Smith model. One of them is the Bradley-Terry model (described in Chapter 3) and the other models incorporate a distance function. We describe only two of these three models in this thesis. The most general model among the three models, the Mallows two-parameter model, is described below.

The Mallows two-parameter model assumes that rankings which have the same distance from a modal ranking π_0, should have the same probability. The Mallows two-parameter model is given by

$$P(\pi_i) = C(\theta, \phi)\, \theta^{d_S(\pi_i,\pi_0)}\, \phi^{d_K(\pi_i,\pi_0)}, \tag{2.1}$$

where $\theta, \phi \in (0, 1)$ and $C(\theta, \phi)$ is a constant to make the probabilities sum to 1. The d_S and d_K are the Spearman and Kendall distances between π and π_0, respectively.

If $\theta = 1$ in Equation (2.1), this yields Mallows ϕ-model,

$$P(\pi_i) = C(\phi)\, \phi^{d_K(\pi_i,\pi_0)}, \tag{2.2}$$

where $0 < \phi \leq 1$. The ranking probability decreases according to increasing

Kendall distance from π_i to π_0. This model also belongs to the class of distance-based models which are described in the next section.

2.5.3 Distance-based Models

Distance-based models use distance functions to measure the discrepancy between two rankings. The probability of an observed ranking is inversely proportional to a distance between the observed ranking, π, and a modal ranking, π_0. The models assume that a modal ranking exists. The advantages of this type of model are simplicity and elegance. However, there are two major weaknesses which are (1) difficulties in incorporating covariates, and (2) the model has only one parameter and therefore lacks flexibility, particularly if many items are compared. These weaknesses mean that distance-based models are of limited use in practice (Lee and Yu, 2010).

Let π and τ be rankings. The usual properties of a distance function, $d(\cdot, \cdot)$, between π and τ are:

(1) reflexivity: $d(\pi, \pi) = 0$

(2) positivity: $d(\pi, \tau) > 0$ if $\pi \neq \tau$

(3) symmetry: $d(\pi, \tau) = d(\tau, \pi)$.

Another property that is required for the ranking data is termed right invariance. The right invariance requirement ensures that the distance is not affected if labelling of items is permuted. Suppose that $\pi(k)$ is the rank given to item k in the ranking π. Let φ be a permutation of the items and define the new ranking $\pi \circ \varphi$ by

$$\pi \circ \varphi(k) = \pi(\varphi(k)).$$

Then the right invariance property is $d(\pi, \tau) = d(\pi \circ \varphi, \tau \circ \varphi)$.

Many distances have been considered for this model such as Kendall tau

and Spearman distances. The Kendall tau distance is given by

$$d_{\mathrm{K}}(\pi,\tau)=\sum_{k<k'} I\left\{[\pi(k)-\pi(k')]\,[\tau(k)-\tau(k')]<0\right\}$$

where $I()$ is the indicator function. This is equivalent to d'_τ in Section 2.3. The Spearman distance is

$$d_{\mathrm{S}}(\pi,\tau)=\left(\sum_{k=1}^{K}[\pi(k)-\tau(k)]^2\right)^{\frac{1}{2}}.$$

Diaconis (1988) discussed many other distances and considered a general class of distance based models. Let ζ be a dispersion parameter where $\zeta \geq 0$ and $C(\zeta)$ denote the normalizing constant. Suppose $d(\pi,\pi_0)$ is an arbitrary right-invariant distance. Then a general distance-based model is

$$P(\pi|\zeta,\pi_0)=C(\zeta)\,e^{-\zeta d(\pi,\pi_0)}.$$

We expect most of the rankers to have rankings close to the modal ranking π_0. Rankings nearer to π_0 have a higher probability of occurrence and this is controlled by ζ.

If the Kendall distance is used in the model, this is equivalent to the Mallows ϕ-model when $\phi=e^{-\zeta}$ in Equation (2.2) (Mallows, 1957).

2.5.4 Multistage Models

Multistage models assume that the ranking process can be decomposed into a sequence of independent stages. Suppose rankers independently rank a set of p objects. The process of ranking for each ranker is decomposed into $p-1$ stages. At stage one, the most preferred object is selected. At the second stage, the most preferred remaining object is selected, and so on until the $(p-1)^{th}$ stage.

The Plackett-Luce model which was mentioned before as an order statistics model also belongs to the class of multistage models. The probability of choosing a particular object k at any stage is conditional on the set $\mathcal{O}$ of objects remaining in each stage. This is discussed in more detail in Chapter 3.

Fligner and Verducci (1988) defined a different kind of probability at each stage which does not depend on the objects remaining at that stage, assuming that the accuracy of the choice made at any stage is independent of the accuracies at the other stages. That is the set of choice probabilities at a particular stage depends only on the stage. The probability of a ranking is affected by the correctness of a ranker's choice at each stage based on how close the selected best object of the remaining objects is to a central ranking π_0. Let V_j denote the number of adjacent transpositions required to move the j^{th} ranked object to have the same ranking as π_0 and let π^{-1} denote an ordering set. For example, suppose $\pi_0^{-1} = (C, A, B, D)$ and $\pi^{-1} = (C, D, B, A)$, then the value of V_j are $V_1 = 0$, $V_2 = 2$, and $V_3 = 1$. The most general model for independent $\boldsymbol{V} = (V_1, \ldots, V_{p-1})$ is

$$P(V_j = m) = p(m, j),$$

where $p(m, j) \geq 0$ and $\sum_{m=0}^{p-j} p(m, j) = 1$. The general multistage model for independent $\boldsymbol{V}$ with $\binom{p}{2}$ parameters is given by

$$P(\pi) = \prod_{j=1}^{p-1} p(V_j, j).$$

This is called the free model (Fligner and Verducci, 1988).

2.5.5 Properties of Ranking Models

Critchlow et al. (1991) defined five properties of ranking models. A brief explanation of these properties is as follows:

(1) Label Invariance

Relabelling of objects does not affect the probability models.

(2) Reversibility

This concept was introduced by Luce (1959). Normally rankers rank the objects from best to worst; however, sometimes the rankers may rank from worst to best. If a model has the reversibility property, it means that the ranking probabilities should be the same. The reverse function,$\gamma(\pi)$, for a ranking, π, of p objects is

$$\gamma(j) = (p+1) - j, \qquad j = 1, \ldots, p.$$

(3) Strong Unimodality or Weak Transposition property

A ranking distribution is called unimodal if there is one ranking, π_0, that has higher probability than any other. Let τ_{ij} be a transposition function in which i and j are interchanged as $\tau(i) = j$, $\tau(j) = i$, and $\tau(m) = m$ for all $m \neq i, j$. Moreover, $\pi \circ \tau_{ij}$ is the permutation that agrees with π except that the ranks assigned to item i and item j are transposed. With a modal ranking π_0 for every pair of item i and j such that

$$\pi_0(i) < \pi_0(j)$$

and any permutation π such that

$$\pi(i) = \pi(j) - 1$$
$$P(\pi) \geq P(\pi \circ \tau_{ij})$$

with equality if $\pi = \pi_0$. This guarantees the probability is non-increasing as π moves one step away from π_0. Then a model is strongly unimodal with modal ranking π_0. For example, suppose a modal ranking is ($C \succ B \succ A \succ D$) where B is more preferred than A. We consider two

rankings which are $(A \succ B \succ C \succ D)$ and $(B \succ A \succ C \succ D)$. Under the strong unimodality property then $P(B \succ A \succ C \succ D)$ should be greater than $P(A \succ B \succ C \succ D)$.

(4) Complete Consensus (Transposition property)

Complete consensus is a stronger version of the unimodality property. It applies to every pair of items (i, j). Suppose that $\pi_0(i) < \pi_0(j)$ and for every π that

$$\pi(i) < \pi(j)$$
$$P(\pi) \geq P(\pi \circ \tau_{ij}).$$

Therefore, the complete consensus implies strong unimodality.

(5) L-decomposability

The idea of L-decomposability (also called Luce-decomposability) is motivated by Luce (1959). The ranking of p objects for a ranker can be decomposed into $p-1$ stages. At stage t, where $t = 1, 2, \ldots, p-1$, the best among the objects remaining at each stage is selected and then the selected object will be removed from the following stages.

Properties of each model

The four classes of models satisfy the first property, label invariance. However, not all models satisfy the other properties.

The Thurstonian model satisfies the reversibility property if the random error distribution is symmetric. The L-decomposability property is difficult to verify because it may not have a closed form as it involves a multiple integral. However, the Plackett-Luce model does satisfy this property because the Plackett-Luce model views rankings as a sequential process. The complete consensus property is satisfied as shown by Savage (1956, 1957) and Henery (1981). Thus, since the complete consensus property is satisfied the strong

unimodality property is also satisfied.

For paired comparison models, Marley (1968) showed that models in this class satisfy the reversibility and L-decomposability properties. Later, the strong unimodal and complete consensus properties were shown to hold (Critchlow et al., 1991). Thus, the paired comparison models satisfy all of the properties.

Properties of the distance-based models are discussed in Critchlow et al. (1991). The distance-based model satisfies all the properties with specific distances e.g. Spearman distance and Kendall distance.

The multistage models do not satisfy the reversibility property but it is obvious that they satisfy L-decomposability. The free model satisfies strong unimodality (Alvo and Yu, 2014).

2.6 Animal Dataset

Our motivating dataset is from an internet survey. The survey was undertaken in order to assess the visual appeal of animal species. This survey was part of a research project at the Durrell Institute of Conservation and Ecology, the University of Kent, in partnership with the Australian Geographic Society. The objective of this survey is to understand what drives people to donate to the conservation of certain species and not others, with the long term aim of improving fundraising for animals.

A total of 385 pictures of species were used in the survey, divided into 4 groups as follows:

Group I: pictures 1 - 97 (97 pictures) are illustrations.
Group II: pictures 98 - 185 (88 pictures) are photographs.
Group III: pictures 186 - 281 (96 pictures) are illustrations.
Group IV: pictures 282 - 385 (104 pictures) are photographs.

The images in first two groups were provided by the organization EDGE (Evo-

lutionarily Distinct and Globally Endangered species). The others were provided by the organization WWF (World Wide Fund for Nature).

The data set was collected over a period of about four months between November 2011 and February 2012. The survey consisted of three parts:

Part I: Ten pictures were randomly selected from one of the groups and the participant ranked them from the most appealing picture to the least, by rearranging the ordering on the screen interactively using the mouse.

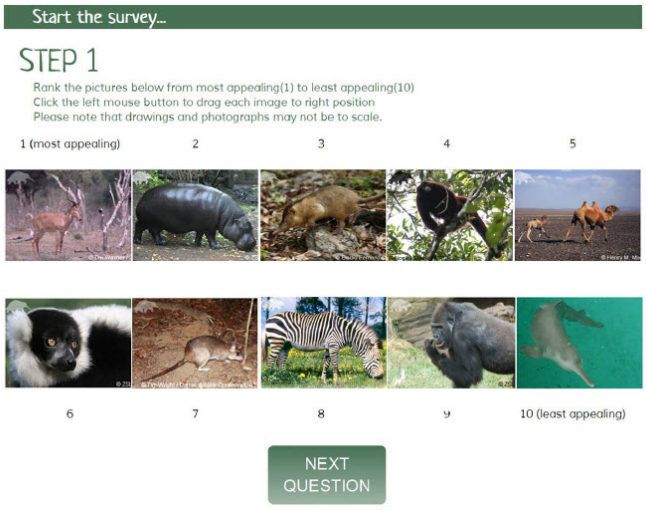


Figure 2.3: Screenshot from the survey

Part II: The participant identified unfamiliar species amongst these 10 pictures.

Part III: The participant provided his/her details which were gender, year of birth and country of origin.

There were 2,040 participants who completed the survey. A small number of observations containing a missing value in gender were removed. Moreover, any cases in which the species in the initial and final order were not the same or when the initial order was exactly the same as the final order were removed from the data set as error records. The reason why the second case was removed is that the probability of the initial ordering being exactly the same as the participant's true preference is very small since there are 10!

different possible orderings of ten items. Therefore it is much more likely that the participant accidentally pressed the "Next Question" button without attempting to rank the species. After cleaning of the data, there were 1,901 participants remaining. The numbers of participants for each group were 450, 468, 529, and 454 for Group I, II, III and IV, respectively.

Each record includes covariates describing item, ranker, and ranker-item covariates. The only item-specific covariate is the animal's type, classified into 3 groups namely mammal, bird, or other.

The ranker-specific covariates are:

(1) Nationality: divided into five groups which are Latin America, North America, Australia, Europe, and other.

   (a) Latin America consists of twelve countries: Argentina, Belize, Brazil, Chile, Colombia, Costa Rica, Cuba, Ecuador, Mexico, Paraguay, Peru, and Uruguay.
   (b) North America consists of two countries: Canada and United States.
   (c) Australia and New Zealand are grouped in Australia nationality.
   (d) Europe consists of thirty-nine countries: Albania, Andorra, Armenia, Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, Ireland, Italy, Jersey, Latvia, Lithuania, Monaco, Netherlands, Norway, Poland, Portugal, Romania, Russia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Serbia, Turkey, Ukraine, and United Kingdom.
   (e) The other group consists of twenty-nine countries: Afghanistan, Algeria, Bermuda, Hong Kong, India, Indonesia, Iran, Israel, Japan, Kenya, Korea South, Laos, Lebanon, Malaysia, Mauritius, Morocco, Myanmar (Burma), Nepal, Pakistan, Philippines, Qatar, Singapore, South Africa, Thailand, Trinidad and Tobago, Tunisia,

Vietnam, Wake Island, and Zimbabwe.

(2) Age is calculated from year of birth and is a continuous covariate (in years).

(3) Gender is a dummy variable where

$$\text{Gender} = \begin{cases} 1, & \text{if female} \\ 0, & \text{if male.} \end{cases}$$

The last type of covariate, ranker-item-specific covariates, are as follows:

(1) Start Position: each participant ranked 10 species which were displayed in two rows as shown in Figure 2.3. The dummy variable Start Position is

$$\text{Start Position} = \begin{cases} 1, & \text{if top row} \\ 0, & \text{otherwise.} \end{cases}$$

(2) Familiarity: each participant indicated the species that they were familiar with. Therefore, Familiarity is a dummy variable where

$$\text{Familiarity} = \begin{cases} 1, & \text{if familiar with the species} \\ 0, & \text{if not familiar with the speices.} \end{cases}$$

2.6.1 Assessment of Ranking Quality

In this section, we investigate whether the participants ranked the given images properly. One question is whether participants only move some species that they have strong opinions about to the top and bottom of the list, while the others are left in the middle.

The normalized Kendall tau distance between the initial and final orderings, d_τ, is calculated in order to investigate this problem. The distance has a value between 0 and 1 where a value of 0 occurs if and only if the original and final orderings are the same and a value of 1 occurs if and only if the final is

the reverse of the initial orderings.

Since the initial ordering in which images were presented was random, we expect it to be unrelated to the final ordering after ranking. In this situation, the sampling distribution of d_τ converges towards a normal distribution as the number of item ranked (p) increases (Kendall, 1938). Therefore we can use the approximate normal distribution of d_τ to assess whether participants are ranking properly. Abdi (2007) stated that the sampling distribution is approximated well by a normal distribution if p is larger than 10; here we have $p = 10$. The normal approximation is given in Kendall (1970). The approximate normal distribution has a mean of 0.5 and a standard deviation

$$\sigma_\tau = \sqrt{\frac{2p+5}{18p(p-1)}}.$$

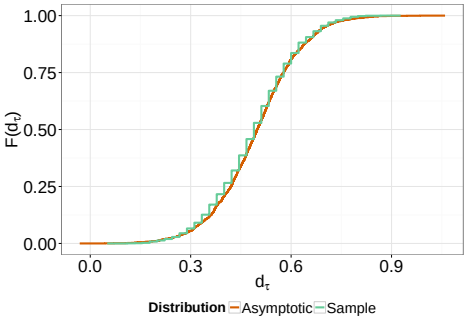
With $p = 10$, the asymptotic null standard deviation of Kendall tau distance is 0.124.

Figure 2.4 shows the empirical cumulative distribution function of the Kendall tau distance and the asymptotic cumulative normal distribution for each group of images. The Kolmogorov-Smirnov test is applied to test the fit of the normal distribution in each group and results are shown in Table 2.2. Table 2.2 shows the Kolmogorov-Smirnov goodness of fit statistic (D)

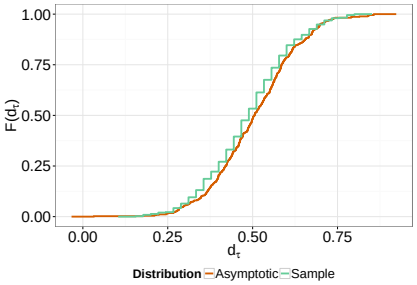
Table 2.2: Kolmogorov-Smirnov test for testing empirical distribution of the Kendall tau distances

	ALL	Group I	Group II	Group III	Group IV
D	0.033	0.049	0.049	0.035	0.033
p-value	0.130	0.470	0.438	0.782	0.896

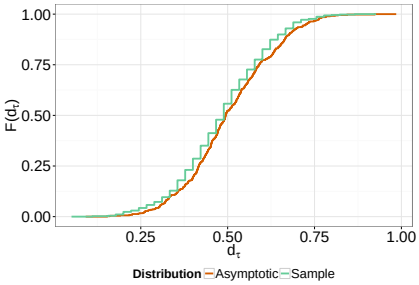
and p-value for each group. The p-value for all participants is smaller than in the separated groups. This is because of the size effect. The sample distribution agrees with the asymptotic distribution at 0.05 significance level in each group. This suggests that the initial orderings are a random permutation of



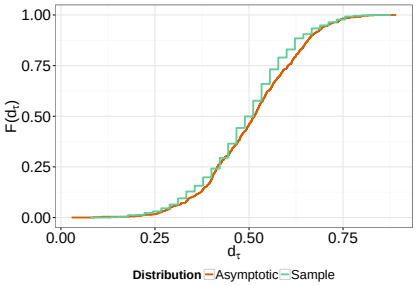
(a) All



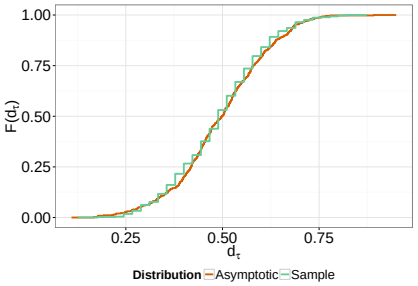
(b) Group I



(c) Group II



(d) Group III



(e) Group IV

Figure 2.4: Cumulative distribution of the Kendall tau distance between initial and final orderings for each group of images.

the participants specific preference and that in this sense the rankings have been done properly.

As we mentioned before, one possible scenario of improper ranked is that the participants only move some species for the top and bottom preference, while the others are left in the middle. Later on, in the Plackett-Luce model, it is assumed that rankings are ranked from best to worst. We explore whether the Kendall tau distance between the initial and final orderings can detect if participants rank in this way. This can be done by doing simulation study. We

generate data according to the Plackett-Luce model with $K = 100$, $p = 10$, and $n = 500$ where each ranking is assumed to rank only top-h and bottom-l. The h and l values are randomly generated integers from 1 to 3. The empirical

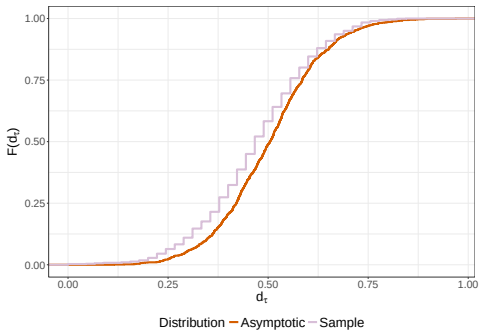


Figure 2.5: Cumulative distribution of the Kendall tau distance between initial and final orderings for the simulated data

cumulative distribution of the Kendall tau distance for the simulated data is shown in Figure 2.5. The results of the Kolmogorov-Smirov test are $D = 0.085$ and p-value $= 0.011$. Therefore, the sample distribution does not agree with the asymptotic distribution at 5% significance level. This means the rankings have not been done properly. The way of investigating the data can distinguish improper ranked data.

2.6.2 Descriptive Statistics

Animal's Type

Animals were classified only to the level of mammal, bird, or the other species. The frequencies of each type are shown in Table 2.3. Table 2.3 is only for

Table 2.3: Frequency for Animal's type

	Mammal	Bird	Other		Mammal	Other
Group III	67	15	14	Group III	67	29
Group IV	75	15	14	Group IV	75	29

Group III and Group IV since there is no information provided for Group I

and Group II. Mostly the animals are from mammal type, therefore, we group bird and other and compare the mammal with other types.

Nationality

As explained above, the dataset used for analysis consisted of 1,901 participants. Table 2.4 shows that most participants were from Europe or North America. There were few from Latin America or Australia; therefore, we combined them with the other groups. After combining, there were three groups of Nationality which were North America, Europe, and other, as shown in the lower half of Table 2.4.

Table 2.4: Frequency for Nationality

Nationality	Group I	Group II	Group III	Group IV
Latin America	18(4.0%)	17(3.6%)	18(3.4%)	18(4.0%)
North America	160(35.6%)	162(34.6%)	196(37.1%)	127(28.0%)
Australia	21(4.7%)	24(5.1%)	20(3.8%)	28(6.2%)
Europe	224(49.8%)	231(49.4%)	261(49.3%)	235(51.8%)
Other	27(6.0%)	34(7.3%)	34(6.4%)	46(10.1%)
Total	450	468	529	454

Nationality	Group I	Group II	Group III	Group IV
North America	160(35.6%)	162(34.6%)	196(37.1%)	127(28.0%)
Europe	224(49.8%)	231(49.4%)	261(49.3%)	235(51.8%)
Other	66(14.7%)	75(16.0%)	72(13.6%)	92(20.3%)
Total	450	468	529	454

Age

Age is a continuous variable. The youngest and oldest participants in the dataset are 7 and 81 years old, respectively as shown in Table 2.5. There is little variation between groups, which is expected since participants were given images from a randomly chosen group. The median of age was 29 and the overall mean was 31.8 years.

Table 2.5: Descriptive statistics of Age (in years)

	Minimum	Median	Mean	Maximum
Group I	7	28	31.3	70
Group II	7	29	32.0	71
Group III	7	29	32.1	73
Group IV	7	29	32.0	81
Overall	7	29	31.8	81

Kernel density plots of Age for participants who ranked the animal images are provided in Figure 2.6. The plots show that males and females have very similar distributions of Age over four groups. Moreover, all the plots are right-skewed, which probably reflects the fact that participants were recruited through social media.

We use Age as a categorical covariate in Chapter 5. Age is divided into two groups by using Age 30 year-old as a threshold. The first group contains the participants from 7 year-old to 30 year-old and the rest belong to the other group. The frequency for Age (as a factor covariate) is shown in Table 2.6.

Table 2.6: Frequency for Age

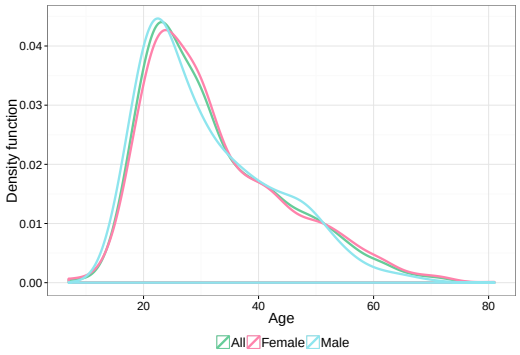
Age	Group I	Group II	Group III	Group IV
$\leq$ 30 year-old	268(59.6%)	258(55.1%)	293(55.4%)	251(55.3%)
$>$ 30 year-old	182(40.4%)	210(44.9%)	236(44.6%)	203(44.7%)
Total	450	468	529	454

Gender

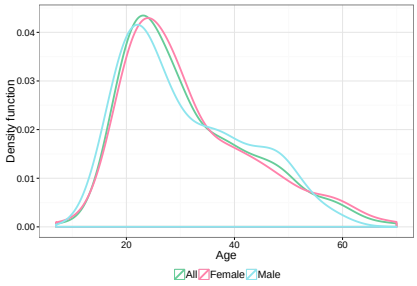
Of the 1,901 participants, 571(30.0%) are male and 1,330(70.0%) are female. Table 2.7 shows the breakdown by group.

Start Position

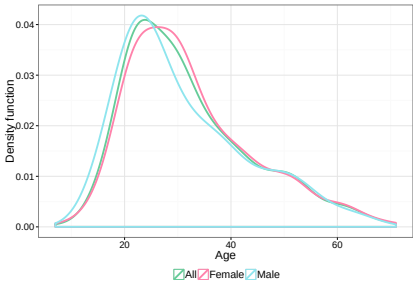
The positions of images are given randomly at the start. We study how participants moved the images on the screen.



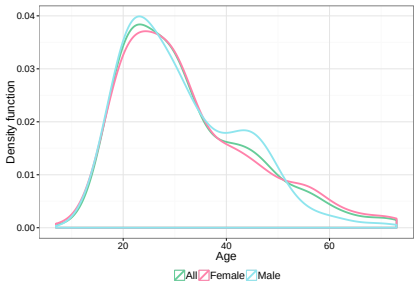
(a) All



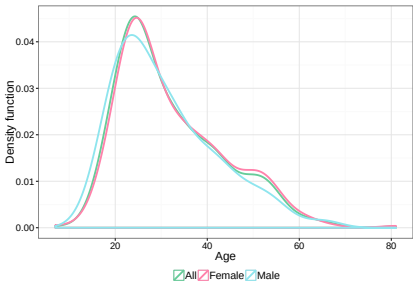
(b) Group I



(c) Group II



(d) Group III



(e) Group IV

Figure 2.6: Kernel density plots of Age by Gender

Table 2.7: Frequency for Gender

Gender	Group I	Group II	Group III	Group IV
Male	134(29.8%)	141(30.1%)	144(27.2%)	152(33.5%)
Female	316(70.2%)	327(69.9%)	385(72.7%)	302(66.5%)
Total	450	468	529	454

Figure 2.7 shows the correlation of the start position and final position where x-axis and y-axis are start position and final position, respectively. The blue colour indicates a high correlation between the start position and the final position. For example, the participants had a tendency not to move images

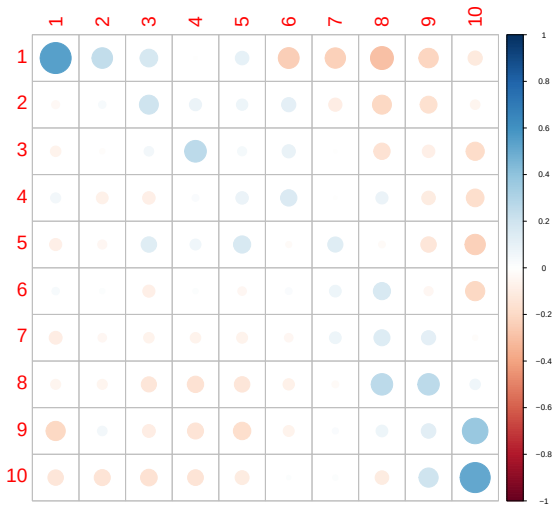


Figure 2.7: The standardized proportion of moving between start position and final position where x-axis is Start Position and y-axis is Final Position

in the most preferred and least preferred position (image 1 and 10). Figure 2.7 also suggests that the images tended to be shuffled in the same row rather than moved between rows.

Familiar Species

During the survey, the participants were asked to indicate which of the species that they had ranked were familiar.

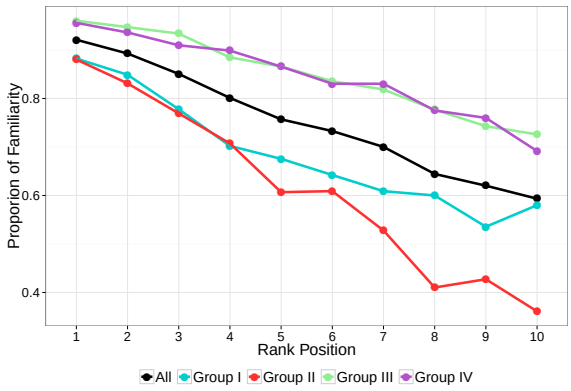


Figure 2.8: Proportion of familiar species in a particular rank position

Figure 2.8 shows the proportion of records that contained a familiar species in each rank position. The proportion decreases steadily as the rank position

increases for all groups and all records. Familiarity is much higher in Group III and IV than in Group I and II as shown in Figure 2.8 and Table 2.8. One possibility is that the animals in these groups, which are from WWF organization, are better known than animals from Group I and Group II. Animals in Group I and Group II are supported by EDGE. These organizations may focus their conservation efforts on different types of species. Table 2.8 indicates that on average the participants are familiar with 6 or 7 of the ten species from Group I and II, and 8 or 9 species from Group III and IV, respectively.

Table 2.8: Mean number of familiar species in the set of ten images (SE in brackets)

Group	I	II	III	IV
Mean	6.85(0.094)	6.13(0.088)	8.49(0.066)	8.45(0.076)

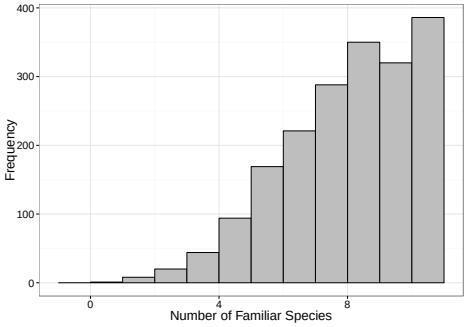
The distributions of number of familiar species across all the records and each group are shown in Figure 2.9. Figure 2.9 shows that the distribution of Group I and II have similar shape while Group III and IV results also have similar shape, but different shape to those of Group I and II. Most of the species are considered familiar in Group III and IV as shown in Figure 2.9c.

Two-sample Kolmogorov-Smirnov tests are applied to test whether the numbers of familiar species from Group I and II, and Group III and IV have the same distributions. The results in Table 2.9 indicate that Group I and Group II do not have the same distribution while Group III and IV have the same distribution of number of familiar species at 0.05 significance level.

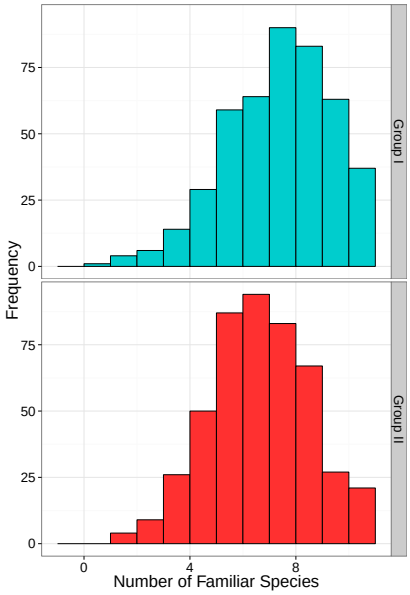
Table 2.9: Kolmogorov-Smirnov test for testing distribution of number of familiar species across all the records

Group	I vs II	III vs IV
D	0.184	0.033
p-value	3×10^{-7}	0.952

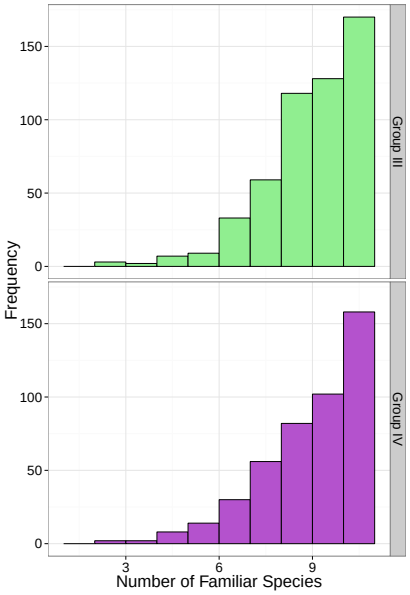
Moreover, Figure 2.10 shows the distribution of the proportion of participants who were familiar with each species and these plots indicate that all



(a) All



(b) Group I and Group II

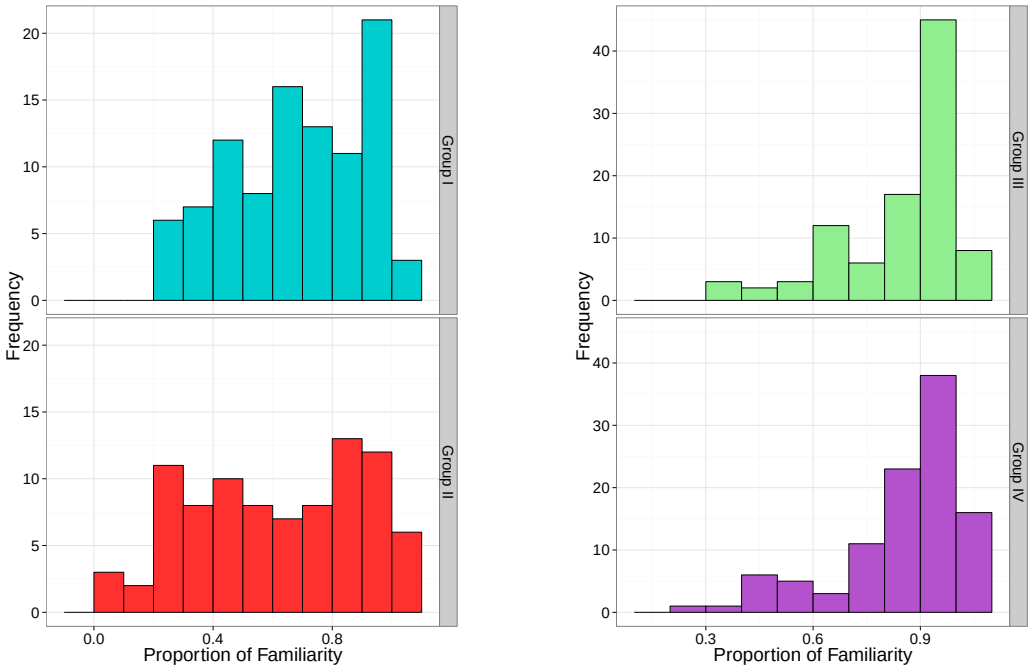


(c) Group III and Group IV

Figure 2.9: Frequency distribution of number of familiar species for each group of images.

the distributions are skew to the left. That means most of the species are recognized by the participants, especially most species in Group III and IV.

The number of times that the species was considered familiar as a proportion of the total number of times that it appeared in the survey is shown in Figure 2.11. It can be observed that the proportion varies among species. There is a difference between Group I and II and Group III and IV. The participants, again, recognized the species in Group III and IV more than Group I and II. In Group II, there are 5 species that less than 20% of participants were familiar with. These species are Anderson's Mouse Opossum, Southern



(a) Group I and Group II

(b) Group III and Group IV

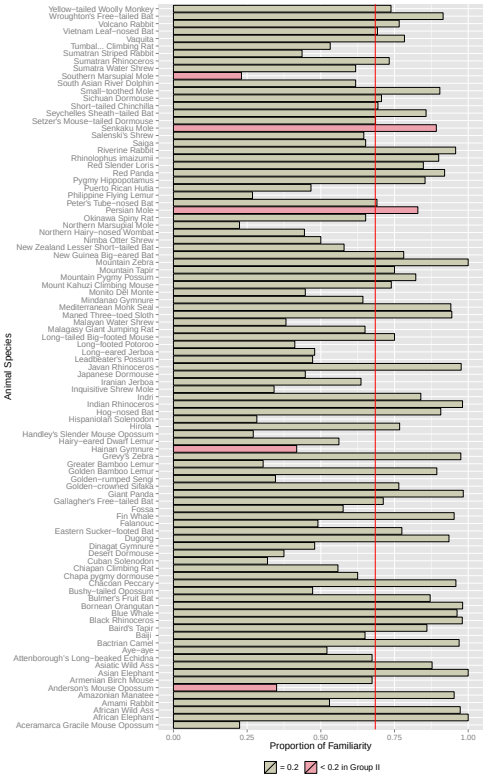
Figure 2.10: Histogram of the proportion of times that a species was considered familiar for each species in each group.

Marsupial Mole, Persian Mole, Hainan Gymnure, and Senkaku Mole which are shown in pink in Figure 2.11a and Figure 2.11b.

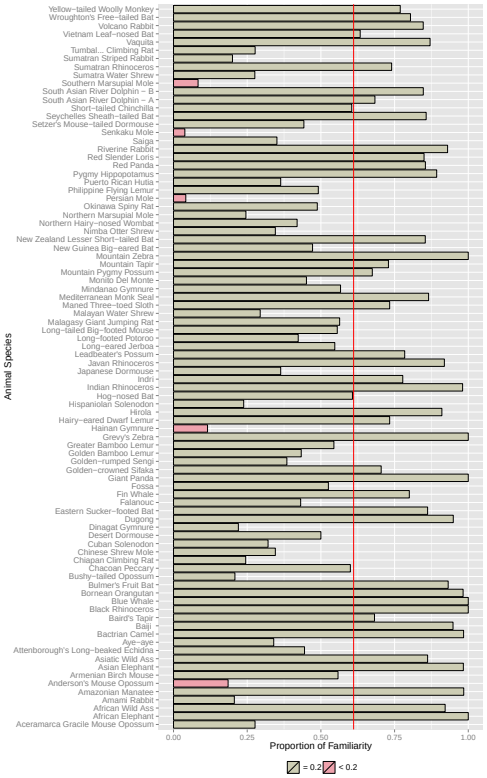
These figures show that most participants are familiar with Persian Mole and Senkaku Mole images in Group I but not in Group II. The percentage of familiarity of Persian Mole in Group I and II are 83 and 4 percent, respectively. Moreover, Senkaku Mole is recognized by 89 and 4 percent of participants who ranked this species in Group I and II, respectively. Note that images from Group I and II are illustrations and photographs, respectively. However it is unclear why there are such large differences.

There are 85 species that appear in both Group I and Group II. Moreover, Group III and Group IV have 96 species that appear in both groups. Figure 2.12b shows that most of the species have almost the same proportion of familiarity between Group III and Group IV.

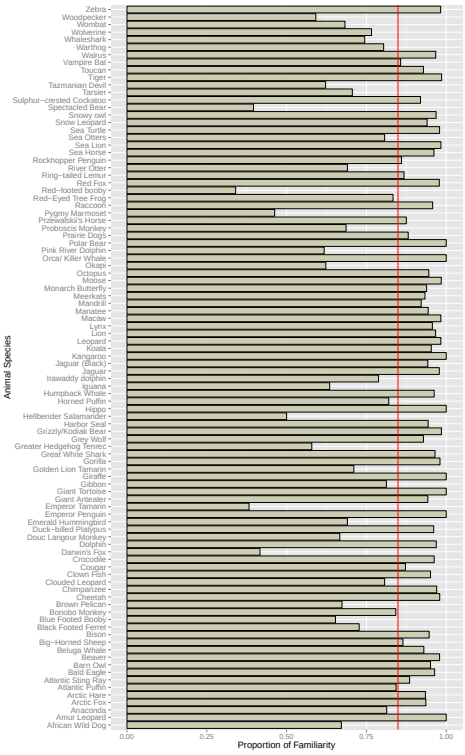
There was no effect of gender on familiarity, the mean number of familiar



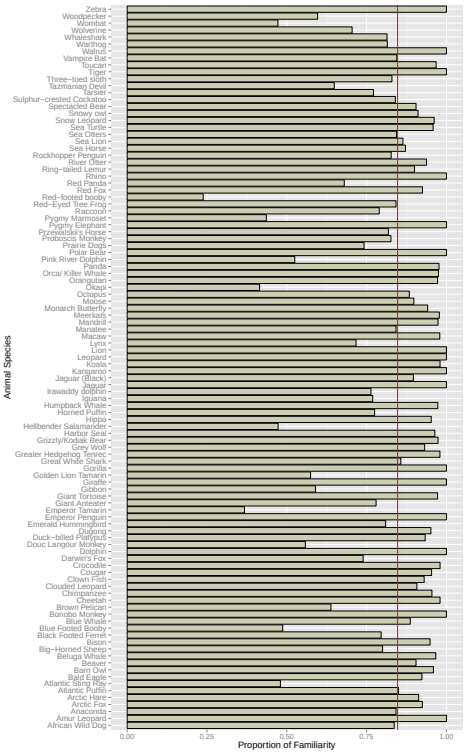
(a) Group I



(b) Group II

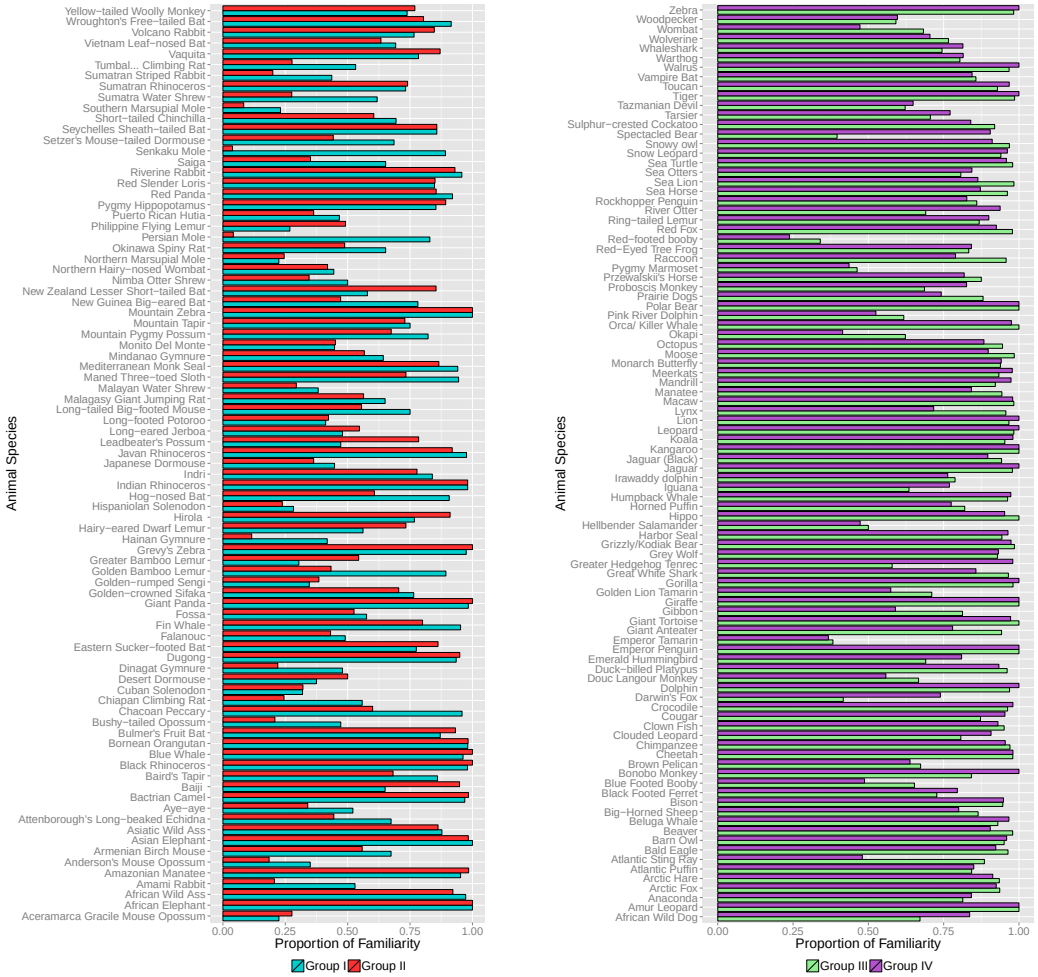


(c) Group III



(d) Group IV

Figure 2.11: Proportion of times that a species was considered familiar. The vertical red line represents the mean proportion of familiar of the group.



(a) Group I vs Group II (b) Group III vs Group IV

Figure 2.12: Proportion of times that a species was considered familiar for the same species in Group I and II, and Group III and IV.

images in the set of 10 images are 7.56(SE=0.055) and 7.41(SE=0.087) images for females and males, respectively.

The effect of age on familiarity is shown by the mean number of familiar images for participants. Age is divided into 6 ranges in order to provide more information as shown in Table 2.10. Most of the age groups are familiar with 6 to 7 images and 8 to 9 images on average for Group I and II, and Group III and IV, respectively; however, the youngest and the oldest groups give different number of familiar images. Participants aged less than 16 years are familiar with 5 to 6 images in Group I and II and with 7 to 8 in Group III and

Table 2.10: Frequency and mean of familiarity for Age in each new age group

Age	Group	Frequency	Mean(SE)
< 16	I	5	6.2(0.663)
	II	10	5(0.537)
	III	9	8(0.745)
	IV	5	6.8(1.068)
16 - 25	I	179	6.84(0.150)
	II	160	6.10(0.155)
	III	188	8.40(0.113)
	IV	158	8.27(0.132)
26 - 35	I	123	6.78(0.175)
	II	151	6.12(0.152)
	III	168	8.41(0.126)
	IV	145	8.59(0.133)
36 - 45	I	78	7.06(0.204)
	II	76	6.04(0.211)
	III	78	8.62(0.150)
	IV	80	8.45(0.164)
46 - 55	I	42	6.36(0.367)
	II	48	6.60(0.252)
	III	54	8.93(0.156)
	IV	51	8.80(0.225)
55 - 65	I	21	7.71(0.437)
	II	19	6(0.501)
	III	22	8.5(0.321)
	IV	12	8.67(0.333)
>65	I	2	7(2)
	II	4	7.25(1.031)
	III	10	8.7(0.473)
	IV	3	7.33(1.764)

IV, while participants aged greater than 65 years are more familiar with the species from Group I and II (7 to 8 images) and from Group III and IV (8 to 9 images). In general, those older than 65 years old (19 participants) are familiar with 8(SE=0.452) images on average in the set of 10 images, while children aged less than 16 years old (29 participants) are familiar with 6.5(SE=0.417) images on average.

2.7 Sushi Dataset

Kamishima (2003) and his colleagues at the National Institute of Advanced Industrial Science and Technology in Japan collected three types of Sushi Preference datasets by using a questionnaire survey method. The three types of dataset were:

- Dataset A: full ranking with 10 sushi flavours
- Dataset B: partial ranking with a total of 100 sushi flavours
- Dataset C: partial rating with a total of 100 sushi flavours.

The Dataset A had only 10 types of sushi which are popular sushi: Shrimp, Sea eel, Tuna, Squid, Sea urchin, Salmon row, Egg, Fatty tuna, Tuna roll and Cucumber roll. The other two datasets contained 100 types of sushi which included these 10 types. The datasets have been used in many ranking studies e.g. Soufiani et al. (2013b), Lu and Boutilier (2011), Bonilla et al. (2010), and Kamishima (2003).

In this thesis, we focus on Dataset B, which involves partial rankings. This dataset contained 100 types and participants were asked to rank 10 types of sushi, which were randomly selected with unequal probabilities. These probabilities were derived from counts of how many of twenty-five sushi restaurants had each sushi type in their menu. The *common* sushi types, which were present in Dataset A, tended to be selected to be ranked more often than the other types as shown in Figure 2.13. Each individual ranked the 10 selected sushi types according to their preferences. There were 5000 individuals who participated this survey.

2.8 Sundarbans Dataset

The final dataset is from a survey that took place in the Sundarbans, one of the largest mangrove forests in the world (UNESCO, 2016). The Sundarbans

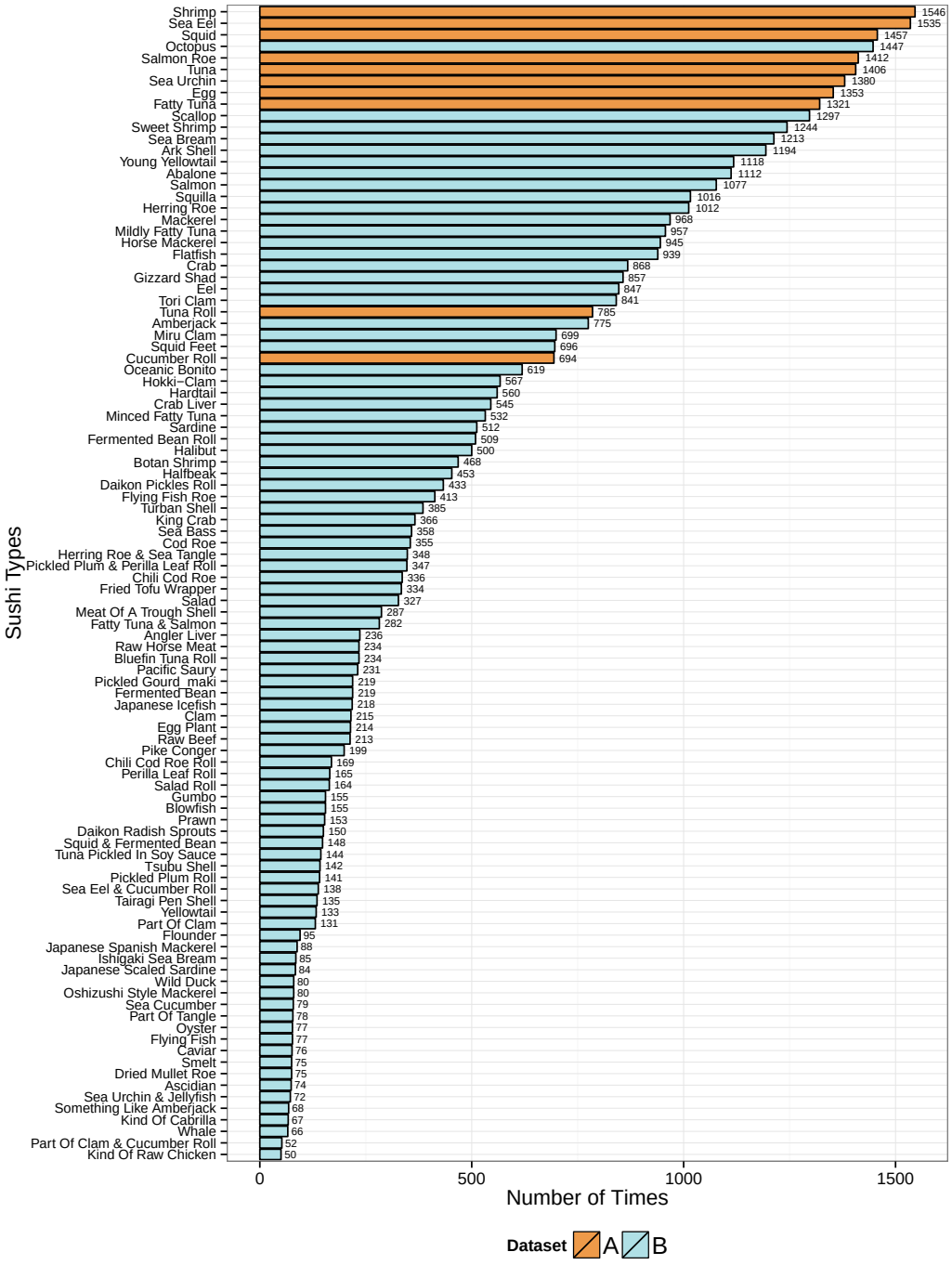


Figure 2.13: Number of times that each sushi type is selected in Dataset B. The orange colour shows the types that are also in Dataset A while the types which appear only in Dataset B are shown in blue.

is located in the south-west of Bangladesh and India, mostly in Bangladesh. The survey was focused on only the Bangladesh Sundarbans because this area is a Class 3 Tiger Conservation Landscape of Global Priority and is one of the world's largest tiger habitats (Inskip et al., 2013). The objective of the

survey was to understand the problems of people who live in, or bordering, this area, especially on human-tiger conflict issues. Data were collected from ten villages which are Bhola, Jewdhara, Khatakhali, Nangli, Terabeka, Bojboja, Kadamtola, Kassiabad, Munshigani, and Tengrakhali. These ten villages were divided into two groups, East and West, according to their location. The first five villages belonged to the West group and the others formed the East group.

The survey was carried out by 2-stage interview. First, interviewees were asked to list all of the problems that they worried about. Second, the interviewees were asked to rank the problems that they had mentioned in the first stage, based on the severity of the problems. Tied rankings were allowed. There are 62 rankings that have ties. A total of 385 participants were interviewed. Interviews were conducted in the Bengali language and the problems identified were translated into English. Then the problems were grouped into 25 categories. Three respondents were removed during this process due to uncertainty about their answers. Moreover, one further respondent was eliminated since there was an error in the record. There were 381 participants remaining.

The 25 categories could be broadly classified into 5 types of problem which were natural, financial, human, social, and physical.

2.8.1 Descriptive Statistics

Number of Problems

The number of problems identified varied between the respondents, ranging from 1 to 7. The average number of problems which the respondents mentioned is 3.15 with standard error 0.060.

A histogram of the number of problems is shown in Figure 2.14. Illustrating that most of the participants listed 1 to 4 problems.

Next, we introduce six ranker-specific covariates which are Gender, Inter-

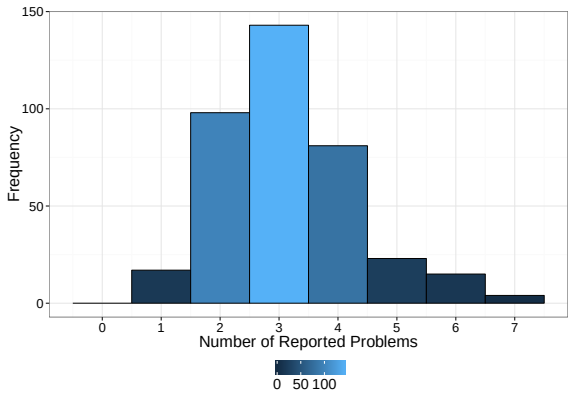


Figure 2.14: Histogram of number of problems.

view Type, Village Location, Age, Education, and Household categories.

Table 2.11: Frequency of Gender and Head of Household

Gender	Frequency				
	Interview Type		Village Location		Total
	Head	Spouse	East	West	
Male	250(94.0%)	0(0%)	147(65.6%)	103(65.6%)	250(65.6%)
Female	16(6.0%)	115(100%)	77(34.4%)	54(34.4%)	131(34.4%)
Total	266(69.8%)	115(30.2%)	224(58.8%)	157(41.2%)	381

Gender

The data consist of 250(65.6%) males and 131(34.4%) females, as shown in Table 2.11.

Interview Type

Interview type consists of two types which are head of household and spouse. The frequency for Interview Type is presented in Table 2.11. The 266 participants are head of households and 115 participants are spouses. Among 266 head of households, there are 250 males and 16 females. All spouses are female.

Villages Location

There are 224 participants from East villages and 157 participants from West villages. Both East and West villages have the same proportion of male and female which are 65.6% and 34.4%, respectively.

Education

Education is recorded as the number of years the participant has spent in education, where 0 means no education. Most of the participants have low education since the distribution is skewed to the right as shown in Figure 2.15. There are 137(36.0%) participants who have no education. School period is

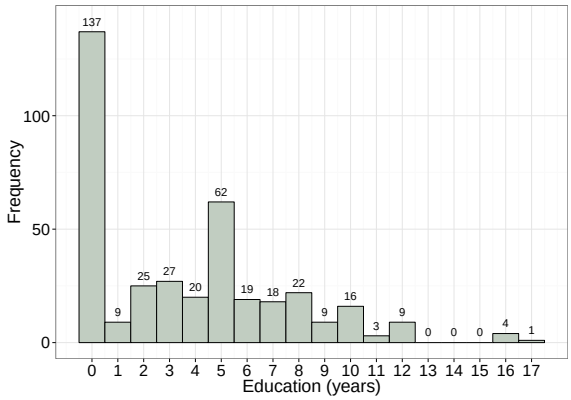


Figure 2.15: Frequency distribution of education

from 1 to 10 years. The 227(59.6%) participants have school-level education. Only 17(4.5%) participants have higher education than school-level.

Age

The youngest and oldest participants are 18 and 82 years old, respectively. The median and mean ages are 40 and 41.7 years, respectively. The distribution of age is present in Figure 2.16. The distribution is slightly skewed to the right.

The distribution of age by gender is provided in Figure 2.17. The age of both male and female distributions are skewed towards the right. The median and mean ages for male are 42 and 45.12 years, respectively. The median

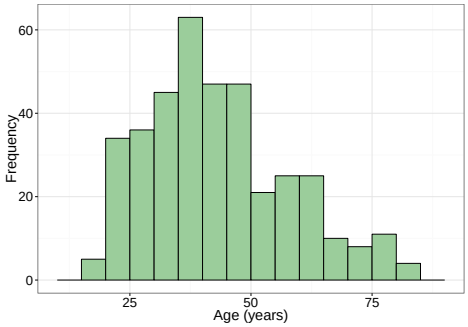
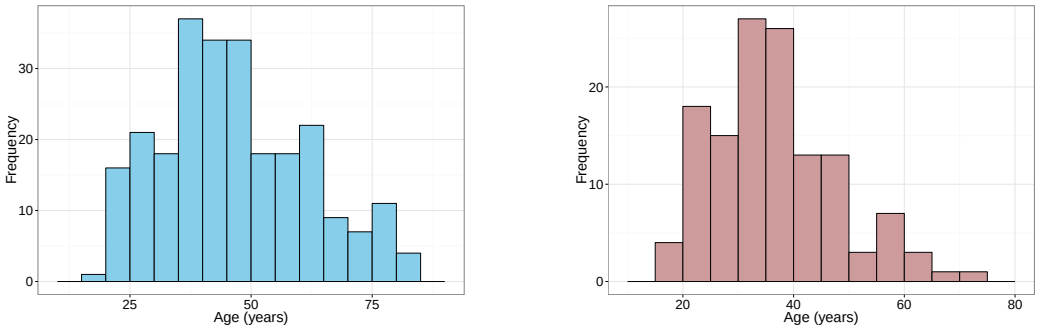


Figure 2.16: Frequency distribution of age.

and mean ages for female are lower than male which are 35 and 35.06 years, respectively.



(a) Male (b) Female

Figure 2.17: Frequency distribution of age by gender

Household Categories

Originally, there were four household categories which indicate the degree of conflict that the household has experienced with tigers. The four categories are fatal human attack, non-fatal human attack, livestock depredation, and no conflict. The frequency of participants who experienced tiger problems is

Table 2.12: Descriptive statistics of number of problems

Category	Frequency
Fatal Attack	95(24.9%)
Non-Fatal Attack	84(22.0%)
Livestock Depredation	102(26.8%)
No Conflict	100(26.2%)

shown in Table 2.12. Later, we group fatal attack household with non-fatal attack household categories because these two categories both involve attacks on people. There are 179(47.0%) households in which family members had been attacked by tigers.

Chapter 3

Models for Partial Ranking

In partial ranking, the main challenge is to decide a global ranking based on partial preferences from rankers. One approach to tackle this kind of data is to assume that the rankings come from a probabilistic model. The most popular statistical models are the Bradley-Terry (BT) model, and the Plackett-Luce (PL) model which is one among several generalized versions of the BT model. The BT model is applicable only for pairwise comparisons, while the PL model allows us to deal with any number of comparisons. Moreover, the PL model is applicable for a complete ranking, or a partial ranking, or a top-h ranking.

In this chapter, we begin with an introduction to the BT model in Section 3.1. The BT model is a popular model for analyzing paired comparisons. This model can be viewed as a logistic regression model. The parameters can be estimated by using maximum likelihood (ML). Many methods can be used to fit this model such as the Newton-Raphson method. However, we follow the early work by Hunter (2004) on the Minorization-Maximization (MM) algorithm. This is because we are going to use this algorithm to fit the more complex model, the PL model in Section 3.2. The observed information matrix can be found through the negative of the Hessian matrix with given data and estimates.

Next in Section 3.2, we explore the PL model. There are two types of

ranking behavior which are forward ranking and backward ranking. In this thesis, we mainly focus on forward ranking. The Luce's Choice Axiom (LCA), which has an important implication – the constant ratio rule, is discussed. The constant ratio rule makes the PL model attractive for partial ranking data. As before, we follow the MM algorithm, which was proposed by Hunter (2004), in order to find the estimates by using ML. Moreover, in a recent work of Caron and Doucet (2012), they proposed the Expectation Maximization (EM) algorithm that works within a Bayesian framework. As for the BT model, the observed information matrix is calculated from the negative of the matrix of second derivatives of the log-likelihood function.

We consider existing packages for the BT and the PL models in the R programming language. Different packages in R, `PLem` and `PLmm` algorithms for both models are examined and compared by using simulated data in Section 3.3. As far as we know, there is no package for computing the observed information matrix of the PL model with partial ranking data. We compare our observed information matrix algorithm with the negative of the Hessian matrix from `optim` function in order to confirm results.

In section 3.4, we discuss the application of the PL model to the Group I data from the Animal dataset. This has been done in order to give an example of interpretation. Afterwards, we perform a bootstrap goodness-of-fit test to check whether the PL model provides a good fit to the Group I data.

In the last section, Section 3.5, we study rank-breaking methods. Soufiani et al. (2013a) and Soufiani and Parkes (2014) studied rank-breaking methods for complete ranking data. The rank-breaking methods are used to break a ranking set into pairwise comparisons. The full, adjacent, and top-h rank-breaking methods are considered in this section. We compare computational times and statistical efficiencies of fitting the BT model to the paired datasets from rank-breaking methods with results of fitting the PL model to the original simulated data. Later in this section, we only focus on the full rank-breaking

method. Khetan and Oh (2016) studied the full rank-breaking method for partial rankings and suggested to use a different weighting instead of equal weighting. They proposed the weighting that is optimal for MSE. We propose other weightings by extending their weighting approach. We present experimental results using both simulated and real-world datasets.

3.1 Bradley-Terry Model

Paired comparisons occur in various fields and Davidson and Farquhar (1976) gave an extensive bibliography on the method of paired comparisons which listed more than 350 papers. Most models for paired comparisons are based on the models of Thurstone (1927) and Bradley and Terry (1952). Here, we focus on the Bradley-Terry model. Bradley and Terry (1952) introduced a model for paired comparisons, where ranking takes place between pairs of items usually drawn from a larger set of items that are of interest. The same idea had been studied before by Zermelo (1929) for estimating the playing strength of participants in chess tournaments (Ebbinghaus, 2008), but despite this the model is generally known as the Bradley-Terry (BT) model.

In the BT model, the probability that item i is preferred to item j is given by

$$P(i \succ j) = \frac{\lambda_i}{\lambda_i + \lambda_j}, \qquad i \neq j$$

where λ_i and λ_j are positive-valued parameters associated with items i and j, respectively.

The model has been applied in many areas including psychology (Tutz, 1986), genetics (Sinsheimer et al., 2000) and sport (Koehler and Ridpath, 1982). In modelling sporting contests, many extensions of the model have been proposed to include factors such as home advantage and current form of players (Agresti, 2002), and to allow the possibility of a tie (Rao and Kupper,

1967). In a recent extension, Cattelan et al. (2013) developed a dynamic paired comparison model for sport tournament data, which allows for time varying abilities in home and away matches. The BT model has also been used for formulating classification problems (Hastie and Tibshirani, 1998).

3.1.1 Connection between the BT Model and Logistic Regression

Suppose that the paired comparisons involve K different items in total and let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)^\intercal$ denote the vector of parameters. It is more common to work with a reparameterized version of the BT model. Letting $v_i = \log(\lambda_i)$, the model is

$$\begin{aligned} P(i \succ j) &= \frac{\exp(v_i)}{\exp(v_i) + \exp(v_j)} \\ &= \frac{1}{1 + \exp(-(v_i - v_j))}, \end{aligned}$$

which is the standard logistic cumulative distribution function evaluated at $v_i - v_j$. Let $\boldsymbol{x} = (x_1, \ldots, x_K)^\intercal$ denote a vector of length K with $x_i = 1$ and $x_j = -1$, and the remaining elements of $\boldsymbol{x}$ set to 0. Also letting $\boldsymbol{v} = (v_1, \ldots, v_K)^\intercal$, the model becomes

$$\text{logit} P(i \succ j) = \boldsymbol{x}^\intercal \boldsymbol{v},$$

where $\boldsymbol{x}^\intercal \boldsymbol{v} = \sum_{k=1}^{K} x_k v_k$ and logit $p = \log \frac{p}{1-p}$, where p is a probability. This is a logistic regression model with a linear predictor containing the unknown parameters $\boldsymbol{v}$.

3.1.2 Maximum Likelihood Estimator for the BT Model

The parameters $\boldsymbol{\lambda}$ can be estimated by using maximum likelihood (ML). In order to use this estimator, the likelihood function is required.

Suppose that there are n independent comparisons between pairs of items and that ties are not allowed. Let n_{ij} denote the number of comparisons between item i and item j and let w_{ij} denote the number of comparisons where item i is preferred to item j, so that $n_{ij} = w_{ij} + w_{ji}$. Also let $w_i = \sum_{j=1, j\neq i}^{K} w_{ij}$ denote the total number of times that item i is preferred in a comparison with another item. Then, since comparisons are assumed to be independent the likelihood function is

$$L(\boldsymbol{\lambda}) = \prod_{1\leq i\neq j\leq K} \left(\frac{\lambda_i}{\lambda_i + \lambda_j}\right)^{w_{ij}}$$

and the log-likelihood function is therefore

$$\begin{aligned} \ell(\boldsymbol{\lambda}) &= \sum_{1\leq i\neq j\leq K} \log\left(\frac{\lambda_i}{\lambda_i + \lambda_j}\right)^{w_{ij}} \\ &= \sum_{i=1}^{K} w_i \log(\lambda_i) - \sum_{1\leq i<j\leq K} n_{ij} \log(\lambda_i + \lambda_j), \end{aligned} \tag{3.1}$$

where $1 \leq i < j \leq K$ denotes the set of ordered pairs $(i, j) \in \{1, \ldots, K\}^2$ such that $i < j$.

The preference probabilities $P(i \succ j)$ are unchanged if all elements of the parameter $\boldsymbol{\lambda}$ are multiplied by a constant. Therefore, in order for the ML estimators to be well-defined, an additional constraint is required such as $\sum_{i=1}^{K} \lambda_i = 1$. Even with this constraint, the ML estimator of $\boldsymbol{\lambda}$ may not exist. This is because if the assumption below does not hold then the estimates are approaching infinity. Ford (1957) showed that $\hat{\boldsymbol{\lambda}}$ will exist if and only if in every partition of the items in two groups, some item in the second group is preferred at least once to some item in the first group.

Minorization-Maximization Algorithm

There are many different algorithms that can be used to fit the BT model (Tsukida and Gupta, 2011), including methods such as iteratively weighted least squares based on the formulation of the BT model as a logistic regression model. Here we discuss one particular algorithm, the Minorization-Maximization (MM) algorithm, because this is also used to fit more complex models discussed later in the thesis.

Lange et al. (2000) and Hunter (2004) showed that the ML estimates for the parameters of the BT model can be obtained by applying the MM algorithm. The MM algorithm is guaranteed to converge to the unique ML estimator under Ford's (1957) assumption mentioned earlier. The MM algorithm creates a surrogate function that minorizes the log-likelihood function and then optimizes the surrogate function. The surrogate function allows maximization of the log-likelihood to be transferred. This is potentially beneficial when the surrogate function is easier to maximize than the log-likelihood function. The log-likelihood function is as in Equation (3.1).

Before moving to the next step, the term *convex* is introduced. Roughly speaking, a convex function is a function that has a bowl shape. A function f is called convex if for all $x, y \in f$ and for any $0 \leq \lambda \leq 1$,

$$f(\lambda y + (1-\lambda)x) \leq \lambda f(y) + (1-\lambda)f(x),$$

and it is called strictly convex if strict inequality holds as shown in Figure 3.1.

In order to construct the surrogate function, we apply the supporting hyperplane property

$$f(x) \geq f(y) + f'(y)(x-y) \qquad \text{for all } x, y > 0, \tag{3.2}$$

where $f(x)$ is a convex function and, regarded as a function of x, the RHS is

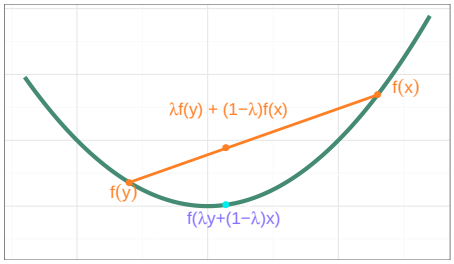


Figure 3.1: f is convex function then $f(\lambda y + (1-\lambda)x) \leq \lambda f(y) + (1-\lambda)f(x)$ for any $0 \leq \lambda \leq 1$

the tangent line at y (as it passes through $(y, f(y))$ and has slope $f'(y)$) and the RHS is the first-order Taylor approximation. Thus, the result is that a convex function lies above its tangent line(s) as shown in Figure 3.2. This is because the first-order Taylor approximation is known as a global underestimator of a convex function (Boyd and Vandenberghe, 2004).

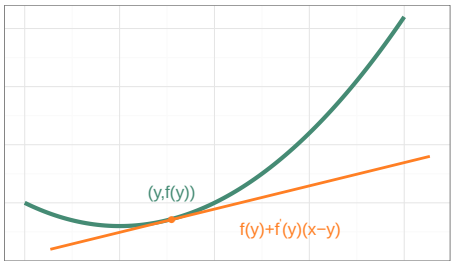


Figure 3.2: f is convex function then $f(x) \geq f(y) + f'(y)(x-y)$ for all $x, y > 0$

The strict convexity of the negative logarithm function implies that for positive x and y and for the choice $f(x) = -\log(x)$, this minorization amounts to

$$-\log(x) \geq 1 - \log(y) - \frac{x}{y} \qquad \text{for all } x, y > 0, \tag{3.3}$$

with equality if and only if $x = y$. We apply this to the second term in the log-likelihood (3.1). Let $x = \lambda_i + \lambda_j$ and $y = \lambda_i^* + \lambda_j^*$ where λ_i^* and λ_j^* are the values from the previous iteration. Then we define $Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*)$ to be the function

$$Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*) = \sum_{i=1}^{K} w_i \log(\lambda_i) - \sum_{1 \leq i < j \leq K} n_{ij} \left[1 - \log(\lambda_i^* + \lambda_j^*) - \frac{\lambda_i + \lambda_j}{\lambda_i^* + \lambda_j^*} \right]$$

$$\equiv \sum_{i=1}^{K} w_i \log(\lambda_i) - \sum_{1 \leq i < j \leq K} n_{ij} \left(\frac{\lambda_i + \lambda_j}{\lambda_i^* + \lambda_j^*} \right),$$

where, since we are interested in maximizing this function with respect to $\boldsymbol{\lambda}$, we omit terms that do not depend on $\boldsymbol{\lambda}$. By the construction of the Q function, $Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*)$ is equal to or less than $\ell(\boldsymbol{\lambda})$ with equality if and only if $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ which implies that

$$\ell(\boldsymbol{\lambda}^{*+1}) \geq Q(\boldsymbol{\lambda}^{*+1}, \boldsymbol{\lambda}^*) \geq Q(\boldsymbol{\lambda}^*, \boldsymbol{\lambda}^*) = \ell(\boldsymbol{\lambda}^*).$$

This sequence of $\boldsymbol{\lambda}^*$ values is therefore guaranteed to increase the log-likelihood. We maximize the Q function by taking the first derivative of this function with respect to λ_i setting this equal to zero and solving for λ_i, which gives

$$\hat{\lambda}_i^{*+1} = \frac{w_i}{\sum_{j=1}^{K} \frac{n_{ij}}{\lambda_i^* + \lambda_j^*}}$$

where $\hat{\lambda}_i^{*+1}$ is the estimated preference value for item i at $(* + 1)^{th}$ iteration. The optimal point is not hard to find because the parameters are estimated separately and explicitly.

So far we have used the constraint $\sum_{i=1}^{K} \lambda_i = 1$. The constraint can be done once at end of the final iteration. An alternative way of introducing a constraint is to treat one of the items as a reference item. We therefore consider this parametrization. The parameters, $\hat{\boldsymbol{\lambda}}$, can be reparameterized following Hunter (2004),

$$\hat{\mu_i} = \log(\hat{\lambda}_i) - \log(\hat{\lambda}_{\text{reference item}}). \tag{3.4}$$

The range of the reparameterized parameters is $-\infty < \hat{\mu}_i < \infty$. The $\hat{\mu}_i$ can be interpreted that if the $\hat{\mu}_i > 0$ it means item i is preferred to the reference item and vice versa if $\hat{\mu}_i < 0$. The reference item has $\hat{\mu}_{\text{reference item}} = 0$.

Observed Information Matrix

Based on the large sample theory of ML estimation, either the observed or the expected information matrix can be used to characterize the parameter estimation performance. However, for ranking models it is usually difficult to calculate the expected information matrix (see further discussion in Chapter 4) and we therefore estimate the variance-covariance matrix of the ML estimator, $\text{var}(\hat{\boldsymbol{\lambda}})$, by the inverse of the observed information matrix ($\mathbf{J}$):

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\lambda}}) &= [\mathbf{J}(\hat{\boldsymbol{\lambda}})]^{-1} \\ &= \begin{bmatrix} \text{var}(\hat{\lambda}_1) & \text{cov}(\hat{\lambda}_1, \hat{\lambda}_2) & \cdots & \text{cov}(\hat{\lambda}_1, \hat{\lambda}_K) \\ \text{cov}(\hat{\lambda}_2, \hat{\lambda}_1) & \text{var}(\hat{\lambda}_2) & \cdots & \text{cov}(\hat{\lambda}_2, \hat{\lambda}_K) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{\lambda}_K, \hat{\lambda}_1) & \text{cov}(\hat{\lambda}_K, \hat{\lambda}_2) & \cdots & \text{var}(\hat{\lambda}_K) \end{bmatrix}.\end{aligned}$$

The elements of the observed information matrix are the negative second derivatives of the log-likelihood function. Let $\lambda_{i(t)}$ be the λ for the item i which is preferred by ranker t and $\lambda_{j(t)}$ be the λ for the item j which is less preferred by ranker t. The log-likelihood function can be written in this form

$$\begin{aligned}\ell(\boldsymbol{\lambda}) &= \sum_{t=1}^{n} \log\left(\frac{\lambda_{i(t)}}{\lambda_{i(t)} + \lambda_{j(t)}}\right) \\ &= \sum_{t=1}^{n} \left[\log(\lambda_{i(t)}) - \log(\lambda_{i(t)} + \lambda_{j(t)})\right].\end{aligned}$$

Considering only a single ranker, we drop the subscript t and then the first and second derivatives are

$$\begin{aligned}\frac{\partial \ell}{\partial \lambda_i} &= \frac{1}{\lambda_i} - \frac{1}{\lambda_i + \lambda_j}, \qquad & \frac{\partial^2 \ell}{\partial \lambda_i^2} &= \frac{1}{-\lambda_i^2} + \frac{1}{(\lambda_i + \lambda_j)^2}, \\ \frac{\partial \ell}{\partial \lambda_j} &= -\frac{1}{(\lambda_i + \lambda_j)}, \qquad & \frac{\partial^2 \ell}{\partial \lambda_j^2} &= \frac{1}{(\lambda_i + \lambda_j)^2},\end{aligned}$$

and because $i \neq j$

$$\frac{\partial^2 \ell}{\partial \lambda_i \partial \lambda_j} = \frac{1}{(\lambda_i + \lambda_j)^2}.$$

Summing the above second derivative terms over all rankers, we get the Hessian matrix. The negative of the Hessian matrix is the observed information matrix of $\hat{\boldsymbol{\lambda}}$.

The observed information matrix of the reparameterized parameters can be found from the re-expressed form of the log-likelihood function:

$$\begin{aligned}\ell(\boldsymbol{\mu}) &= \sum_{t=1}^{n} \log\left(\frac{\exp(\mu_{i(t)})}{\exp(\mu_{i(t)}) + \exp(\mu_{j(t)})}\right) \\ &= \sum_{t=1}^{n} \left[\mu_{i(t)} - \log\left(\exp(\mu_{i(t)}) + \exp(\mu_{j(t)})\right)\right].\end{aligned}$$

The first and second derivatives for a single ranker are

$$\begin{aligned}\frac{\partial \ell}{\partial \mu_i} &= 1 - \frac{\exp(\mu_i)}{\exp(\mu_i) + \exp(\mu_j)}, & \frac{\partial^2 \ell}{\partial \mu_i^2} &= -\frac{\exp(\mu_i + \mu_j)}{(\exp(\mu_i) + \exp(\mu_j))^2} \\ \frac{\partial \ell}{\partial \mu_j} &= -\frac{\exp(\mu_j)}{\exp(\mu_i) + \exp(\mu_j)}, & \frac{\partial^2 \ell}{\partial \mu_j^2} &= -\frac{\exp(\mu_i + \mu_j)}{(\exp(\mu_i) + \exp(\mu_j))^2}\end{aligned}$$

and since $i \neq j$

$$\frac{\partial^2 \ell}{\partial \mu_i \partial \mu_j} = \frac{\exp(\mu_i + \mu_j)}{(\exp(\mu_i) + \exp(\mu_j))^2}.$$

3.2 Plackett-Luce Model

The Plackett-Luce (PL) model generalizes the BT model, which is only for pairwise comparisons, to a model for any number of ranked items. The PL model was proposed independently by Luce (1959) and Plackett (1975). Plackett (1975) was inspired by horse races. Luce (1959) established a choice based axiomatic foundation for this model, Luce's Choice Axiom (LCA), to describe individual choice behaviour based on a general axiom. We describe LCA later in Section 3.2.1. The extension of LCA led to a model of an individual's ten-

dency to choose one object over another. Plackett (1975) proposed a series of increasingly complex models, the Luce model is equivalent to the first-order model in this series (Critchlow et al., 1991). The PL model is appropriate for partial or incomplete rankings, such as horse racing and auto car racing. In the preference area, the items are ranked from *best* to *worst*, such that there are no ties in the ranking. This kind of ranking is called forward ranking. There is another way of ranking, in the opposite direction, which is backward ranking. Luce (1959) stated that forward and backward ranking do not lead to the same result since the PL model is irreversible.

The PL model has been applied to many fields including horse racing (Plackett, 1975), Irish election data (Gormley and Murphy, 2008), label ranking (Cheng et al., 2010), and including time variation e.g. individuals' preferences may change over time (Baker and McHale, 2015). In psychology, the PL model is also a popular model for investigating the preferences of a specific population or for studying how people make choices under uncertainty (Hino et al., 2010; Tran et al., 2016).

In the PL model, the probability of the ranking $\rho_i \equiv (\rho_{i1} \succ \rho_{i2} \succ \cdots \succ \rho_{ip_i})$ is

$$
\begin{aligned}
P(\rho_i; \boldsymbol{\lambda}) &= \frac{\lambda_{\rho_{i1}}}{\lambda_{\rho_{i1}} + \cdots + \lambda_{\rho_{ip_i}}} \times \frac{\lambda_{\rho_{i2}}}{\lambda_{\rho_{i2}} + \cdots + \lambda_{\rho_{ip_i}}} \times \cdots \times \frac{\lambda_{\rho_{ip_i-1}}}{\lambda_{\rho_{ip_i-1}} + \lambda_{\rho_{ip_i}}} \times \frac{\lambda_{\rho_{ip_i}}}{\lambda_{\rho_{ip_i}}} \\
&= \prod_{j=1}^{p_i} \frac{\lambda_{\rho_{ij}}}{\sum_{m=j}^{p_i} \lambda_{\rho_{im}}} \qquad (3.5) \\
&= \prod_{j=1}^{p_i-1} \frac{\lambda_{\rho_{ij}}}{\sum_{m=j}^{p_i} \lambda_{\rho_{im}}}, \qquad (3.6)
\end{aligned}
$$

where $\lambda_{\rho_{ij}}$ is a positive value indicating the preference for item ρ_{ij}. The PL model views ranking as a sequential process as shown in Equation (3.5). First, the rankers choose their preferred item, then they continually choose their preferred item from those that remain until the ranking is complete so that

the PL model belongs to the family of multistage ranking models (Marden, 1995). The reduced form of the PL model in Equation (3.6) arises because the last term in the full form always equals one, the probability that the last item is ranked first when there is only one item left to rank. We can re-express Equation (3.5) as

$$P(\rho_i; \boldsymbol{\mu}) = \prod_{j=1}^{p_i} \frac{\exp(\mu_{\rho_{ij}})}{\sum_{m=j}^{p_i} \exp(\mu_{\rho_{im}})}, \tag{3.7}$$

where $\mu_{\rho_{ij}}$ is reparameterized parameter as in the previous section.

The reverse PL model for backward ranking can be seen as choice by elimination (Tran et al., 2014). That means the ranking process is reversed – the least preferred item is ranked/eliminated first. Tversky (1972) introduced this choice by elimination and after that it has been studied in the psychological area. The advantage is that near the end of the process only the best items remain and comparisons should be easier when there are only a small number of items to compare. This is different from the PL model because in the PL model the best items are compared against all other items. Tran et al. (2014) introduced a reverse PL model as follows

$$P\left(\rho_i; \boldsymbol{\lambda}\right) = \prod_{j=1}^{p_i} \frac{\exp(-\mu_{\rho_{ij}})}{\sum_{m=1}^{j} \exp(-\mu_{\rho_{im}})}.$$

Tran et al. (2014) assumed that the probability of an item being eliminated is inversely proportional to its worth.

The PL model satisfies LCA (Marden, 1995). This result follows from Yellott (1977). Yellott (1977) showed that the PL model satisfies the LCA through a Thurstonian model. This is because the PL model is the Thurstonian model based on the Gumbel distribution. Moreover, because the Gumbel distribution is asymmetric the PL model does not satisfy reversibility property, as mentioned before.

3.2.1 Luce's Choice Axiom

Luce (1959) proposed an axiom which is an assumption about how rankers make choices. LCA is a probabilistic choice theory. Let S and T be subsets of items with $S \subset T$ and suppose that an item is chosen from T. Let $P_T(S)$ be the probability that the chosen element lies in S. The probability axioms are

(i) For $S \subset T, 0 \leq P_T(S) \leq 1$

(ii) $P_T(T) = 1$

(iii) If $R, S \subset T$ and $R \cap S = \emptyset$ then $P_T(R \cup S) = P_T(R) + P_T(S)$.

Let $x \in T$ and $P_T(x)$ denote the probability that x is selected. The probability axiom (iii) implies that

$$P_T(S) = \sum_{x \in S} P_T(x).$$

The probability axioms do not indicate how probabilities of selection are related over different sets. For example, how a ranker selects an object from a smaller set when the same object is also in a larger set of alternatives. This connection is necessary for a theory of choice. Thus, LCA investigates this connection. Let $P(x, y)$ stand for $P_{\{x,y\}}(x)$ when $x \neq y$ and then $P(x, y) + P(y, x) = 1$. The axiom has two parts as follows:

(i) If $P(x, y) \neq 0, 1$ for all $x, y \in T$ then for $R \subset S \subset T$

$$P_T(R) = P_S(R) \cdot P_T(S)$$

(ii) If $P(x, y) = 0$ for some $x, y \in T$ then for every $S \subset T$

$$P_T(S) = P_{T-\{x\}}(S - \{x\}).$$

Part (i) of LCA states that the probability of choosing the set of alternatives R from T is the same as the probability of choosing R from S multiplied by the probability of choosing S from T. This can be viewed as a conditional

probability

$$\begin{aligned} P_T(R|S) &= P_S(R) \\ &= \frac{P_T(R)}{P_T(S)}, \end{aligned}$$

where $P_T(R|S)$ is the probability that R is chosen from S when the larger set T is available and $P_T(S) > 0$. For example, suppose that T is the set of desserts where $T = \{\text{shortcake, ice cream, pudding}\}$, S is a subset of T ($S = \{\text{shortcake, pudding}\}$), and R has only one element which is shortcake. According to part (i), the probability of choosing shortcake from S, is the same as the conditional probability of choosing shortcake from S when the whole menu, T, is available.

Part (ii) allows us to delete x from T without impacting the choice probability if x is never preferred in pairwise choices. For example, if pudding is never chosen in preference to shortcake, then the choice among pudding, shortcake, and ice cream can be safely reduced to the choice of shortcake and ice cream.

The implication of LCA from LCA (i) is that if $P(x, y) \neq 0, 1$ for all $x, y \in T$ then for any $S \subset T$ such at $x, y \in S$,

$$\frac{P(x, y)}{P(y, x)} = \frac{P_S(x)}{P_S(y)}.$$

That is when LCA holds for T and its subsets, the ratio $\frac{P_S(x)}{P_S(y)}$ is independent of S. It implies that the relative probabilities of choosing between two items is independent among the choices available. For example, suppose that $\mathcal{O} = \{A, B, C, D\}$ and $S \subset T$ where $S = \{A, B\}$, and $T = \mathcal{O}$ then by the independence of choosing item among the choice

$$\frac{P_{\{A,B\}}(A)}{P_{\{A,B\}}(B)} = \frac{P_{\{A,B,C\}}(A)}{P_{\{A,B,C\}}(B)} = \frac{P_{\mathcal{O}}(A)}{P_{\mathcal{O}}(B)}.$$

LCA states that the relative probability of choosing A in preference to B should not depend on whether other items are in the set of choices. When more items are introduced, we expect the absolute probabilities of choosing A or choosing B to decrease. However, according to LCA, this ratio of probabilities of any two alternatives should remain the same when expanding to the set of all items. In other words, the ratio of the probability of choosing one item to the probability of choosing another should be constant and this relationship can be called the constant ratio rule. The constant ratio rule is a probabilistic version of the Independence from Irrelevant Alternatives (IIA) from Arrow (1951). However, this property is unrealistic when items are very similar or substitutes for the others (Yu, 2000). Considering the red-bus-blue-bus problem as an example, a person has a choice of going to work by driving a car or taking a blue bus then $\mathcal{O} = \{\text{Car, Blue Bus}\}$. By assuming that the probabilities of taking a car and taking a blue bus are equal such that $P_{\mathcal{O}}(\text{Car}) = P_{\mathcal{O}}(\text{Blue Bus}) = \frac{1}{2}$ then the ratio of probabilities is

$$\frac{P_{\mathcal{O}}(\text{Car})}{P_{\mathcal{O}}(\text{Blue Bus})} = 1.$$

Later a new bus, red bus, is introduced then $\mathcal{O}' = \{\text{Car, Blue Bus, Red Bus}\}$ and the person considers the red bus to be the same as the blue bus. The probability of taking the blue bus is the same as taking the red bus then

$$\frac{P_{\mathcal{O}'}(\text{Blue Bus})}{P_{\mathcal{O}'}(\text{Red Bus})} = 1.$$

However, with the constant ratio rule, the ratio of $\frac{P_{\mathcal{O}'}(\text{Car})}{P_{\mathcal{O}'}(\text{Blue Bus})}$ remains the same when the red bus is introduced. This ratio will remain the same if

$$P_{\mathcal{O}'}(\text{Car}) = P_{\mathcal{O}'}(\text{Blue Bus}) = P_{\mathcal{O}'}(\text{Red Bus}) = \frac{1}{3},$$

then

$$\frac{P_{\mathcal{O}'}(\text{Car})}{P_{\mathcal{O}'}(\text{Blue Bus})} = 1 \quad \text{and} \quad \frac{P_{\mathcal{O}'}(\text{Blue Bus})}{P_{\mathcal{O}'}(\text{Red Bus})} = 1.$$

This is hard to believe in real life since we would expect the probability of taking the car to remain the same when the red bus is introduced. The probability of taking bus is shared between the blue and the red buses then the probabilities are

$$P_{\mathcal{O}'}(\text{Car}) = \frac{1}{2} \quad \text{and} \quad P_{\mathcal{O}'}(\text{Blue Bus}) = P_{\mathcal{O}'}(\text{Red Bus}) = \frac{1}{4}.$$

In this case, the IIA property underestimates and overestimates the probabilities of taking the car and taking buses, respectively. The ratio of probabilities of taking car and blue bus changes when the red bus is introduced rather than remaining constant as stated in the IIA. Therefore, the IIA property can be violated when items are substituted.

The constant ratio rule is very attractive for partial ranking data because it implies that we can get information about overall preferences from the partial rankings.

Yellott (1977) showed that a Thurstonian model satisfies the LCA only if the distribution of preferences is a Gumbel distribution with fixed scale parameter, which is equivalent to the PL model.

The Gumbel distribution, or the extreme value distribution, is right-skewed and has two parameters, a location parameter μ and a scale parameter β. The probability density function (PDF) $f(x|\mu, \beta)$ and cumulative distribution function (CDF) $F(x; \mu, \beta)$ are given by

$$f(x; \mu, \beta) = \frac{z}{\beta} e^{-z},$$

$$F(x; \mu, \beta) = e^{-z},$$

respectively, where $z(x) = e^{-\frac{x-\mu}{\beta}}$ and $x \in (-\infty, \infty)$. The Gumbel distribution with fixed location of 0 and fixed scale of 1, which is termed the standard Gumbel distribution, has PDF and CDF as follows

$$f(x) = e^{-(x+e^{-x})}, \qquad x \in (-\infty, \infty)$$
$$F(x) = e^{-e^{-x}}, \qquad x \in (-\infty, \infty).$$

Assume that all individuals make the ranking decision according to their preferences where a high ranking implies a high value. The value of item k for individual i, U_{ik}, is defined as

$$U_{ik} = V_{ik} + \epsilon_{ik}, \qquad k = 1, \ldots, K$$

where V_{ik} is a constant variable and ϵ_{ik} is an error term which is a random unobserved variable. If the error terms are independent and have a standard Gumbel distribution then the probability that an individual's value of item l is less than z is

$$P(U_l \leq z) = P(\epsilon_l \leq z - V_l),$$

where, for convenience, the subscript i for indicating the individual is omitted. Then the probability that the value of item k is larger than the value of item l is

$$\begin{aligned} P(U_l \leq U_k) &= P(\epsilon_l \leq U_k - V_l) \\ &= e^{-e^{-(V_k+\epsilon_k-V_l)}}, \qquad k, l \in 1, \ldots, K \text{ and } k \neq l. \end{aligned}$$

Now, we are interested in which of the two items, k and l will be preferred. An individual ranks the item with *greater* value higher than the other item. Moreover, if U_k and U_l are continuous variables then $P(U_k = U_l) = 0$. Thus,

the probability that item k is preferred to item l where $k \neq l$ is

$$\begin{aligned}
P(k \succ l) &= P(U_k > U_l) \\
&= \int_{-\infty}^{\infty} P(U_k > U_l \mid \epsilon_k) f_\epsilon(\epsilon_k)\, d\epsilon_k \\
&= \int_{-\infty}^{\infty} e^{-e^{-(V_k+\epsilon_k-V_l)}} e^{-\epsilon_k} e^{-e^{-\epsilon_k}}\, d\epsilon_k \\
&= \int_{-\infty}^{\infty} e^{-e^{-\epsilon_k}(e^{-(V_k-V_l)}+1)} e^{-\epsilon_k}\, d\epsilon_k \\
&= -\int_{\infty}^{0} e^{-t(e^{-(V_k-V_l)}+1)}\, dt, \quad t = e^{-\epsilon_k} \text{ and } dt = -e^{-\epsilon_k} d\epsilon_k \\
&= \int_{0}^{\infty} e^{-t(e^{-(V_k-V_l)}+1)}\, dt \\
&= \left[-\frac{e^{-t(e^{-(V_k-V_l)}+1)}}{e^{-(V_k-V_l)}+1} \right]_{t=0}^{t=\infty} \\
&= \frac{1}{e^{-(V_k-V_l)}+1} \\
&= \frac{e^{V_k}}{e^{V_k}+e^{V_l}}
\end{aligned}$$

where $\infty < t < 0$ (Train, 2003). We can extend this to obtain the probability that the value of item k is the largest among all items ranked by the individual. Under the assumption that ϵ_k's are independent and identically Gumbel distributed, that means the U_k's are independent where $k = 1, \ldots, K$ (Cramer, 2003). Then

$$\begin{aligned}
P(U_k > U_l, \forall k \neq l) &= \int_{-\infty}^{\infty} \prod_{j \neq k}^{K} e^{-e^{-(U_k-V_j)}} e^{-\epsilon_k} e^{-e^{-\epsilon_k}}\, d\epsilon_k \\
&= \int_{-\infty}^{\infty} \prod_{j=1}^{K} e^{-e^{-(V_k+\epsilon_k-V_j)}} e^{-\epsilon_k}\, d\epsilon_k \\
&= \int_{-\infty}^{\infty} e^{-e^{-\epsilon_k}\sum_{j=1}^{K} e^{-(V_k-V_j)}} e^{-\epsilon_k}\, d\epsilon_k
\end{aligned}$$

substitute $t = e^{-\epsilon_k}$

$$dt = -e^{-\epsilon_k} d\epsilon_k$$

$\infty < t < 0$ then

$$\begin{aligned} P(U_k > U_l, \forall k \neq l) &= \int_0^{\infty} e^{-t\sum_{j=1}^{K} e^{-(V_k - V_j)}} dt \\ &= \left[-\frac{e^{-t\sum_{j=1}^{K} e^{-(V_k - V_j)}}}{\sum_{j=1}^{K} e^{-(V_k - V_j)}} \right]_0^{\infty} \\ &= \frac{1}{\sum_{j=1}^{K} e^{-(V_k - V_j)}} \\ &= \frac{e^{V_k}}{\sum_{j=1}^{K} e^{V_j}}. \end{aligned}$$

The overall probability of the ranking of p items out of K items can be expressed by using all derived results. Following Beggs et al. (1981) and letting ρ_j be the item index that ranked in j^{th} position then the probability is

$P(U_{\rho_1} > U_{\rho_2} > \cdots > U_{\rho_p})$

$$\begin{aligned} &= \int_{-\infty}^{\infty} \int_{-\infty}^{U_{\rho_1}} \int_{-\infty}^{U_{\rho_2}} \cdots \int_{-\infty}^{U_{\rho_{p-1}}} \prod_{j=1}^{p} e^{-e^{-(U_{\rho_j} - V_{\rho_j})}} e^{-(U_{\rho_j} - V_{\rho_j})} \, dU_{\rho_p} \cdots U_{\rho_1} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{U_{\rho_1}} \int_{-\infty}^{U_{\rho_2}} \cdots \int_{-\infty}^{U_{\rho_{p-2}}} C e^{-e^{-U_{\rho_{p-1}}}(e^{V_{\rho_{p-1}}} + e^{V_{\rho_p}})} e^{-(U_{\rho_{p-1}} - V_{\rho_{p-1}})} \\ &\qquad dU_{\rho_{p-1}} \cdots U_{\rho_1}, \text{ where } C = \prod_{j=1}^{p-2} e^{-e^{(U_{\rho_j} - V_{\rho_j})}} e^{(U_{\rho_j} - V_{\rho_j})} \\ &= -e^{V_{\rho_{p-1}}} \int_{-\infty}^{\infty} \int_{-\infty}^{U_{\rho_1}} \int_{-\infty}^{U_{\rho_2}} \cdots \int_{e^{-U_{\rho_{p-2}}}}^{\infty} C e^{-t(e^{V_{\rho_{p-1}}} + e^{V_{\rho_p}})} \, dU_{\rho_{p-1}} \cdots U_{\rho_1}, \\ &\qquad t = e^{-V_{\rho_{p-1}}} \\ &= \frac{e^{V_{\rho_{p-1}}}}{e^{V_{\rho_{p-1}}} + e^{V_{\rho_p}}} \int_{-\infty}^{\infty} \int_{-\infty}^{U_{\rho_1}} \int_{-\infty}^{U_{\rho_2}} \cdots \int_{-\infty}^{U_{\rho_{p-3}}} C e^{-e^{-U_{\rho_{p-2}}}(e^{V_{\rho_{p-1}}} + e^{V_{\rho_p}})} \\ &\qquad dU_{\rho_{p-2}} \cdots U_{\rho_1}, \\ &= \cdots = \prod_{j=2}^{p-1} \left(\frac{e^{V_{\rho_j}}}{\sum_{m=j}^{p} e^{V_{\rho_m}}} \right) \int_{-\infty}^{\infty} e^{-e^{-U_{\rho_1}}(\sum_{m=2}^{p} e^{V_{\rho_m}})} e^{-e^{-(U_{\rho_1} - V_{\rho_1})}} e^{-(U_{\rho_1} - V_{\rho_1})} dU_{\rho_1} \\ &= \prod_{j=2}^{p-1} \left(\frac{e^{V_{\rho_j}}}{\sum_{m=j}^{p} e^{V_{\rho_m}}} \right) \int_{-\infty}^{\infty} e^{-e^{-U_{\rho_1}}(\sum_{m=1}^{p} e^{V_{\rho_m}})} e^{-(U_{\rho_1} - V_{\rho_1})} dU_{\rho_1} \end{aligned}$$

substitute $t = e^{-U_{\rho_1}}$

$$dt = -e^{-U_{\rho_1}} dU_{\rho_1}$$

$\infty < t < 0$ then

$$
\begin{aligned}
P(U_{\rho_1} > U_{\rho_2} > \cdots > U_{\rho_p}) &= \prod_{j=2}^{p-1} \left(\frac{e^{V_{\rho_j}}}{\sum_{m=j}^{p} e^{V_{\rho_m}}} \right) e^{V_{\rho_1}} \int_0^{\infty} e^{-t(\sum_{m=1}^{p} e^{V_{\rho_m}})} dt \\
&= \prod_{j=2}^{p-1} \left(\frac{e^{V_{\rho_j}}}{\sum_{m=j}^{p} e^{V_{\rho_m}}} \right) e^{V_{\rho_1}} \left[-\frac{e^{-t(\sum_{m=1}^{p} e^{V_{\rho_m}})}}{\sum_{m=1}^{p} e^{V_{\rho_m}}} \right]_0^{\infty} \\
&= \prod_{j=1}^{p} \left(\frac{e^{V_{\rho_j}}}{\sum_{m=j}^{p} e^{V_{\rho_m}}} \right).
\end{aligned}
$$

The result matches the Equation (3.7) where $V_{\rho_{\cdot j}} = \mu_{\rho_{ij}}$. Therefore, this is the probability of a ranking in the PL model.

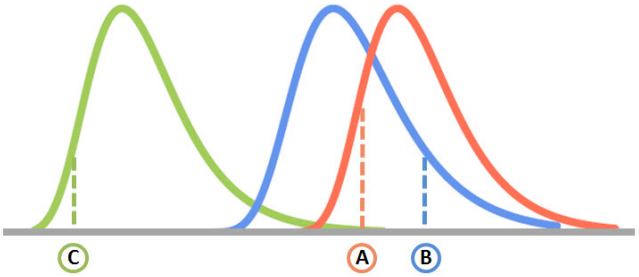


Figure 3.3: Probability density functions for items A, B, and C

As an illustration, Figure 3.3 shows the hypothetical preference distributions of items A, B, and C. Figure 3.3 shows a rank order among three items, A, B, and C, based on a particular sample of values form these distributions which leads to the ranking $B \succ A \succ C$.

3.2.2 EM Algorithm and MM Algorithm

Computation of the ML estimator of the PL model is a problem that already has been considered in the literature from both classical and Bayesian perspectives e.g. Hunter (2004), Guiver and Snelson (2009), Caron and Doucet (2012). The ML estimator can be determined only by numerical methods (Plackett, 1975). In addition, the ML estimator requires the assumption that no item is always ranked first or always ranked last in all comparisons, in order to

prevent the estimates from approaching infinity (Marden, 1995).

The Expectation Maximization (EM) and MM algorithms for fitting the PL model to ranking data, which are proposed by Caron and Doucet (2012) and Hunter (2004), respectively, are typically used. Standard optimization procedures such as the Newton-Raphson method can also be used; however, Hunter (2004) reported that this method is slower and less practical. The Newton-Raphson method is not well-behaved even though the log-likelihood function is strictly concave in the reparameterized parameter space. The possible reasons why the Newton-Raphson method fails are sensitive starting values and over-shooting. The EM and MM algorithms perform well for both complete and partial ranking data. The EM algorithm is a special case of the MM algorithm and both algorithms are guaranteed to converge to the unique maximum of the likelihood function (Hunter, 2004). The algorithms proceed iteratively to find the estimated parameter, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)^\top$, which gives the preference value for each item.

The log-likelihood function can be written as

$$\ell(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \left[\log(\lambda_{\rho_{ij}}) - \log \left(\sum_{m=j}^{p_i} \lambda_{\rho_{im}} \right) \right]. \tag{3.8}$$

Hunter (2004) proposed an MM algorithm for the PL model. The MM algorithm uses surrogate minimizing functions of the log-likelihood function to define an iteration. That is, the optimization is performed on a surrogate function rather than on the log-likelihood itself. Each iteration in the MM algorithm involves two steps as follows:

(1) Minorization step

As in the BT model, Equation (3.3) is also considered here. We apply this to the second term in the log-likelihood. Let $x = \sum_{m=j}^{p_i} \lambda_{\rho_{im}}$ and $y = \sum_{m=j}^{p_i} \lambda^*_{\rho_{im}}$ where $\lambda^*_{\rho_{im}}$ denotes the estimate of $\lambda_{\rho_{im}}$ from the previous

iteration. The inequality in Equation (3.3) becomes

$$-\log\left(\sum_{m=j}^{p_i}\lambda_{\rho_{im}}\right) \geq 1-\log\left(\sum_{m=j}^{p_i}\lambda^*_{\rho_{im}}\right)-\frac{\sum_{m=j}^{p_i}\lambda_{\rho_{im}}}{\sum_{m=j}^{p_i}\lambda^*_{\rho_{im}}},$$

and the surrogate objective function is given by

$$\begin{aligned} Q\left(\boldsymbol{\lambda},\boldsymbol{\lambda}^*\right) &= \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log(\lambda_{\rho_{ij}})+1-\log\left(\sum_{m=j}^{p_i}\lambda^*_{\rho_{im}}\right)-\frac{\sum_{m=j}^{p_i}\lambda_{\rho_{im}}}{\sum_{m=j}^{p_i}\lambda^*_{\rho_{im}}}\right] \\ &\equiv \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log(\lambda_{\rho_{ij}})-\frac{\sum_{m=j}^{p_i}\lambda_{\rho_{im}}}{\sum_{m=j}^{p_i}\lambda^*_{\rho_{im}}}\right] \qquad (3.9) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log(\lambda_{\rho_{ij}})-c^*_{ij}\sum_{m=j}^{p_i}\lambda_{\rho_{im}}\right], \end{aligned}$$

where

$$c^*_{ij}=\frac{1}{\sum_{m=j}^{p_i}\lambda^*_{\rho_{im}}}.$$

Equation (3.9) contains only the terms that depend on elements of $\boldsymbol{\lambda}$.

(2) Maximization step

The maximization of $Q(\boldsymbol{\lambda},\boldsymbol{\lambda}^*)$ with respect to λ_k can be done explicitly since the elements of the parameter vector $\boldsymbol{\lambda}$ are separated. Differentiating the surrogate objective function, $Q(\boldsymbol{\lambda},\boldsymbol{\lambda}^*)$, with respect to λ_k gives

$$\frac{\partial Q}{\partial\lambda_k}=\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\eta_{ijk}-c^*_{ij}\delta_{ijk}\right],$$

where η_{ijk} and δ_{ijk} are indicator functions defined as follows

$$\eta_{ijk}=\begin{cases}1, & \text{if } \rho_{ij}=k\\ 0, & \text{otherwise,}\end{cases}$$

and

$$\delta_{ijk}=\begin{cases}1, & \text{if } k\in\{\rho_{ij},\ldots,\rho_{ip_i}\}\\ 0, & \text{otherwise.}\end{cases}$$

In other words, η_{ijk} is the indicator of the event that item k is not ranked last and δ_{ijk} is the indicator of the event that item k receives a rank no better than j^{th} position by ranker i. Then setting the derivative of the surrogate objective function to zero yields

$$\begin{aligned}\hat{\lambda}_k^{*+1} &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\eta_{ijk}}{\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}c_{ij}^{*}\,\delta_{ijk}} \\ &= \frac{w_k}{\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}c_{ij}^{*}\,\delta_{ijk}} \end{aligned} \tag{3.10}$$

where w_k is the number of rankings in which item k is not ranked last.

Hunter (2004) also proved that the MM algorithm is guaranteed to converge to the unique ML estimator if the following assumption holds:

"Assumption 1. In every possible partition of the items into two nonempty subsets, some item in the second set ranks higher than an item in the first set at least once."

This assumption is an extension of the assumption from Ford (1957) in the BT model. Hunter (2004) concluded that the MM algorithm is guaranteed to converge to the unique ML estimator if Assumption 1 holds.

Algorithm 1 is pseudo code for estimating the PL model, which we have implemented in the R programming language.

Algorithm 1 PLmm algorithm

1: initialize parameter estimates $\boldsymbol{\lambda}^{(0)} = (\frac{1}{K}, \ldots, \frac{1}{K})$ and $h = 0$
2: repeat
3: compute $c_{ij}^{(h)}$ based on $\boldsymbol{\lambda}^{(h)}$
4: increment h
5: compute $\boldsymbol{\lambda}^{(h)}$ by using Equation (3.10)
6: until converged
7: normalize parameters to satisfy $\sum_{i=1}^{K}\lambda_i = 1$

Later, Caron and Doucet (2012) introduced a set of latent variables that enabled the standard EM algorithm to be used to find estimated parameters. Acceleration techniques for the EM algorithm can be applied in order to make

the algorithm converge faster. Moreover, by doing this, it allows the algorithm to work in a Bayesian framework. The two steps of the EM algorithm are as follows:

(1) Expectation step (E-step)

The latent variable is introduced to define the EM and data augmentation step. Let $Z_{\rho_{ij}}$ be an independent variable, exponentially distributed with rate parameter $\lambda_{\rho_{ij}}$ where

$$Z_{i\rho_{i1}} < Z_{i\rho_{i2}} < \cdots < Z_{i\rho_{ip_i}}.$$

Let $Z = \{Z_{i\rho_{ij}} : i = 1, \ldots, n, \rho_{ip_i} \in \mathcal{O}_i\}$ where $\mathcal{O}_i$ is a set of items for ranker i. The latent variables can be introduced as

$$p\left(\boldsymbol{z}|\boldsymbol{\lambda}\right) = \prod_{i=1}^{n}\prod_{j=1}^{p_i-1} Exp\left(z_{ij}; \sum_{m=j}^{p_i}\lambda_{\rho_{im}}\right), \tag{3.11}$$

where Exp is the exponential distribution function. Equation (3.11) is used in implementation of the posterior optimization. The log-likelihood function becomes

$$\begin{aligned}\ell\left(\boldsymbol{\lambda}\right) &= \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log\left(\lambda_{\rho_{ij}}\right) - \log\left(\sum_{m=j}^{p_i}\lambda_{\rho_{im}}\right)\right] \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log\left(\sum_{m=j}^{p_i}\lambda_{\rho_{im}}\right) - \left(\sum_{m=j}^{p_i}\lambda_{\rho_{im}}\right)z_{ij}\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log\left(\lambda_{\rho_{ij}}\right) - \left(\sum_{m=j}^{p_i}\lambda_{\rho_{im}}\right)z_{ij}\right].\end{aligned}$$

The Q function can be constructed by assigning an additional term, the prior for $\boldsymbol{\lambda}$ (Caron and Doucet, 2012),

$$p(\boldsymbol{\lambda}) = \prod_{k=1}^{K} Gamma\left(\lambda_k; a, b\right),$$

where *Gamma* is the gamma distribution with shape and scale parameters, a and b. Caron and Doucet (2012) showed that the Q function is given by

$$\begin{aligned} Q\left(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*\right) &= E_{Z|\lambda^*}\left[\ell(\boldsymbol{\lambda})\right] + \log\left(p(\boldsymbol{\lambda})\right) \\ &\equiv \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left[\log\left(\lambda_{\rho_{ij}}\right) - \frac{\sum_{m=j}^{p_i}\lambda_{\rho_{im}}}{\sum_{m=j}^{p_i}\lambda^*_{\rho_{im}}}\right] + \\ &\quad \sum_{k=1}^{K}\left[(a-1)\log(\lambda_k) - b\lambda_k\right]. \end{aligned}$$

This function is the same as the majorizing function in Hunter (2004) where $a = 1$ and $b = 0$ because the logarithm of the prior term equals 0.

(2) Maximization step (M-step)

Maximizing $Q\left(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*\right)$, where $\boldsymbol{\lambda}^*$ is value from previous iteration, can be done iteratively by

$$\begin{aligned} \hat{\lambda}_k^{*+1} &= \frac{(a-1+w_k)}{\left[b + \sum_{i=1}^{n}\left(\sum_{j=1}^{p_i-1}\frac{\delta_{ijk}}{\sum_{m=j}^{p_i}\lambda^*_{\rho_{im}}}\right)\right]} \\ &= \frac{(a-1+w_k)}{\left[b + \sum_{i=1}^{n}\sum_{j=1}^{p_i-1}\left(c^*_{ij}\delta_{ijk}\right)\right]}, \end{aligned}$$

where c^*_{ij}, δ_{ijk}, and w_k are defined as previous in the MM algorithm.

Caron and Doucet (2012) claimed that the MM algorithm from Hunter (2004) is a special case of the EM algorithm. However, the EM algorithm is known as a special case of the MM algorithm in general. Since the Q functions are the same in both algorithms and $\ell(\boldsymbol{\lambda}^{(h+1)}) \geq \ell(\boldsymbol{\lambda}^{(h)})$, the $\boldsymbol{\lambda}^{(h)}$ is a stationary point of $\ell(\boldsymbol{\lambda}^{(h+1)})$ if $\ell(\boldsymbol{\lambda}^{(h)}) = \ell(\boldsymbol{\lambda}^{(h+1)})$ where h is the current iteration. Moreover, Caron and Doucet (2012) showed that their algorithm can be used even when Assumption 1 is not met by following a Bayesian approach with specific a and b parameters in prior. This is because the items that violate the assumption above have extra information from prior. Therefore,

these estimates do not approach infinity.

Another assumption is mentioned in Hunter (2004) for the strict concavity of the log-likelihood function under the reparameterization. The assumption is:

> "Assumption 2. In every possible partition of the items into two nonempty subsets, some item in the second set is compared with some item in the first set at least once."

Hunter (2004) stated that Assumption 2 is necessary and sufficient for the strict concavity of log-likelihood function under the reparameterization. The log-likelihood function is not *strictly* concave as a function of the original $\boldsymbol{\lambda}$ parameters. The log-likelihood function is strictly concave as a function of the $\boldsymbol{\mu}$'s. Hunter (2004) also suggested to use the reparameterization parameters. The MM algorithm does not change after reparameterization (Hunter, 2004). The expression of the reparameterized parameters is shown in Equation (3.4).

3.2.3 Observed Information Matrix

The ML estimator for the PL model is calculated by maximizing the log-likelihood function

$$\ell(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \sum_{j=1}^{p_i} \left[\log(\lambda_{\rho_{ij}}) - \log \left(\sum_{m=j}^{p_i} \lambda_{\rho_{im}} \right) \right], \tag{3.12}$$

which assumes that the rankings done by different rankers are independent. Thus the formulas below relate to a single ranker. The observed information matrix can be found from summing these expressions over all rankers. The second derivative of the log-likelihood function for each ranker can be found from the first derivative, which is

$$\frac{\partial \ell}{\partial \lambda_{\rho_{ir}}} = \frac{1}{\lambda_{\rho_{ir}}} - \sum_{j=1}^{r} \frac{1}{\sum_{m=j}^{p_i} \lambda_{\rho_{im}}}, \quad r = 1, \ldots, p_i.$$

The second derivative can, therefore, be written as

$$\frac{\partial^2 \ell}{\partial \lambda_{\rho_{ir}}^2} = \frac{1}{-\lambda_{\rho_{ir}}^2} + \sum_{j=1}^{r} \frac{1}{\left(\sum_{m=j}^{p_i} \lambda_{\rho_{im}}\right)^2}, \quad r = 1, \ldots, p_i$$

and, assuming that m is greater than r, the off-diagonal elements of the Hessian matrix can be expressed as

$$\frac{\partial^2 l}{\partial \lambda_{\rho_{ir}} \partial \lambda_{\rho_{im}}} = \sum_{j=1}^{r} \frac{1}{(\sum_{m>j}^{p_i} \lambda_{im})^2}, \qquad m = r+1, \ldots, p_i.$$

The observed information matrix is the negative of the matrix of second derivatives of the log-likelihood function. Moreover, the observed information matrix for the parameters from the PL model is a singular matrix since the parameter preferences, $\boldsymbol{\lambda}$, are only determined up to a multiplicative constant. Therefore, one item, e.g. item 1, should be left out of the information matrix in order to find the inverse and item 1 is considered as baseline. The item 1 is referred to the reference item as in reparameterized parameters. The inverse of the information matrix is the variance-covariance matrix.

The observed information matrix of the reparameterized parameter, μ_k from Equation (3.4), from the log-likelihood function in Equation (3.12) can be found by using chain rule:

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}} = \frac{\partial \ell}{\partial \boldsymbol{\lambda}} \times \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}}$$

then the second derivative is

$$\frac{\partial^2 \ell}{\partial \mu_i \partial \mu_j} = \frac{\partial^2 \ell}{\partial \lambda_k \partial \lambda_r} \times \frac{\partial \lambda_k}{\partial \mu_i} \times \frac{\partial \lambda_r}{\partial \mu_j} + \frac{\partial \ell}{\partial \lambda_k} \times \frac{\partial^2 \lambda_k}{\partial \mu_i \partial \mu_j},$$

where i, j, k, and r are indices of the items. The observed information matrix

is

$$\begin{aligned}\mathbf{J}(\boldsymbol{\mu}) &= -\frac{\partial^2 \ell}{\partial \mu_i \partial \mu_j} \\ &= -\frac{\partial^2 \ell}{\partial \lambda_k \partial \lambda_r} \times \frac{\partial \lambda_k}{\partial \mu_i} \times \frac{\partial \lambda_r}{\partial \mu_j} - \frac{\partial \ell}{\partial \lambda_k} \times \frac{\partial^2 \lambda_k}{\partial \mu_i \partial \mu_j} \\ &= \mathbf{J}(\boldsymbol{\lambda}) \times \frac{\partial \lambda_k}{\partial \mu_i} \times \frac{\partial \lambda_r}{\partial \mu_j} - \frac{\partial \ell}{\partial \lambda_k} \times \frac{\partial^2 \lambda_k}{\partial \mu_i \partial \mu_j}. \end{aligned} \tag{3.13}$$

Note that $\lambda_k = \frac{\exp(\mu_k)}{\sum_{m=1}^{K} \exp(\mu_m)}$ and consider each element in Equation (3.13). The first and the fourth terms can be found as described previously. We consider the other terms. Considering $\frac{\partial \lambda_k}{\partial \mu_i}$ when $i = k$

$$\begin{aligned}\frac{\partial \lambda_k}{\partial \mu_i} &= \frac{\partial \lambda_k}{\partial \mu_k} \\ &= \frac{\sum_{m=1}^{K} \exp(\mu_m)\ \exp(\mu_k) - \exp(\mu_k)\ \exp(\mu_k)}{\left(\sum_{m=1}^{K} \exp(\mu_m)\right)^2} \\ &= \lambda_k - \lambda_k^2,\end{aligned}$$

and when $i \neq k$

$$\begin{aligned}\frac{\partial \lambda_k}{\partial \mu_i} &= \frac{0 - (\exp(\mu_k)\ \exp(\mu_i))}{\left(\sum_{m=1}^{K} \exp(\mu_m)\right)^2} \\ &= -\lambda_k \lambda_i.\end{aligned}$$

The derivative $\frac{\partial \lambda_r}{\partial \mu_j}$ follows in the same way as $\frac{\partial \lambda_k}{\partial \mu_i}$ and the results for $j = r$ and $j \neq r$ are $\lambda_r - \lambda_r^2$ and $-\lambda_r \lambda_j$, respectively. The last term, $\frac{\partial^2 \ell}{\partial \mu_i \partial \mu_j}$, has five possible expressions, depending on which, if any, of i, j, and k are equal.

(1) $i = j = k$

$$\frac{\partial^2 \lambda_k}{\partial \mu_i \partial \mu_j} = \frac{\partial^2 \lambda_k}{\partial \mu_k^2}$$

$$= \frac{\left(\sum_{m=1}^{K} \exp(\mu_m)\right) \exp(\mu_k) - \exp(\mu_k)\exp(\mu_k)}{\left(\sum_{m=1}^{K} \exp(\mu_m)\right)^2} -$$

$$\frac{1}{\left(\sum_{m=1}^{K} \exp(\mu_m)\right)^4} \left[\left(\sum_{m=1}^{K} \exp(\mu_m)\right)^2 2\exp(\mu_k)\exp(\mu_k) - \right.$$

$$\left. (\exp(\mu_k))^2 2 \left(\sum_{m=1}^{K} \exp(\mu_m)\right) \exp(\mu_k) \right]$$

$$= \lambda_k - \lambda_k^2 - \left(2\lambda_k^2 - 2\lambda_k^3\right)$$

$$= \lambda_k (1 - \lambda_k)(1 - 2\lambda_k).$$

(2) $i \neq j \neq k$

$$\frac{\partial^2 \lambda_k}{\partial \mu_i \partial \mu_j} = 0 + \frac{\exp(\mu_k)\exp(\mu_i)\exp(\mu_j) 2 \left(\sum_{m=1}^{K} \exp(\mu_m)\right)}{\left(\sum_{m=1}^{K} \exp(\mu_m)\right)^4}$$

$$= 2\lambda_k \lambda_i \lambda_j.$$

(3) $(i = j) \neq k$

$$\frac{\partial^2 \lambda_k}{\partial \mu_i \partial \mu_j} = \frac{\partial^2 \lambda_k}{\partial \mu_i^2}$$

$$= -\frac{1}{\left(\sum_{m=1}^{K} \exp(\mu_m)\right)^4} \left[\left(\sum_{m=1}^{K} \exp(\mu_m)\right)^2 \exp(\mu_k)\exp(\mu_i) - \right.$$

$$\left. \exp(\mu_k)\exp(\mu_i) 2 \left(\sum_{m=1}^{K} \exp(\mu_m)\right) \exp(\mu_i) \right]$$

$$= -\lambda_k \lambda_i + 2\lambda_k \lambda_i^2$$

$$= -\lambda_k \lambda_i (1 - 2\lambda_i).$$

(4) $(i = k) \neq j$

$$\frac{\partial^2 \lambda_k}{\partial \mu_k \partial \mu_j} = 0 - \frac{\exp(\mu_k)\exp(\mu_j)}{\left(\sum_{m=1}^{K} \exp(\mu_m)\right)^2} - \frac{1}{\left(\sum_{m=1}^{K} \exp(\mu_m)\right)^4} \left[0 \quad - \right.$$

$$\exp(\mu_k)\exp(\mu_k)2\left(\sum_{m=1}^{K}\exp(\mu_m)\right)\exp(\mu_j)\Bigg]$$
$$= -\lambda_k\lambda_j + 2\lambda_k^2\lambda_j$$
$$= -\lambda_k\lambda_j(1-2\lambda_k).$$

(5) $(j=k)\neq i$

$$\frac{\partial^2\lambda_k}{\partial\mu_k\partial\mu_i} = 0 - \frac{\exp(\mu_k)\exp(\mu_i)}{\left(\sum_{m=1}^{K}\exp(\mu_m)\right)^2} - \frac{1}{\left(\sum_{m=1}^{K}\exp(\mu_m)\right)^4}\Bigg[0 \quad -$$
$$\exp(\mu_k)\exp(\mu_k)2\left(\sum_{m=1}^{K}\exp(\mu_m)\right)\exp(\mu_i)\Bigg]$$
$$= -\lambda_k\lambda_i + 2\lambda_k^2\lambda_i$$
$$= -\lambda_k\lambda_i(1-2\lambda_k).$$

The Expression (3), (4), and (5) have the same form. They differ by index values.

Another way to find the observed information matrix of the reparameterized parameters is to find it directly from Equation (3.7) then the log-likelihood becomes

$$\ell(\boldsymbol{\mu}) = \sum_{j=1}^{p_i}\left[\mu_{\rho_{ij}} - \log\left(\sum_{m=j}^{p_i}\exp(\mu_{\rho_{im}})\right)\right]. \tag{3.14}$$

Similarly to the derivation of $\mathbf{J}(\boldsymbol{\lambda})$, we consider only a single ranker to find the first and second derivatives of Equation (3.14). The first derivative is:

$$\frac{\partial\ell}{\partial\mu_{\rho_{ir}}} = 1 - \sum_{j=1}^{r}\frac{\exp(\mu_{\rho_{ir}})}{\sum_{m=j}^{p_i}\exp(\mu_{\rho_{im}})}, \quad r = 1,\ldots,p_i$$

Then the second derivative is

$$\frac{\partial^2\ell}{\partial\mu_{\rho_{ir}}^2} = -\sum_{j=1}^{r}\frac{\left(\sum_{m=j}^{p_i}\exp(\mu_{\rho_{im}})\right)\exp(\mu_{\rho_{ir}}) - (\exp(\mu_{\rho_{ir}}))^2}{\left(\sum_{m=j}^{p_i}\exp(\mu_{\rho_{im}})\right)^2}, \quad r = 1,\ldots,p_i$$

and the off-diagonal elements are

$$\frac{\partial^2 \ell}{\partial \mu_{\rho_{ir}} \partial \mu_{\rho_{it}}} = \sum_{j=1}^{r} \frac{\exp(\mu_{\rho_{ir}}) \exp(\mu_{\rho_{it}})}{\left(\sum_{m=j}^{p_i} \exp(\mu_{\rho_{im}})\right)^2}, \quad t = r+1, \ldots, p_i$$

3.3 Packages in R for the Bradley-Terry and the Plackett-Luce Models

The standard models for analyzing partial ranking data are the BT and the PL models. Both the BT and the PL models have existing packages in the R programming language. In this section, we compare the computational times and efficiency of the existing packages with our algorithms. We implement code to compute the observed information matrix and then compare results with the `optim` function. All experiments were conducted on a Toshiba notebook with Intel Core i5-3210M and 8 GB RAM in the R programming language.

3.3.1 Packages in R for the Bradley-Terry Model

The BT model is included in several existing packages in R, including `prefmod`, `RankResponse`, and `BradleyTerry2`. Among them, the `BradleyTerry2` package, which is implemented by Turner and Firth (2012), appears to be the most widely used package. This package is an extension of the earlier package `BradleyTerry` (Firth, 2008). The `BTm` function from the `BradleyTerry2` package can also incorporate covariates. Another algorithm, `BTmm`, we coded in R, based on the original code in Matlab by Caron and Doucet (2012). Since the BT model is a special case of the PL model when there are only two items, any algorithm for fitting the PL model can apply to the paired data. We translated Matlab code of Caron and Doucet (2012) for `PLem` and we also implemented the `PLmm`. Both are based on the Algorithm 1; however, they are different in matrix structure. We ran these four algorithms with the same simulated data

in order to compare computational time and mean square error (MSE) of the parameter estimates.

We use simulations to assess the performance for both small and large datasets. The small dataset is simulated with 10 items ($K = 10$) and this dataset contains 200 pairwise comparisons ($n = 200$). The large dataset contains 100 items ($K = 100$) with 10,000 pair comparisons ($n = 10,000$), to make sure that the algorithms have enough data to make the BT model converges.

Table 3.1: Computational times of `BradleyTerry2`, `BTmm`, `PLem`, and `PLmm`

	Time (seconds)			
	`BradleyTerry2`	`BTmm`	`PLem`	`PLmm`
Small data	0.01	2.92	0.42	0.03
Large data	2.85	9.67	21.34	5.78

Table 3.1 shows the computational times (in seconds) of the four algorithms. The `BradleyTerry2` algorithm is the fastest, followed by the `PLmm`, `PLem`, and `BTmm` for the small data. Moreover, the `BradleyTerry2` is also the fastest algorithm for the large data, while the `PLem` is a lot slower. The `BradleyTerry2` package uses the `glm` function in `stats` package. The reason why the `PLem` performs poorly here is that the `PLem` is for the PL model; therefore, it has unnecessary loops for paired data. Moreover, `PLmm` is based on another way of introducing matrices of data and ran faster than `PLem` in both cases.

To evaluate these four algorithms, the MSE is calculated in order to compare the estimated parameters ($\hat{\boldsymbol{\lambda}}$) with true parameter values. The results are given in Table 3.2. Table 3.2 indicates that all algorithms give almost the same MSE. The computational time is therefore considered as main criterion. The `BradleyTerry2` is the best according to computational times for both small and large datasets; however, we have to transform the datasets into a specific format. This is computationally demanding when K increases. Therefore,

Table 3.2: MSE of `BradleyTerry2`, `BTmm`, and `PLmm`

	MSE			
	`BradleyTerry2`	`BTmm`	`PLem`	`PLmm`
Small data	6.913e-03	6.912e-03	6.913e-03	7.027e-03
Large data	6.779e-06	6.729e-06	6.779e-06	6.855e-06

the `PLmm` is chosen to fit the paired dataset because it is the second fastest algorithm and it is more convenient in terms of data format.

3.3.2 Packages in R for the Plackett-Luce Model

The PL model is not implemented in many existing packages. Lee and Yu (2013) developed the `pmr` package; however, the `pl` function in this package only works for full ranking data. Later, Chen and Soufiani (2013) implemented the `StatRank` package. The `Estimation.PL.MLE` function works for both full and partial ranking data. This function is for the PL model with a flipped Gumbel distribution. However, the flipped Gumbel distribution can be seen as a Gumbel distribution where the smaller λ is more preferred. The `Estimation.PL.MLE` is forced to work for a specific number of iterations. The default number of iterations is 10. This means that the algorithm is forced to work even though it already converged in less than 10 iterations. This is a big drawback because the algorithm will always work up to this number of iterations. Moreover, for a large number of items, it is hard for the algorithm to converge within 10 iterations, but we can change this default number. Thus, the `Estimation.PL.MLE` does not check the convergence.

The computational times, which are shown in Table 3.3, are computed by applying the `Estimation.PL.MLE`, the `PLem`, and the `PLmm` on both synthetic and real world datasets. We simulate two kinds of dataset as in the previous section. First, for a small number of items, we generated 200 rankings with $K = 20$ and each ranker ranking a random sample of $p = 5$ items. The

second data contain 500 rankings with $K = 100$ and $p = 10$ items. The real dataset, NASCAR is also used since this dataset is provided in the `StatRank` package. Moreover, Hunter (2004) used this dataset as a test dataset as well. The NASCAR dataset contains records of auto racing in the United States. In the 2002 season, there were 87 drivers who participated in 36 races. However, Assumption 1 is violated and therefore we removed four drivers and then there are 83 drivers and 36 races in the dataset. We set the number of iterations in the `Estimation.PL.MLE` equal to the number of iterations that the `PLem` algorithm used.

Table 3.3: Computational times of `StatRank`, `PLem` and `PLmm`

	Time (seconds)		
	`StatRank`	`PLem`	`PLmm`
Small data	14.54	0.42	0.48
Large data	415.72	1.37	15.34
NASCAR	4295.53	0.11	12.52

Table 3.3 shows that the `PLem` algorithm performs much faster than the `StatRank` package. `StatRank` does not take much computational time for the small dataset but still much slower than alternatives. Moreover, when the dataset becomes large, the `StatRank` package works even slower than the others. `StatRank` suffers from large computational time because the function contains a lot of loops while the `PLem` and the `PLmm` algorithms work in matrix form. The `PLem` and the `PLmm` algorithms are different in data structure. The `PLem` algorithm and `PLmm` algorithm use almost the same computational times for the small dataset. When the dataset becomes larger the `PLem` algorithm is faster than the `PLmm` algorithm. Because of this clear difference in performance, the `PLem` algorithm is used when covariates are not considered in later sections. The `PLmm` algorithm will be extended in Chapter 5 since it is easier to introduce covariates to the algorithm.

We mentioned about the `pmr` package for full ranking earlier. We also

Table 3.4: Computational times of `StatRank`, `pmr`, `PLem`, and `PLmm`

	Time (seconds)			
	`StatRank`	`pmr`	`PLem`	`PLmm`
Small data	0.83	60.04	0.52	0.58

compare these three algorithms with the `pl` function from that package. A dataset with 500 rankings and $K = p = 6$ is generated. The computational times are shown in Table 3.4. The `PLem` and `PLmm` algorithms are the fastest, followed by the `StatRank` package while the `pmr` package is by far the worst among them. The `pmr` package uses the `optim` function; however, the log-likelihood code contains double loops. This is likely to be the reason why the `pmr` package is slow.

3.3.3 The `optim` Function in R and Our Algorithm for Computing the Observed Information Matrix

As far as we know, there is no existing package in R which can calculate the observed information matrix for the PL model with partial ranking data. We implemented the `PLinfm` algorithm. Standard errors are computed from the Hessian matrix from the `optim` function when Hessian option is set to TRUE and the `PLinfm` algorithm. Results are shown in Figure 3.4. Both algorithms

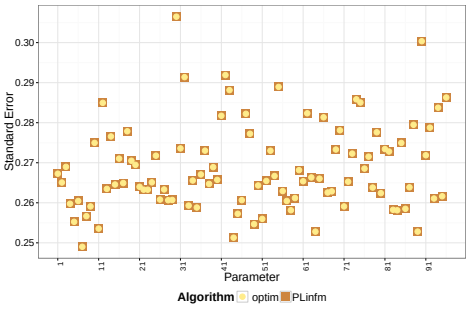


Figure 3.4: Adjacent ranking method

give the same standard error of each parameter. We use the `PLinfm` algorithm to calculate the observed information matrix for the $\hat{\boldsymbol{\mu}}$ later in this chapter.

3.4 Results of Fitting the PL Model

The PL model is applied to the Group I data from the Animal dataset in order to estimate the preference of each animal species, λ_k. Each participant was asked to rank the given animal images according to his/her preferences in descending order. The reparameterized parameters ($\hat{\boldsymbol{\mu}}$) and their standard errors were computed by using Baji as the reference species and hence $\hat{\mu}_{\text{Baji}} = 0$.

Table 3.5: Top five and bottom five values according to $\hat{\mu}_k$ when fitting the PL model to the Group I data from the Animal dataset

Animal Species	Contests	Average Rank	$\hat{\lambda}_k$	$\hat{\mu}_k$	SE($\hat{\mu}_k$)
Red Panda	50	2.200	0.057	1.959	0.278
Giant Panda	60	2.367	0.046	1.745	0.265
African Elephant	37	2.973	0.031	1.364	0.286
Fin Whale	42	3.143	0.029	1.294	0.275
Asian Elephant	63	3.000	0.028	1.245	0.254
⋮					
Mindanao Gymnure	42	8.000	0.003	-1.052	0.282
Eastern Sucker-footed Bat	40	7.175	0.003	-1.182	0.301
Chiapan Climbing Rat	43	8.116	0.002	-1.226	0.284
New Guinea Big-eared Bat	64	7.609	0.002	-1.252	0.265
Southern Marsupial Mole	39	7.641	0.002	-1.359	0.307

Table 3.5 shows the number of contests, average place, estimated parameter ($\hat{\lambda}_k$), reparameterized parameter ($\hat{\mu}_k$), and standard error of reparameterized parameter (SE($\hat{\mu}_k$)). Considering, for example, the Red Panda; its picture is ranked by 50 rankers and the average place is 2.2 which is the highest average place. The reparameterized parameter, $\hat{\mu}_{\text{Red Panda}}$, is 1.959 and it can be interpreted that Red Panda has substantially higher preference than Baji. Moreover, the estimated parameter, $\hat{\lambda}_{\text{Red Panda}}$, has the highest value which is 0.057. This means that Red Panda is the most appealing animal species among 97 animal species in Group I data.

It is clear from Table 3.5 that the preference order according to average

(a) Red Panda

(b) Baji

(c) Southern Marsupial Mole

Figure 3.5: Pictures of Red Panda, Baji, and Southern Marsupial Mole

place is not the same as the order based on $\hat{\mu}$. This is because the average place does not consider the effect of the partial ranking. For example, if animals are ordered by average place then Asian Elephant has higher preference than Fin Whale since the scores are 3 and 3.143, respectively while ordering according to $\hat{\mu}$, the result is the other way around.

The 95% confidence intervals for $\hat{\mu}$ are plotted in Figure 3.6. Figure 3.6 indicates that Red Panda and Giant Panda are the most preferred species but their 95% confidence intervals overlap the next six species which are African Elephant, Fin Whale, Asian Elephant, Vaquita, Blue Whale, and Amazonian Manatee. However, the overlaps between these next six species are less than the overlap between Red Panda and Giant Panda. There is a small gap between the least four favorite species and the species ranked above, these species are Eastern Sucker-Bat, Chiapan Climbing Rat, New Guinea Big-eared Bat, and Southern Marsupial Mole. Moreover, two of the highest standard errors are in the last four species. The SE($\hat{\mu}$) of Southern Marsupial Mole has the largest value and is followed by the Eastern Sucker-footed Bat which are 0.307 and 0.301, respectively.

3.4.1 Goodness-of-Fit for the PL Model for the Group I Data from the Animal Dataset

We perform a bootstrap goodness-of-fit test in order to check whether the Group I data from the Animal data can be fitted by the PL model. The

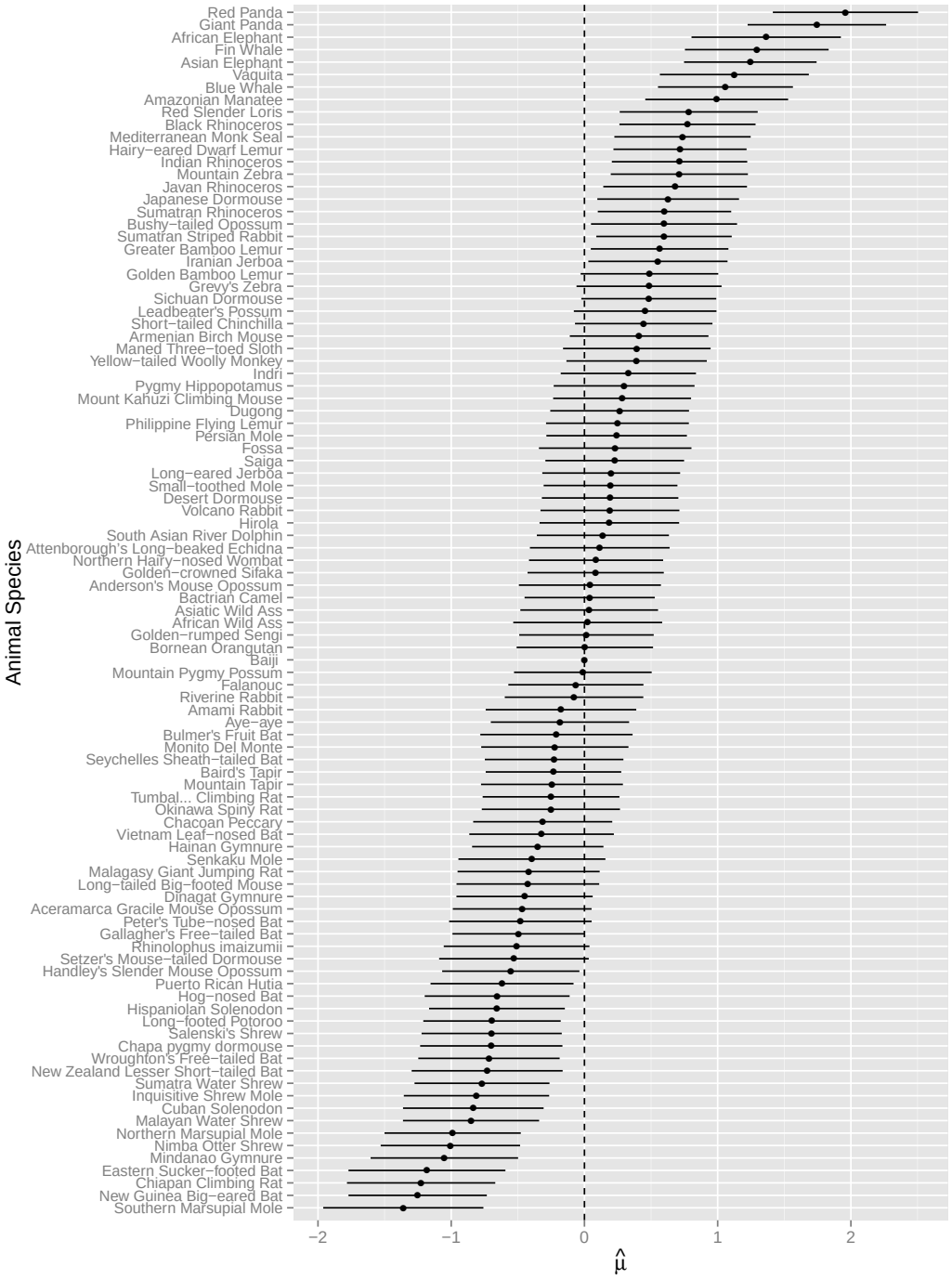


Figure 3.6: The 95 % confidence interval of $\hat{\mu}$ for the Group I data from the Animal Dataset

bootstrap sample is 10000 ($B = 10000$).

Figure 3.7 shows the Kendall tau distance from the bootstrapping. The two-sided p-value is 0.072. We conclude that the PL model is an appropriate model for fitting the Group I data from the Animal dataset at 5% significance

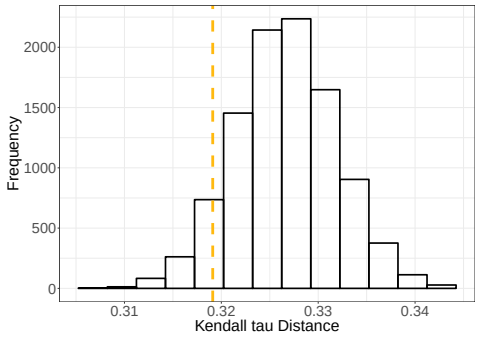


Figure 3.7: Histogram of Kendall tau distance from the bootstrapping goodness-of-fit for the PL model where dashed line is the Kendall tau distance from the Group I data

level.

The IOS value is approaching to zero. This leads us to compute the two-sided p-value. The two-sided p-value is 0.264 and this gives the same conclusion as the Kendall tau distance that the PL model is a suitable model for fitting the Group I data.

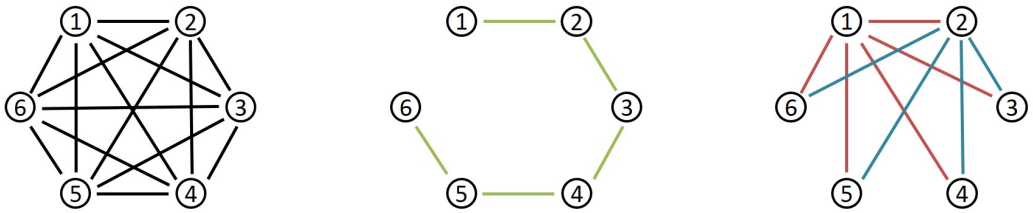
3.5 Rank Breaking

Another approach to deal with partial rankings is to break them into pairwise comparisons and treat the pairs as if they were independent. Soufiani and Parkes (2014) proposed a *rank-breaking* methodology that breaks the full ranking into a subset of pairwise comparisons according to rank positions. The BT model, which is less complicated than the PL model, can then be used instead of the PL model. We apply the rank-breaking method to partial ranking data and compare results of fitting the PL model to the original ranking data and fitting the BT model to a collection of pairs generated by rank-breaking.

3.5.1 Three Methods of Rank Breaking

Soufiani and Parkes (2014) presented a rank-breaking by using an undirected breaking graph. The nodes in the graph are ranked positions. Soufiani et al.

(2013a) proposed five methods of rank breaking. In this thesis, we consider only three methods which are full rank-breaking, adjacent rank-breaking, and top-h rank-breaking as shown in Figure 3.8.



(a) Full breaking (b) Adjacent breaking (c) Top-2 breaking

Figure 3.8: Three methods of rank-breaking when $p = 6$

1 Full Rank-Breaking

The full rank-breaking method considers all possible paired comparisons; therefore, the number of pairwise comparisons generated is $\binom{p_i}{2}$ where p_i is the number of items for ranker i to rank.

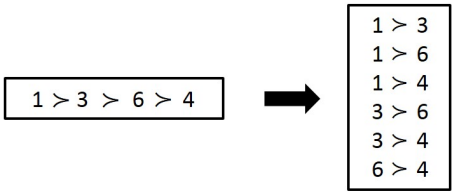


Figure 3.9: Full ranking method

Figure 3.9 shows an example of full rank-breaking method for a ranking of 4 items. Thus, the original ranking generates 6 pairwise comparisons.

2 Adjacent Rank-Breaking

The adjacent rank-breaking method considers only adjacent positions. The number of pair comparisons after applying the adjacent method to the partial ranking is $p_i - 1$ pairs. An example of applying the adjacent rank-breaking method to a ranking set is given in Figure 3.10.

In other words, pairwise comparisons come from

$$\{\rho_{i1}, \rho_{i2}\}, \{\rho_{i2}, \rho_{i3}\}, \ldots, \{\rho_{i(p_1-1)}, \rho_{\rho_{i,p_i}}\}.$$

Figure 3.10: Adjacent ranking method

3 Top-h Rank-Breaking

The position-h rank-breaking method is a special case of the full rank-breaking method (when $h = p_i$) but this breaking considers only items ranked up to h^{th} position compared to items in lower positions. The value of h can be any number between 1 and p_i. For example, in Figure

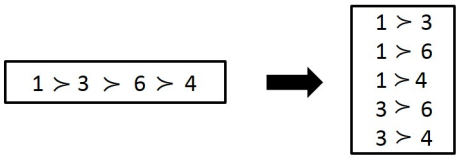


Figure 3.11: Top-h ranking method

3.11, if $h = 2$ then the Top-2 rank-breaking method is applied. The ranking gives five pairwise comparisons.

3.5.2 Rank Breaking with Unequal Weights

Later, Khetan and Oh (2016) extended Soufiani and Parkes's (2014) idea by applying weightings to the pairs from the full rank-breaking method. These pairwise comparisons can come from either full or partial ranking data. Here, again the pairwise comparisons were treated as if they are independent. They used a directed acyclic graph (DAG) to present a partial ordering and let $\mathcal{G}_i$ denote the DAG of an ordering from ranker i. The term "separator" is introduced. A separator is a node that can partition the rest of the nodes into two parts. Let a be a set of separator items. Moreover, let A_{top} be the set of items that are ranked higher than the separator item and A_{bottom} be the set of items that are ranked lower than the separator item where A_{top} is allowed to be the empty set but A_{bottom} cannot be empty. For example, given an ordering

set {1, 3, 6, 4} from ranker 1, the $\mathcal{G}_1$ is shown in Figure 3.12. There are 3

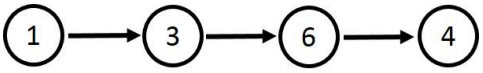
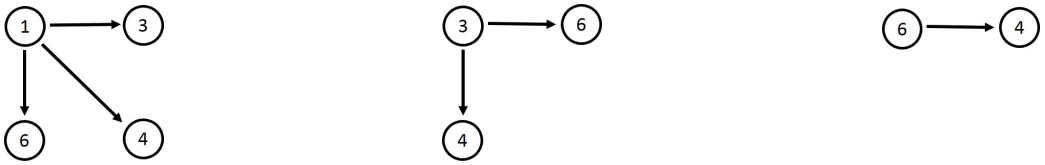


Figure 3.12: The $\mathcal{G}_1$ for the ordering {1, 3, 6, 4}

separators which are $a = \{1, 3, 6\}$ in this example.

Let l_i denote the number of separators and G_{i,a_j} denote the rank-breaking graph for ranking i with separator a_j where $j = 1, \ldots, l_i$. From the previous example then $l_1 = 3$ and $G_{1,\cdot}$ is given in Figure 3.13. Thus, the number of



(a) $G_{1,1}$ (b) $G_{1,3}$ (c) $G_{1,6}$

Figure 3.13: Rank-breaking graphs ($G_{1,\cdot}$) from the full rank-breaking method for $\mathcal{G}_1$

rank-breaking graphs for ranker i is equal to l_i rank-breaking graphs. In our case, our datasets do not have tied ranking, therefore, $l_i = p_i - 1$. We can rewrite G_{i,a_j} as $G_{i,\rho_{ij}}$ where $j = 1, \ldots, p_i - 1$.

Khetan and Oh (2016) suggested that the pairwise comparisons should not have the same weighting. If the equal weighting is given, it means the dependencies in the original data are ignored. They also suggested that all pairwise comparisons in each $G_{i,\rho_{ij}}$ have the same weighting; however, the weights differ between the $G_{i,\rho_{ij}}$. Let $w_{i,\rho_{ij}}$ be weight for the $G_{i,\rho_{ij}}$ and $w_{i,\rho_{ij}}$ has a positive value. The $w_{i,\rho_{ij}}$ is computed from

$$w_{i,\rho_{ij}} = \frac{1}{l_i - j + 1}. \tag{3.15}$$

The log-likelihood function of the BT model with these weightings becomes

$$\ell\left(\boldsymbol{\mu}\right)=\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}w_{i,\rho_{ij}}\left[\sum_{k,k'\in G_{i,\rho_{ij}}}\left(\mu_k-\log\left(\exp\left(\mu_k\right)+\exp\left(\mu_{k'}\right)\right)\right)\right],$$

where item k is preferred to item k'. The estimates can be found via the MM algorithm as for the BT model with equal weightings then

$$\hat{\mu}_k=\frac{\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}w_{i,\rho_{ij}}\sum_{k\succ k'\in G_{i,\rho_{ij}}}\eta_{k\in G_{i,\rho_{ij}}}}{\sum_{i=1}^{n}\sum_{j=1}^{p_i-1}w_{i,\rho_{ij}}\sum_{k,k'\in G_{i,\rho_{ij}}}\frac{\eta_{k,k'\in G_{i,\rho_{ij}}}}{\exp(\mu_k^*)+\exp(\mu_{k'}^*)}},$$

where $\eta_{k\in G_{i,a}}$ is an indicator function such that

$$\eta_{k\in G_{i,\rho_{ij}}}=\begin{cases}1, & \text{if } k\in G_{i,\rho_{ij}}\\ 0, & \text{otherwise.}\end{cases}$$

Observed Information Matrix

The observed information matrix of the reparameterized parameters is the negative of the Hessian matrix and it is a positive semi-definite matrix. The Hessian matrix can be found by using the second derivative of the log-likelihood function. Summing these over all rankers will obtain the Hessian matrix. The first derivative is:

$$\frac{\partial\ell}{\partial\mu_k}=\sum_{j=1}^{p_i-1}w_{i,\rho_{ij}}\left[\sum_{k,k'\in G_{i,\rho_{ij}}}\left(1-\frac{\exp(\mu_k)}{\exp(\mu_k)+\exp(\mu_{k'})}\right)\right],$$

then the second derivative is:

$$\frac{\partial^2\ell}{\partial\mu_k^2}=\sum_{j=1}^{p_i-1}w_{i,\rho_{ij}}\left[\sum_{k,k'\in G_{i,\rho_{ij}}}-\frac{\exp(\mu_k+\mu_{k'})}{\left(\exp(\mu_k)+\exp(\mu_{k'})\right)^2}\right],$$

and since $k \succ k'$ then

$$\frac{\partial^2 \ell}{\partial \mu_k \partial \mu_{k'}} = \sum_{j=1}^{p_i - 1} w_{i,\rho_{ij}} \left[\sum_{k,k' \in G_{i,a}} \frac{\exp(\mu_k + \mu_{k'})}{(\exp(\mu_k) + \exp(\mu_{k'}))^2} \right].$$

This is Hessian matrix and the observed information matrix is the negative of the Hessian matrix.

3.5.3 Consistency

Inappropriate use of the rank-breaking method can introduce bias and lead to inconsistent estimators. Soufiani and Parkes (2014) developed a general condition to determine whether the rank-breaking will provide a consistent estimator for the PL model for complete ranking data. Soufiani et al. (2013a) and Soufiani and Parkes (2014) proved that a rank-breaking graph is consistent if and only if it satisfies this property. Recall that the nodes in a rank-breaking graph represent positions in the original ranking. The property is that if a node a is connected to any node b where $b > a$, then node a must be connected to all the nodes c when $c > a$. Following this property, the full rank-breaking and top-h rank-breaking methods are consistent for the PL model but the adjacent rank-breaking is inconsistent for the PL model. Thus, the adjacent rank-breaking will not provide good estimates of the PL model.

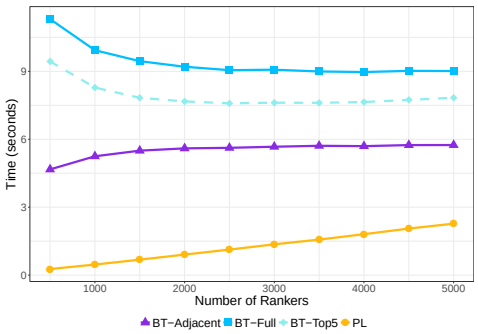
In the recent work of Khetan and Oh (2016), they applied the full rank-breaking method to partial ranking sets. Khetan and Oh (2016) used the proof from Soufiani et al. (2013a) and Soufiani and Parkes (2014). The rank-breaking graph $G_{i,\rho_{ij}}$ has a separator node which is connected to all other items in the graph that are ranked below it. Thus, the Khetan and Oh's (2016) rank-breaking graph satisfies the property. Khetan and Oh (2016) stated that all pairwise outcomes, in the rank-breaking graphs, give a consistent estimator when all pairwise comparisons in each $G_{i,\rho_{ij}}$ have the same weighing.

Weightings between rank-breaking graphs can be different. They showed that the full rank-breaking method with unequal weighting for each rank-breaking graph is consistent.

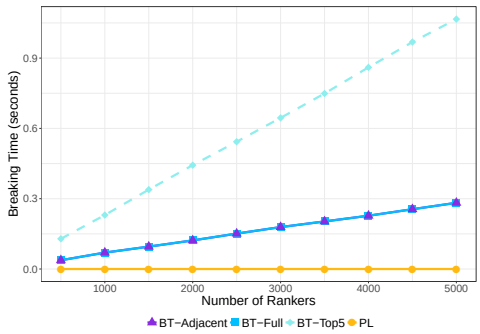
3.5.4 Experimental Results

Synthetic Data: Compare Rank-Breaking Methods in Section 3.5.1

Synthetic data were generated as follows: Suppose that there are $K = 100$ items. We generate n partial rankings with $p = 10$ according to the PL model with true parameter values, $\boldsymbol{\lambda}$. Then for each ranking, we apply the three methods of rank-breaking from Soufiani and Parkes (2014). The `PLmm` algorithm is applied to the pairwise comparisons due to the data format and to the original synthetic data samples of size 500 to 5000 rankers, in steps of 500. This process is repeated 500 times and the average computation time is shown in Figure 3.14 while the average of the Kendall tau correlation and the average of the MSE are presented in Figure 3.15.



(a) Computational time of breaking into pairs and fitting the BT and the PL models to synthetic data



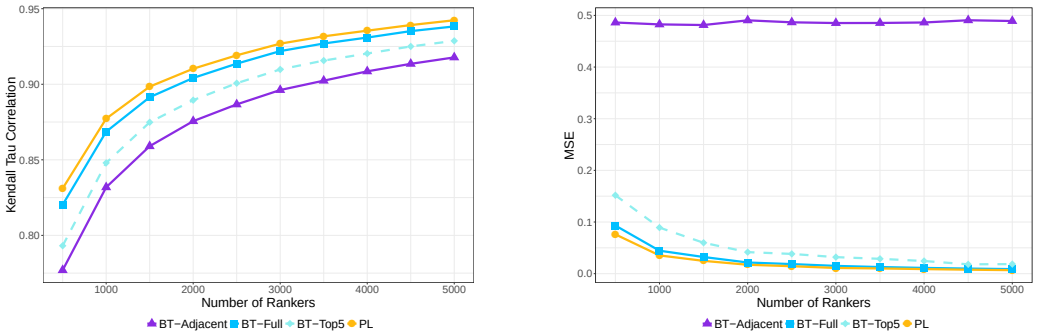
(b) Computational time for breaking synthetic data into pairs

Figure 3.14: Computational time of breaking into pairs and fitting the BT and the PL models to synthetic data

Figure 3.14a shows the overall computational times of breaking into pairs and fitting the BT and the PL models on synthetic data. The computational times of the three rank-breaking methods are presented in Figure 3.14b. Figure

3.14a illustrates that the PL and BT-Adjacent (adjacent rank-breaking) take more time when the number of rankers becomes higher, while for the other two method, BT-Full (full rank-breaking) and BT-Top5 (top-5 rank-breaking), the running time decreases with increasing in n. This is because the algorithms need more iterations to converge when the number of rankers is small for BT-Full and BT-Top5. The PL is the fastest method among them.

All rank-breaking methods need more computational time for breaking the data into pairs as the number of rankers increases as presented in 3.14b. We use the `Breaking` function from the `StatRank` package in order to decompose into pairwise comparisons. The BT-Top5 is the slowest method while the adjacent and full rank-breakings use almost the same amount of time. The breaking time for PL remains zero for all sample sizes since the PLmm algorithm is applied to the original data.



(a) Kendall Tau Correlation criterion

(b) MSE criterion

Figure 3.15: The average of Kendall tau correlation and average of MSE criteria when applied the PL model to original synthetic data and the BT model to pairwise data from BT-Full (full breaking), BT-Adjacent (adjcent breaking), and BT-Top5 (top-5 breaking)

The MSE and Kendall tau correlation are employed to evaluate and compare the fitting results, as shown in Figure 3.15. Both figures give the same conclusion that the accuracy of all the methods improves as number of rankers increases in which the PL performs best and is followed by BT-Full, BT-Top5 and BT-Adjacent. It is clear in Figure 3.15b that the BT-Adjacent does not perform well under the MSE criterion. The BT-Full performs almost as well

as PL as the number of rankers increases. However, the efficiency of Top-h rank-breaking method depends upon the h value.

The paired t-test is applied to evaluate whether the averages of Kendall tau correlation from the PL model and the BT model with full rank-breaking method are different. The differences between these are significant at 1% level which confirms that the PL model performs better than the BT model with the full rank-breaking.

Moreover, adjacent breaking provides poor estimates even when compared to the other methods as the data size grows. This reflects the fact that the adjacent rank-breaking estimator is inconsistent.

Synthetic Data: Full Rank-Breaking with Equal and Unequal Weights

The synthetic data were generated as previously stated in order to compare the performance of different weightings. Only the full rank-breaking method is used to break the synthetic data into pairwise comparisons. The `PLmm` algorithm is applied to paired data and original synthetic data. One of the sets of unequal weights is from Khetan and Oh (2016). We explore other weightings by extending Equation (3.15). The reasons why we explore other weightings are (1) to confirm empirically that the original weighting was optimal for MSE as mentioned by Khetan and Oh (2016) and (2) to see if it was also optimal for other criteria (Kendall tau correlation and logarithm of determinant of the observed information matrix) or whether alternative weightings might do better. The different powers, which are shown in Table 3.6, are taken of the weight in Equation (3.15). This is an empirical idea to see how these weightings affect the estimates from the BT model. We rescale these new weights by introducing a constant value,

$$c_i = \sum_{j=1}^{p_i-1} w_{i,\rho_{ij}} \, (l_i - j - 1),$$

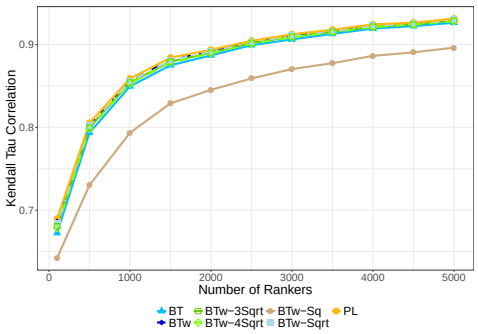
Table 3.6: Unequal weights from Equation (3.15) for full rank-breaking pairs

Weighting method	Power	Name
(1) -	1	BTw
(2) Square root	$\frac{1}{2}$	BTw-Sqrt
(3) Cube root	$\frac{1}{3}$	BTw-3Sqrt
(4) 4^{th} root	$\frac{1}{4}$	BTw-4Sqrt
(5) Square (Sq)	2	BTw-Sq

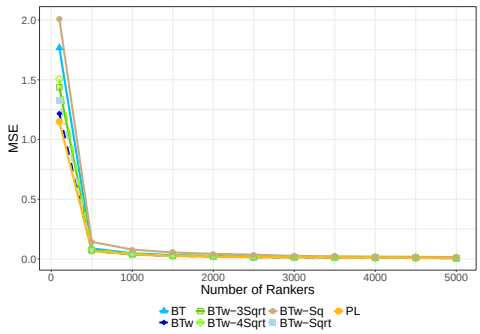
then the new weight is

$$w'_{i,\rho_{ij}} = (p_i - 1)\frac{w_{i,\rho_{ij}}}{c_i}.$$

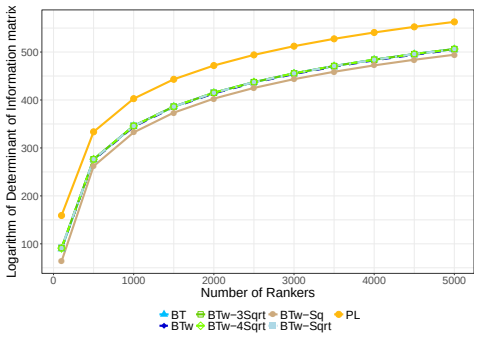
The $w'_{i,\rho_{ij}}$ is used instead of $w_{i,\rho_{ij}}$ in analysis.



(a) Kendall Tau Correlation criterion



(b) MSE criterion



(c) Logarithm of Determinant of Information matrix criterion

Figure 3.16: The average of Kendall tau correlation, MSE, and logarithm of the determinant of the observed information matrix criteria when applied the PL model to original synthetic data and the BT model to pairwise data with BT, BTw, BTw-Sqrt, BTw-3Sqrt, BTw-4Sqrt and BTw-Sq weightings from Table 3.6, respectively

The averages of Kendall tau correlation, MSE, and the logarithm of de-

terminant of the observed information matrix are shown in Figure 3.16 where BT is the BT model with equal weight and the BTw stands for the BT model with different weights.

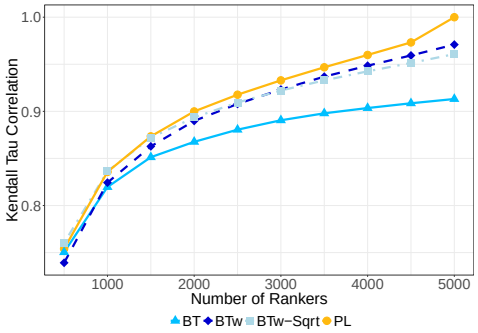
Figure 3.16a presents that the BT model with square weighting (BTw-Sq) performs the worst while the other methods are comparable to the PL model. The MSE criterion gives almost the same conclusion as shown in Figure 3.16b. When number of rankers is small (n=100), it is obvious that the BTw performs best and is followed by BTw-Sqrt, BTw-3Sqrt, BTw-4Sqrt, BT, and BTw-Sq. However, all methods give almost the same results when number of rankers is large. The logarithm of the determinant of the observed information shows that the PL model is the best among all methods while the BT model with weightings, except BTw-Sq, and equal weightings yield the same performance, while the BTw-Sq gives the poorest performance.

Sushi Dataset: Full Rank Breaking with Equal and Unequal Weights

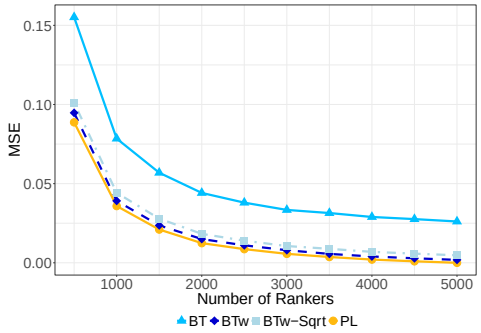
The Sushi dataset is considered here. We recall that there are 5000 participants who ranked sushi flavours where $K = 100$, and $p = 10$. The 'true' parameter values are obtained by applying the `PLmm` algorithm to all 5000 ranking sets.

In order to compare performances of different weightings, the three evaluating criteria used previously are considered. The estimates are calculated after every 500 rankers from 500 to 5000 rankers. We randomly choose 500 rankers from the original Sushi dataset at each point with replacement. This process is repeated 200 times after that we calculate average of correlation, MSE, and logarithm of determinant of the observed information matrix values.

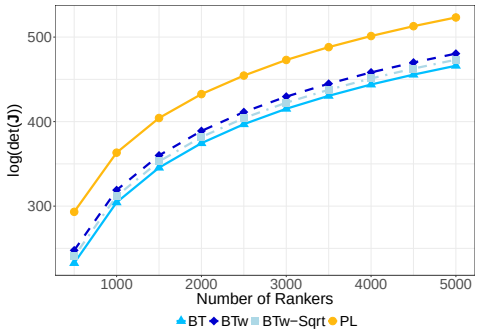
We compare only three weightings which are BT, BTw, and BTw-Sqrt. The other weightings are excluded as the BTw-3Sqrt and BTw-4Sqrt give almost the same results as BTw-Sqrt, and BTw-Sq performed poorly with the synthetic data. Results are presented in Figure 3.17 where the BTw and BTw-Sqrt are weighting (1) and (2) in Table 3.6, respectively. Both MSE and the



(a) Kendall Tau Correlation criterion



(b) MSE criterion



(c) Logarithm of Determinant of Information matrix criterion

Figure 3.17: The average of Kendall tau correlation, MSE, and logarithm of the determinant of the observed information matrix criteria when applied the PL model to Sushi dataset and the BT model to full rank-breaking data with BTw and BTw-Sqrt weightings in Table 3.6

logarithm of the determinant of the observed information matrix criteria reveal the same conclusion while the correlation criterion gives a different conclusion at the beginning. The MSE and logarithm of the determinant of the observed information matrix, in Figure 3.17b and Figure 3.17c, respectively, show that the BTw performs much better than equal weight for MSE, while it performs slightly better than equal weight for the logarithm of the determinant of the observed information matrix. However, Figure 3.17a, correlation criterion, indicates that BTw-Sqrt gives a slightly better result when the number of rankers is less than 2500.

The 95% confidence intervals for $\hat{\mu}$ are shown in Figure 3.18. This figure shows that the confidence interval for the $\hat{\mu}_{\text{PL}}$ are smaller than the $\hat{\mu}_{\text{BTw}}$ and the $\hat{\mu}_{\text{BTw-Sqrt}}$. Moreover, most of the $\hat{\mu}_{\text{PL}}$ lie between the $\hat{\mu}_{\text{BTw}}$ and the

$\hat{\mu}_{\text{BTw-Sqrt}}$.

The PL model and the BT model with BTw and BTw-sqrt weightings are applied to the Group I data from the Animal dataset and the Sushi dataset. Table 3.7 shows that the Kendall tau correlation and the MSE from BTw-Sqrt

Table 3.7: Kendall tau correlations and MSE for the BT model with BTw and BTw-Sqrt weightings when compared with the PL model

		Correlation	MSE
Group I	BTw	0.948	0.015
	BTw-Sqrt	0.956	0.004
Sushi	BTw	0.971	0.002
	BTw-Sqrt	0.961	0.005

weighting provide better results from the Animal dataset. The BTw weighting performs better in these criteria with the Sushi dataset. The total number of rankers of the Group I data is 450 which is less than 2500 rankers. This may be reason why the BTw-Sqrt weighting performs better than the BTw weighting.

3.6 Conclusions

In this chapter, we focus on two models for analyzing ranking data. The BT model is for pairwise comparisons data and the PL model is for complete or partial ranking data. We follow the MM algorithm from Hunter (2004) in order to obtain the ML estimates for the parameters of both the BT and the PL models. The PL models satisfy LCA (Marden, 1995). The essential implication of LCA is the constant ratio rule which is important for analyzing partial ranking data, since this rule implies that information about overall preferences can be found from the partial rankings.

We translated two Matlab codes of Caron and Doucet (2012) for the BT model and the PL model. These algorithms are called the `BTmm` and the `PLem`, respectively. We also implemented our own code, the `PLmm`, for the PL model.

The `PLem` and the `PLmm` algorithms require the same data format and they are based on matrices. However, they are different in the way of using matrices. In the `PLmm` algorithm, we produce more matrices than in the `PLem` algorithm.

We compare algorithms from different existing packages in R with each other and with our algorithms. The experiments for the BT model show that the `BTmm` algorithm is slower than the existing `BradleyTerry2` package. Moreover, for the PL model, the `PLem` algorithm is the fastest algorithm, followed by the `PLmm` algorithm. Both of the algorithms for the PL model have much faster computational times than the existing packages.

The `PLinfm` algorithm for computing the observed information is implemented. The standard errors which are calculated from the observed information matrix from the `PLinfm` algorithm and the `optim` function are compared in order to confirm our algorithm. The results show that they give the same standard errors.

The `PLem` and the `PLinfm` algorithms are applied to the Group I data from the Animal dataset as an illustration on how to interpret the results. The bootstrap goodness-of-fit tests with both the Kendall tau distance and IOS statistic show that the PL model is an appropriate model for fitting the Group I data.

Next, we explore the rank-breaking methods which were introduced by Soufiani et al. (2013a). The rank-breaking method will reduce complexity of analyzing of p items for a partial ranking to a set of pairwise comparisons. We apply three rank-breaking methods which are full, adjacent, and top-h to synthetic data. The full rank-breaking outperforms the other rank-breaking methods in both running time and quality of parameter estimates.

We further explore the full rank-breaking method with different weightings. Most of the weightings improve the estimates when compared with the non-weighting on the simulated data. Among the proposed weightings, the

BTw-Sqrt performs the best in the simulated data. Thus, we compare the BTw with the BTw-Sqrt on a real dataset, the Sushi dataset. The results show that the BTw-Sqrt is slightly better than BTw when n is less than 2500 rankers in term of the Kendall tau correlation criterion. However, the MSE criterion gives a different conclusion, that the BTw weighting performs better than the BTw-Sqrt. We prefer comparing the Kendall tau correlation because in many real world applications the parameter values are of less importance than the true rankings of the items. Thus, the BTw-Sqrt is a better option to use when number of rankers is small in this situation. Moreover, we further investigate the BTw and the BTw-Sqrt weightings by applying these weightings to the pairs from the Group I data from the Animal dataset. The Kendall tau correlation and MSE show that our BTw-Sqrt performs better than the BTw.

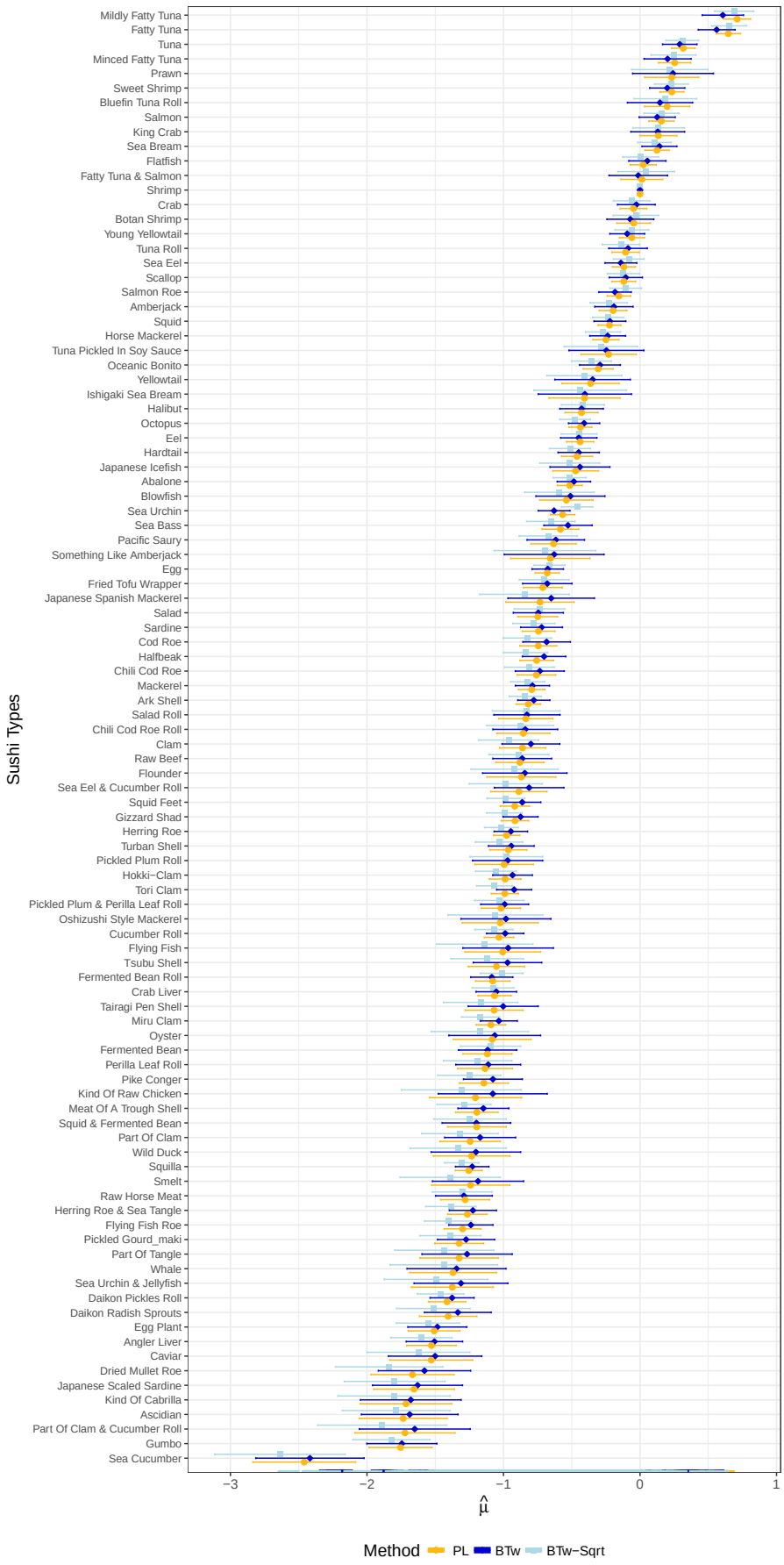


Figure 3.18: The 95% confidence interval of $\hat{\mu}$ of the PL model and the BT model with weighting (1) and (2) from Table 3.6 for the Sushi dataset

Chapter 4

Preference Learning

In many applications, ranking information is collected sequentially from each successive participant ranking a subset of the items. For example, in the Animal dataset, participants were asked to rank subsets of ten images. These subsets were chosen at random from all the images available. The general question addressed in this chapter is whether the rankings which have already been completed can be used to make a more effective choice of future subsets of items to be ranked than simply choosing at random. Moreover, we also aim to elicit rankings from as few participants as possible. Our main objective is to find methods that can efficiently pick subsets for ranking which quickly lead us to good estimates of the preference parameters.

This chapter presents methods for selecting informative subsets which contain more than two objects. We review some related work in Section 4.1. We describe why we are interested in this problem as presented in Section 4.2. We give small examples to illustrate our motivation. Section 4.3 discusses estimation of the logarithm of the expected information matrix by using a multiple regression model. In Section 4.4, we explore three existing criteria and three proposed methods for selecting a suitable subset. The existing criteria are adapted from experimental design framework, D-optimality and E-optimality. We also consider another criterion proposed by Soufiani et al. (2013b) which

we call the Wald criterion. Due to computational problems with the existing criteria, we propose three methods which are simpler to implement. Experimental results are shown in Section 4.5. First, the existing criteria were applied on simulated data. The data was generated under the PL model with small and large number of items, $K = 6$ and $K = 100$, respectively. In order to compare performance, random selection of subsets was also used. Second, the proposed methods were applied to simulated data with $K = 100$. These three proposed methods are again compared with random selection. Lastly, we compare the results from the existing criteria with $K = 100$ with the results from the proposed methods.

We change some of the notations here where $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ from Chapter 3 become $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$, respectively.

4.1 Related Work

The problem of item selection has been considered in the literature on paired comparisons. The scheduling method was introduced by Aftab et al. (2011). The result showed that the scheduling method performed better than random selection for pairwise comparison. Later, Pfeiffer et al. (2012) proposed an adaptive elicitation method for pairwise comparisons that was based on the Thurstone-Mosteller model. This study was done by using Amazon Mechanical Turk (MTurk). MTurk is a system in which internet users are paid a small fee for completing an online task, in this case doing a comparison. The adaptive method was compared with the random choice of pairs. Better results were obtained with the adaptive method. It increased the accuracy of estimating parameters more quickly. Both studies revealed that, for paired comparison data, if suitable pairs were selected then the estimated parameters converged more rapidly to the true values.

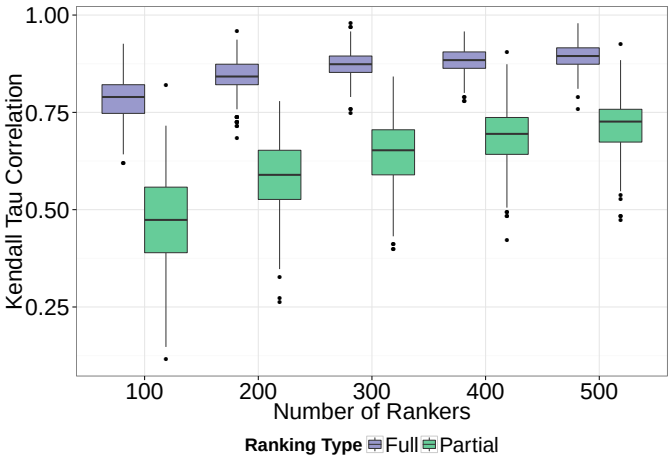
Soufiani et al. (2013b) extended these studies to multiple comparisons. They showed that criteria from experimental design can be adapted for using in this framework. Two such criteria are D-optimality and E-optimality. They also proposed a new method based on the t-test. Their simulation experiments used data generated from the Normal distribution and compared results with the random selection of subsets. The results showed that the D-optimality and E-optimality criteria sometimes perform worse than random selection. The proposed t-test criterion performed better than existing criteria.

4.2 Motivation

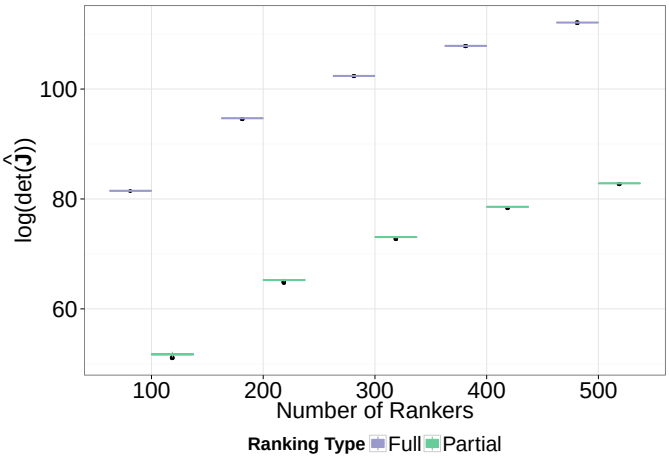
The main question of this study is how to find an effective way to elicit information about preferences. If the number of items is small, then rankers can be asked to undertake a full ranking. The problem occurs when there are too many items and it is impossible for rankers to rank them all. Thus, we need to choose the subset of items that provides the most useful information.

The difference in information gained between full and partial ranking is illustrated in Figure 4.1, which shows the Kendall tau correlation between estimated parameters and true parameters and the logarithm of the determinant of the observed information matrix ($\log(\det(\mathbf{J}(\hat{\boldsymbol{\lambda}})))$) after every 100 rankings starting from 100 to 500. The full and partial rankings were generated under a PL model for a total of 20 items in which each ranker ranked all of the items for the full ranking data, while only 6 items were required to be ranked for the partial ranking data. The subsets of 6 items were randomly selected for each ranker. The true parameter values for items 1 to 20 were $\pi_k \propto k$, $k = 1, \ldots, 20$.

We simulated 500 times at each point. As expected, the full and partial ranking data reveals the same trend that the Kendall tau correlation increases as the number of rankers increases, but the correlations from the full ranking



(a) Kendall tau correlation



(b) Logarithm of the Determinant of the Observed Information matrix

Figure 4.1: Boxplot of Kendall tau correlation between estimated parameters and true values and the logarithm of the determinant of the observed information matrix for the PL model when fitted to the full and partial simulated datasets

data are higher than from the partial ranking data when the data has the same number of rankers (Figure 4.1a). The boxplots show that the estimated Kendall tau correlation varies considerably between simulation runs, particularly for partial rankings, and variability gradually decreases as the number of rankers increases.

Figure 4.1b shows that $\log(\det(\mathbf{J}(\hat{\boldsymbol{\lambda}})))$ also increases as the number of rankers increases. However, the $\log(\det(\mathbf{J}(\hat{\boldsymbol{\lambda}})))$ does not vary much between simulation runs. Moreover, if we change the x-axis to be $\log(n)$, it does not

have a linear relationship with the $\log(\det(\mathbf{J}))$. This gives us the idea to estimate the $\log(\det(\mathbf{J}(\hat{\boldsymbol{\lambda}})))$ in Section 4.3.

To increase the Kendall tau correlation between the estimated parameters and the true ranks and the $\log(\det(\mathbf{J}(\hat{\boldsymbol{\lambda}})))$ for the partial ranking dataset, a suitable subset of items must be chosen to be ranked at each stage. A small example is given for illustrative purpose. Suppose we are interested in the preferences of 6 items but require only 4 items to be ranked at a time. The true parameters for item 1 to item 6 are generated from a uniform distribution and then sorted into ascending order. The true values are $\boldsymbol{\pi} = (0.014, 0.018, 0.107, 0.271, 0.289, 0.300)^\intercal$ and if we let item 1 be the reference item then $\boldsymbol{\lambda} = (0, 0.244, 2.007, 2.942, 3.005, 3.043)^\intercal$. Initial data was obtained by simulating the ranking of 20 randomly chosen subsets of size 4. The reason why we need these preliminary rankings is to have enough initial data to ensure that the PL model converges, to give a value of $\hat{\boldsymbol{\lambda}}$. We then calculate the expected information matrix ($\mathbf{I}_{\text{new subset}}$), evaluated at $\hat{\boldsymbol{\lambda}}$, for each possible subset. This involves calculating the observed information matrix for each possible ranking of the subset and calculating a weighted average of these matrices where the weights are the estimated probabilities of the different rankings. Moreover, the estimates $\hat{\boldsymbol{\lambda}}$ from the PL model are also required in this computation. The estimates and the initial data both affect the expected information matrix. To reduce the variability from the initial data, we repeat this process many times. The process is repeated 500 times ($n_{\text{sim}} = 500$). The algorithm is described in Algorithm 2.

Before presenting the results from Algorithm 2, we give a small example showing how to calculate the expected information matrix. We consider $K = 4$, $p = 3$, and $n_{\text{init}} = 20$. First, we generate an initial dataset under the PL model where true parameter values are generated from a uniform distribution. The PL model is fitted to the initial data and $\hat{\boldsymbol{\lambda}} = (0, 0.472, 2.142, 2.980)^\intercal$.

Algorithm 2 Find the expected information matrix for an extra subset

Given K, p, n_{init}, and n_{sim}
Find all possible subsets of p items from k items (n_C)
for $t = 1$ to n_{sim} **do**
    Generate an initial dataset under the PL model with an appropriate
    number of individuals (n_{init})
    Estimate the preference parameters ($\hat{\boldsymbol{\lambda}}$)
    for $i = 1$ to n_C **do**
        Find all possible orderings (n_P) of subset i
        for $j = 1$ to n_P **do**
            Update $\mathbf{I}_{\text{new}}(\hat{\boldsymbol{\lambda}})$ weighted by probability of ordering j
        end for
    end for
end for
Calculate mean of $\log(\det(\mathbf{I}_{\text{new}}(\hat{\boldsymbol{\lambda}})))$ of each subset

The subset {1, 2, 3} is considered; therefore, there are 6 possible orderings. We calculate the probability and the observed information matrix for each ordering as shown in Table 4.1. If K and/or p are large, This approach leads to computational problems.

Table 4.1: Probability for each ordering when $K = 4$ and $p = 3$

	Probability	$\mathbf{J}(\hat{\boldsymbol{\lambda}})$	Probability $\times$ $\mathbf{J}(\hat{\boldsymbol{\lambda}})$
$(1 \succ 2 \succ 3)$	0.014	$\mathbf{J}_1(\hat{\boldsymbol{\lambda}})$	$0.014 \times \mathbf{J}_1(\hat{\boldsymbol{\lambda}})$
$(1 \succ 3 \succ 2)$	0.076	$\mathbf{J}_2(\hat{\boldsymbol{\lambda}})$	$0.076 \times \mathbf{J}_2(\hat{\boldsymbol{\lambda}})$
$(2 \succ 1 \succ 3)$	0.015	$\mathbf{J}_3(\hat{\boldsymbol{\lambda}})$	$0.015 \times \mathbf{J}_3(\hat{\boldsymbol{\lambda}})$
$(2 \succ 3 \succ 1)$	0.129	$\mathbf{J}_4(\hat{\boldsymbol{\lambda}})$	$0.129 \times \mathbf{J}_4(\hat{\boldsymbol{\lambda}})$
$(3 \succ 1 \succ 2)$	0.294	$\mathbf{J}_5(\hat{\boldsymbol{\lambda}})$	$0.294 \times \mathbf{J}_5(\hat{\boldsymbol{\lambda}})$
$(3 \succ 2 \succ 1)$	0.472	$\mathbf{J}_6(\hat{\boldsymbol{\lambda}})$	$0.472 \times \mathbf{J}_6(\hat{\boldsymbol{\lambda}})$
Sum	1	-	$\mathbf{I}(\hat{\boldsymbol{\lambda}})$

In Figure 4.2, the 15 possible subsets are presented on x-axis while the y-axis shows the mean of the logarithm of the determinant of $\mathbf{I}_{\text{new subset}}(\hat{\boldsymbol{\lambda}})$. The determinant of the observed information matrix is the reciprocal of the generalized variance which is the determinant of the variance-covariance matrix. This value clearly shows how much information is provided by adding one extra subset since the determinant can be thought as a measure of volume. In

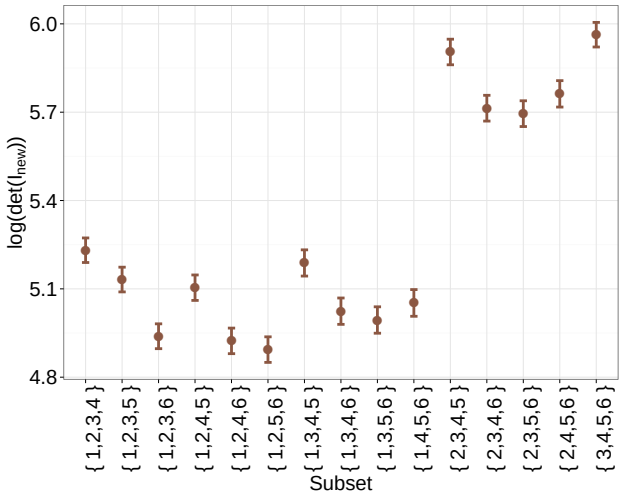


Figure 4.2: The average of the logarithm of determinant of the expected information matrix when a single extra subset is added to the initial data

this example, the last subset, which contains items $\{3, 4, 5, 6\}$, is chosen to be the next subset since this subset gives the largest average gain in expected information, in other words, gives the most information. Therefore, if we choose a good subset, we will get better estimated parameters in a shorter time.

4.3 Estimated Logarithm of Determinant of Expected Information Matrix

In this section, we try to estimate the expected information matrix given the values of n, K and p. We cannot compute the expected information matrix when K is large because of excessive computational times. The objective is to find the relationship among n, K, and p and extend this to find approximate relationships for the observed information matrix.

In general, information is additive if experiments are independent. Then we can write $\mathbf{I}_n(\boldsymbol{\lambda})$, the expected information matrix for sample size n, in terms of $\mathbf{I}_1(\boldsymbol{\lambda})$:

$$\mathbf{I}_n(\boldsymbol{\lambda}) = n \times \mathbf{I}_1(\boldsymbol{\lambda}).$$

Therefore,

$$\det(\mathbf{I}_n(\boldsymbol{\lambda})) = n^{(K-1)} \det(\mathbf{I}_1(\boldsymbol{\lambda})),$$

where the term $n^{(K-1)}$ arises because the information matrix has rank $K-1$ (Section 3.2.3). Therefore

$$\begin{aligned} \log\left(\det(\mathbf{I}_n(\boldsymbol{\lambda}))\right) &= (K-1)\log(n) + \log(\det(\mathbf{I}_1(\boldsymbol{\lambda}))) \\ &= (K-1)\log(n) + \text{constant}. \end{aligned} \tag{4.1}$$

Instead of $\mathbf{I}_1(\boldsymbol{\lambda})$, we can write this term as $\mathbf{I}_{n_{\min}}(\boldsymbol{\lambda})$ where $n_{\min}$ is a minimum sample size. The $n_{\min}$ is introduced here to confirm that we have enough rankings for the PL model to converge. Then the Equation (4.1) becomes

$$\log\left(\det\left(\mathbf{I}_n(\boldsymbol{\lambda})\right)\right) = (K-1)\log(n) + \text{constant} - K\log(n_{\min}) + \log(n_{\min}). \tag{4.2}$$

In order to get rid of $n_{\min}$ terms, we include K as one of covariates in the model. We combine the last terms in Equation (4.2) with the constant term. We get the same equation as in Equation (4.1). Therefore, we do not need to know $n_{\min}$ for estimating the $\log\left(\det\left(\mathbf{I}_n(\boldsymbol{\lambda})\right)\right)$.

The relationship in Equation (4.1) can also be applied to the observed information matrix. The observed information matrix is dependent on the data which involves n, K, and p. Thus, we can estimate $\log(\det(\mathbf{J}))$ given these parameters. This study investigates the effect of varying n, K, and p on the value of $\log(\det(\mathbf{J}))$, using multiple regression. The multiple regression model is a good model to start with because it is simple. The dependent variable is $\log(\det(\mathbf{J})) - (K-1)\log(n)$ since we expect $\det(\mathbf{J})$ to be proportional to n^{K-1} from Equation (4.1). The independent variables that we consider are p, K, $\log\binom{p}{2}$, $\log\left(\frac{p}{K}\right)$, $\log\left(\frac{n}{K}\right)$, and some interaction terms between pairs of these

independent variables. Backward elimination and stepwise variable selection methods are used to find the best model.

We start by fixing $K = 100$, and simulate data based on varying n and p under the PL model. The true parameter values are randomly generated from a uniform distribution. The settings of the simulated data, with fixed $K = 100$, and varying p and n are shown in Table 4.2. We simulate under these settings 20 times and apply the multiple regression model to these data.

Table 4.2: Values of p and n used to simulate data with fixed $K = 100$.

K	p	n
100	10	200, 300, 400, 500
100	20	200, 300, 400, 500
100	30	200, 300, 400, 500
100	40	200, 300, 400, 500
100	50	200, 300, 400, 500
100	60	200, 300, 400, 500
100	70	100, 200, 300, 400
100	80	100, 200, 300, 400
100	90	100, 200, 300, 400
100	100	100, 200, 300, 400

Table 4.3: Values of K and n used to simulate data with fixed $p = 10$.

K	p	n
10	10	50, 100, 150, 200
20	10	100, 200, 300, 400
30	10	100, 200, 300, 400
40	10	200, 300, 400, 500
50	10	200, 300, 400, 500
70	10	200, 300, 400, 500
80	10	300, 400, 500
90	10	200, 300, 400, 500
100	10	200, 300, 400, 500, 600, 700
150	10	300, 400

Table 4.4 shows the multiple regression result. The terms $\log\binom{p}{2}$ and $\log\left(\frac{p}{K}\right)$ are chosen by both selection methods, which are backward elimination and stepwise variable, and they are significant at the 0.1% significance level while $\log\left(\frac{n}{K}\right)$ is significant at 1% level. This shows that $\log\binom{p}{2}$, $\log\left(\frac{p}{K}\right)$, and

Table 4.4: Regression estimates when fixed $K = 100$

	Coefficient	SE	t-value	p-value
Intercept	-3207.742	43.232	-74.199	< 0.001
$\log\binom{p}{2}$	375.595	5.078	73.963	< 0.001
$\log\left(\frac{p}{K}\right)$	-649.112	10.353	-62.700	< 0.001
$\log\left(\frac{n}{K}\right)$	-0.295	0.106	-2.782	0.006

$\log\left(\frac{n}{K}\right)$ have an effect on $\log(\det(\mathbf{J}))$. The R^2_{adj} is 0.9997. The fitted regression

model is

$$\log(\widehat{\det(\mathbf{J}_{\text{Reg}})}) = (K-1)\log(n) - 3207.742 + 375.595 \log\binom{p}{2} - 649.112 \log\left(\frac{p}{K}\right) - 0.295 \log\left(\frac{n}{K}\right) \tag{4.3}$$

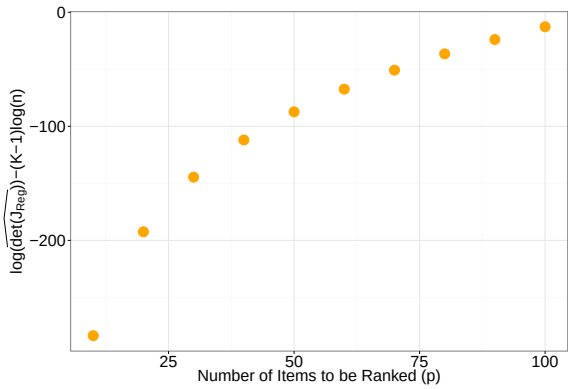


Figure 4.3: The estimated values $\log(\widehat{\det(\mathbf{J}_{reg})}) - (K-1)\log(n)$ from the regression model in Equation (4.3) with $p = 10, 20, \ldots, 100$, $K = 100$ and $n = 200$

In order to see the relationship between p and $\log(\widehat{\det(\mathbf{J}_{\text{Reg}})}) - (K-1)\log(n)$, we set $K = 100$ and $n = 200$ to plot those values as displayed in Figure 4.3. This figure indicates that when increasing p, the value of $\log(\widehat{\det(\mathbf{J}_{\text{Reg}})}) - (K-1)\log(n)$ also increases.

We next explore the effect of varying K but fixing p. The datasets are simulated by fixing $p = 10$ and varying K and n as presented in Table 4.3. Each setting is generated 20 times. Both model selection methods agree that K, $\log\left(\frac{p}{K}\right)$, and $\log\left(\frac{n}{K}\right)$ should be included in the model. However, $\log\left(\frac{n}{K}\right)$ is not significant at the 5% level. After we remove $\log\left(\frac{n}{K}\right)$ from the model, the $R^2_{\text{adj}} = 0.9987$ remains the same. The estimated regression coefficients are shown in Table 4.5.

We plot the estimated values $\log(\widehat{\det(\mathbf{J}_{reg})}) - (K-1)\log(n)$ given $p = 10$ against K in Figure 4.4. We get less information when K increases and p is fixed. This is because we have more parameters to estimate when we have more

Table 4.5: Regression estimates when fixed $p = 10$

	Coefficient	SE	t-value	p-value
Intercept	31.469	0.393	80.010	< 0.001
$\log\left(\frac{p}{K}\right)$	-46.420	0.604	-76.860	< 0.001
K	-4.242	0.013	-338.590	< 0.001

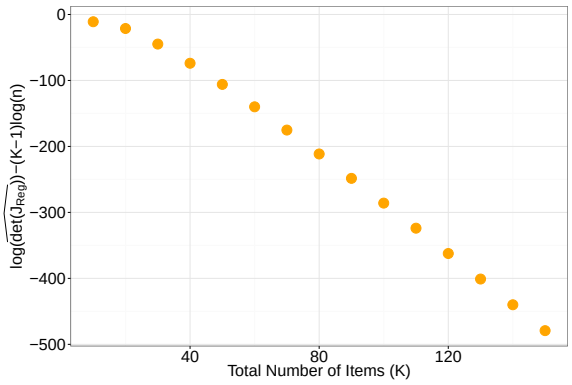


Figure 4.4: Plot of $\log(\widehat{\det(\mathbf{J}_{reg})}) - (K-1)\log(n)$ from the regression model in Table 4.5 against K with fixed $p = 10$.

items involved but each ranker still provides the same amount of information, $p = 10$ items out of K. Thus, the $\log(\widehat{\det(\mathbf{J}_{reg})}) - (K-1)\log(n)$ has a negative relationship to K.

Finally, we vary both K and p as shown in Table 4.6. Twenty sets of data were simulated according to each of these settings and multiple regression models were fitted to this dataset. The regression model that gave the highest R^2_{adj}, $R^2_{adj} = 0.9992$, included all terms in Table 4.7.

Table 4.6: The values of K, p, and n used to simulate data when varied both K and p.

K	p	n
10	4, 5, 7, 10	50, 100, 200
20	5, 10, 15,20	100, 200, 300
50	10, 15, 35, 50	200, 300, 400
70	10, 20, 30, 40, 50 , 60, 70	200, 300, 400
90	10, 20, 30, 40, 50 , 60, 70, 80, 90	200, 300, 400
100	10, 20, 30, 40, 50 , 60, 70, 80, 90,100	200, 300, 400

Both model selection methods give the same result which is that K, p,

$\log\binom{p}{2}$, $\log\left(\frac{p}{K}\right)$, $K \times p$, and $K \times \log\binom{p}{2}$ are all included in the model. All

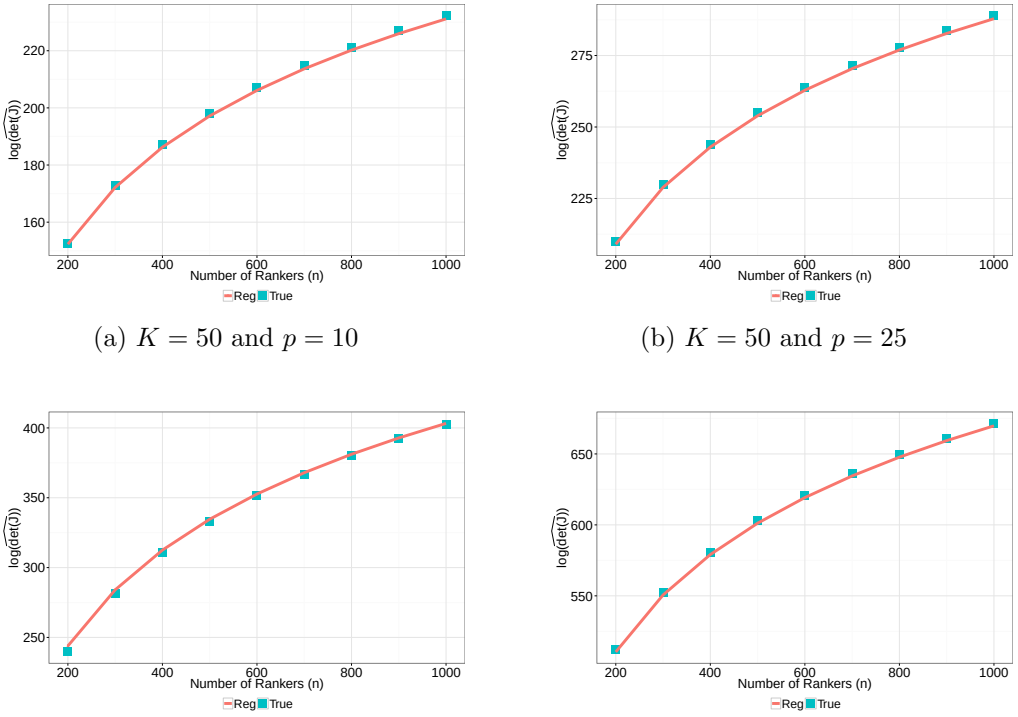
Table 4.7: Regression estimates where the dependent variable is $\log(\det(\mathbf{J}) - (K-1)\log(n))$ from simulations that varied both K and p

	Coefficient	SE	t-value	p-value
Intercept	-30.854	0.838	-36.838	< 0.001
K	-6.386	0.024	-264.954	< 0.001
p	0.241	0.028	8.741	< 0.001
$\log\binom{p}{2}$	15.821	0.279	56.729	< 0.001
$\log\left(\frac{p}{K}\right)$	-37.303	0.534	-69.909	< 0.001
$K \times p$	-0.005	0.001	-17.958	< 0.001
$K \times \log\binom{p}{2}$	0.645	0.004	168.338	< 0.001

terms in the table are significant at the 0.1% significance level. The estimated logarithm of the determinant of the observed information matrix from the regression model is

$$\log(\widehat{\det(\mathbf{J}_{\text{Reg}})}) = (K-1)\log(n) - 30.854 - 6.386K + 0.241p + 15.821\log\binom{p}{2} - 37.303\log\left(\frac{p}{K}\right) - 0.005(K \times p) + 0.645\left(K \times \log\binom{p}{2}\right). \tag{4.4}$$

We compare $\log(\widehat{\det(\mathbf{J}_{\text{Reg}})})$ from Equation (4.4) with the average of $\log(\det(\mathbf{J}))$ from the PL model when we simulated 100 times, which we refer to as true values. We plot four different settings in Figure 4.5 to compare $\log(\widehat{\det(\mathbf{J}_{\text{Reg}})})$ with $\log(\det(\mathbf{J}_{\text{True}}))$. The estimated logarithm of the determinant of the observed information from the regression model captures the trend of the true values well. Thus, the regression model provides a good approximation to the logarithm of the determinant of the information matrix. The regression model can be used as a guideline to predict the logarithm of the determinant of the information matrix as it is easy and quick to compute.



(a) $K = 50$ and $p = 10$

(b) $K = 50$ and $p = 25$

(c) $K = 100$ and $p = 10$

(d) $K = 100$ and $p = 100$

Figure 4.5: Plot of the estimated logarithm of the determinant of the observed information matrix from Equation (4.4) against true values.

4.4 Elicitation Criteria

In order to find the suitable next subset to be ranked, we consider three existing criteria. Two criteria are classical criteria used in experimental design and another criterion is the Wald criterion, which was introduced by Soufiani et al. (2013b). We also propose three new methods in this section.

4.4.1 Experimental Design Criteria

In the experimental design framework, many criteria have been introduced to quantify the performance of different designs, and these can be adapted to assess the performance of different elicitation schemes in our case. The main idea is to minimize the variance, which corresponds to maximizing the information. A "large" information matrix implies more efficient estimators. To measure how "large" an information matrix is, we require a scalar function

of the matrix. Different functions give raise to different optimality criteria. Well-known optimality criteria can be used. Two criteria are considered here, namely D-optimality and E-optimality.

D-optimality

The most popular criterion in the optimum experimental design framework is D-optimality which aims at maximizing the determinant of the expected information matrix which can be written as:

$$\max \det(\mathbf{I}).$$

Moreover, in order to ensure a convex optimization problem, it is given by

$$\max(\log(\det(\mathbf{I}))) \text{ or } \min(-\log(\det(\mathbf{I}))),$$

(Atkinson et al., 2007, Chapter 10).

Additionally, as the variance-covariance matrix is the inverse of the information matrix, the D-optimality criterion is equivalent to minimizing the determinant of the variance-covariance matrix. The adaptive algorithm proposed by Pfeiffer et al. (2012) is closely related to the D-optimality criterion and their method works well for pairwise comparisons. Later, Soufiani et al. (2013b) suggested that D-optimality might not be a suitable choice for preference rankings; however, they studied a different setting than this. They focused on General Random Utility models and generated datasets from a normal distribution.

In our case, we aim to find the next subset. To achieve this, we need to know the probabilities of all possible orderings. Then the expected information matrix for each possible subset is calculated. The subset that gives the information matrix with the greatest logarithm of determinant is chosen.

Details are given in Section 4.5.1.

E-optimality

The E-optimality criterion is quite widely used in optimum experimental design as well. This criterion chooses the subset that maximizes the minimum eigenvalue of the expected information matrix and it can be expressed as:

$$\max \min \left(eigenvalue \left(\mathbf{I} \right) \right).$$

In the experimental design area, the E-optimality criterion is used for designs in which all factors are qualitative, while D-optimality is mostly used for quantitative factors. The E-optimality criterion is expected to be more suitable than the D-optimality criterion in our case.

As for the D-optimality criterion, we need to find the expected information for each subset. The eigenvalues are computed from this matrix and we choose the subset that maximizes the minimum eigenvalue. See Section 4.5.1.

4.4.2 Wald Criterion

Soufiani et al. (2013b) proposed a new elicitation criterion which we refer to the Wald criterion. This criterion is based on the idea of the t-test to compare $\boldsymbol{\lambda}-$values in a pairwise comparison. The larger the value of the Wald criterion, the easier it is to distinguish item$_i$ from item$_j$. The Wald criterion is as follows:

$$\text{Wald}_{ij} = \frac{\mid \hat{\lambda}_i - \hat{\lambda}_j \mid}{\sqrt{\text{var}\left(\hat{\lambda}_i\right) + \text{var}\left(\hat{\lambda}_j\right) - 2\text{cov}\left(\hat{\lambda}_i, \hat{\lambda}_j\right)}}.$$

The subsets that contain the pair which has minimum value overall the Wald$_{ij}$ are selected. We choose a subset among the selected subsets which has maximum value of sum of all possible pairs based on the item i in the selected subset. For example, $K = 4$ and $p = 3$, if (1, 2) has minimum value, there are

two selected subsets which are (1, 2, 3) and (1, 2, 4). We calculate Wald_{12} + Wald_{13} + Wald_{23} and Wald_{12} + Wald_{14} + Wald_{24} and choose the subset that has greater value.

4.4.3 Proposed Selection Methods

Random selection is the most popular method for selecting subsets of items to be ranked; however, random selection ignores the information available from the previously ranking sets. Three proposed methods are introduced here and these methods will make use of the previous rankings. We suggest a systematic way to perform an effective way of choosing the next subset. The proposed methods are started with random selection and then use the ranked sets to find new subsets.

Suppose for illustration that there are 16 items and that each ranker ranks a subset of 4 items. First, the 16 items are randomly divided into 4 subsets so that each item belongs to only one subset. Each ranker gives their preference ranking of one subset. These are starting ranking sets as presented in Figure 4.6.

	Rank Position			
	1	2	3	4
Ranker1	1	2	3	4
Ranker2	5	6	7	8
Ranker3	9	10	11	12
Ranker4	13	14	15	16

Figure 4.6: The starting ranking sets

Method I

Method I selects the first item in each ranking set, shown in the blue box in Figure 4.7, to be a new subset. Then selects items ranked second (the orange box) as another new subset and so on.

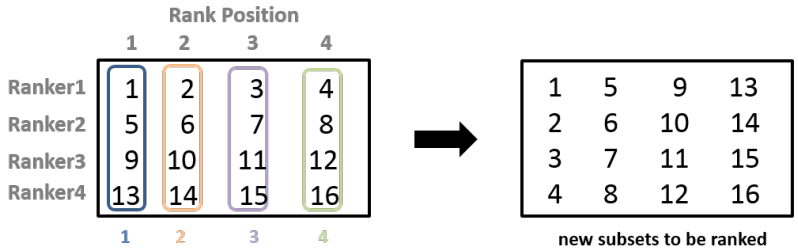


Figure 4.7: Method I

In order to perform Method I, it needs at least p starting ranked sets. The pseudo code of Method I is shown in Algorithm 3 which explains how Method I works. We repeat this algorithm by using the output, the *rankedSubsets*, as the starting data for the next iteration.

Algorithm 3 Method I

1: Given a starting data which contain all K items with p ranking sets
2: **for** $j = 1$ to p **do**
3: $newSubsets \leftarrow$ a subset of all items with rank j^{th} in the starting data
4: **end for**
5: **return** *newSubsets*

Method II

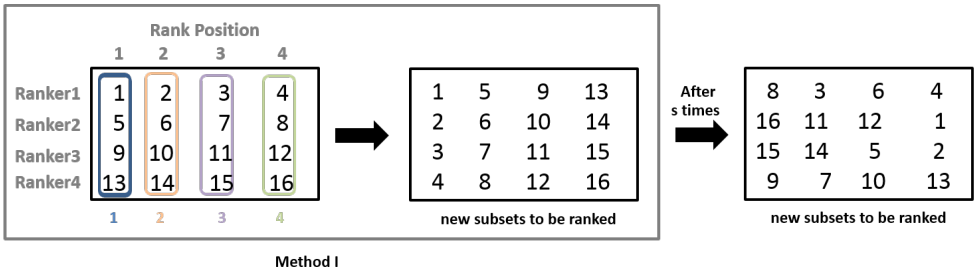


Figure 4.8: Method II

Method II is a mixed method between Method I and random selection. Method II will perform like Method I for s times and after that random selection is applied. Random selection is considered after s times because the data from the proposed method may make the PL model converge for a smaller number of rankers and after that random selection, which performed well in

Soufiani et al. (2013b), is used. Again Method II needs at least p starting ranking sets since the first part of this method is Method I.

Method III

Methods III consists of two steps. The first step is dividing the starting ranked sets into 2 groups as shown in red and yellow boxes in Figure 4.9.

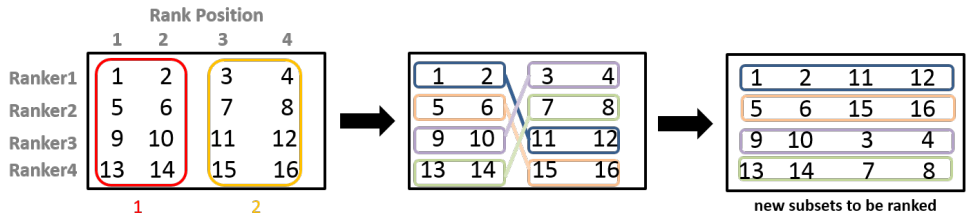


Figure 4.9: Method III

The first group contains the items in 1^{st} to $\lceil \frac{p}{2} \rceil^{th}$ rank position and the remaining items belong to the second group. The second step is to match the first group from ranker i with the second group from ranker $i+2$. We get p new subsets. Like the previous methods, this method requires at least p starting ranking sets. The algorithm is presented in Algorithm 4.

Algorithm 4 Method III

1: Given a starting data with a total number of K items and p ranking sets
2: $group1 \leftarrow 1^{st}$ to $\lceil \frac{p}{2} \rceil^{th}$ rank in the starting data
3: $group2 \leftarrow \lfloor \frac{p}{2} \rfloor^{th}$ to p^{th} rank in the starting data
4: **for** $j = 1$ to p **do**
5: $\quad newSubsets \leftarrow$ a subset from ranker j in $group1$ and a subset from ranker $j+2$ in $group2$
6: **end for**
7: **return** $newSubsets$

4.5 Simulation Study

4.5.1 Evaluation of the D-optimality, E-optimality, and Wald Criteria

In application for finding the next subset, the generating process is presented in Algorithm 5. Algorithm 5 processes through all of the possible subsets (n_C) and all of the possible orderings (n_P). $\mathbf{J}$ denotes the observed information matrix from the data and $\mathbf{I}_{\text{new}}$ is the expected information matrix when adding each possible subset. We target the next subset of items that can provide the best expected improvement upon each criterion to the current estimation.

Algorithm 5 D-optimality and E-optimality criteria

1: Given an initial dataset with an appropriate number of individuals (n_{init})
2: Estimate the preference parameters ($\hat{\boldsymbol{\lambda}}^{(0)}$) and compute $\mathbf{J}^{(0)}$ based on the initial dataset
3: **for** $t = 1$ to n_{extra} **do**
4: Find all possible subsets of p items from k items (n_C)
5: **for** $i = 1$ to n_C **do**
6: Find all possible orderings (n_P) of subset i
7: **for** $j = 1$ to n_P **do**
8: Update $\mathbf{I}_{\text{new}}$ to include ordering j weighted by probability of ordering j based on the current estimates, $\hat{\boldsymbol{\lambda}}^{(t-1)}$
9: **end for**
10: **end for**
11: Choose the subset that fulfils the criterion based on $\mathbf{J}^{(t-1)} + \mathbf{I}_{\text{new}}$
12: Generate a ranking of chosen subset based on the true values or ask a ranker to rank chosen subset
13: Add to the current dataset and recalculate $\hat{\boldsymbol{\lambda}}^{(t)}$ and $\mathbf{J}^{(t)}$
14: **end for**

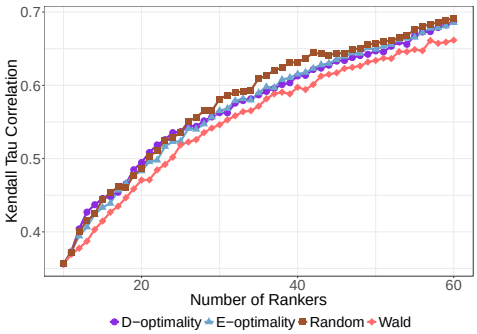
Simulation Study: Small Number of Items

An initial study is based on a small example with $K = 6$ items and subsets of size $p = 4$ since it is easier to understand and requires less computational time.

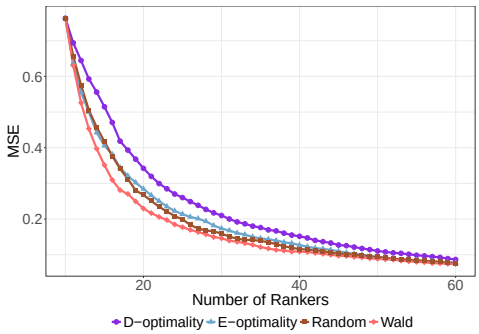
We generate $n_{\text{init}} = 20$ rankers' preference rankings of randomly chosen

subsets of 4 items, based on the PL model and use these rankings as initial data. Since the 4 items in each subset are chosen from a total of 6 items, there are 15 possible subsets ($n_C = 15$) and each subset has 4! possible orderings ($n_P = 24$).

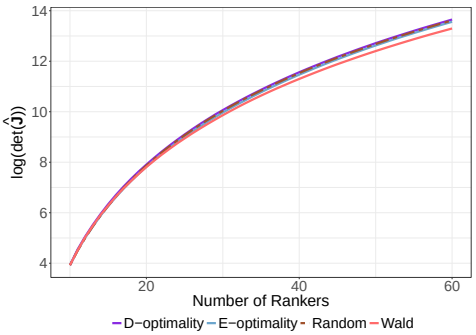
The true parameters are randomly generated from a uniform distribution between 0 and 1. Initial data is obtained by simulating 10 rankings of size 4 under the PL model.



(a) Kendall Tau Correlation



(b) MSE



(c) Logarithm of the Determinant of the Observed Information matrix

Figure 4.10: The average of Kendall tau correlation, MSE, and logarithm of the determinant of the observed information matrix of parameter estimates from different criteria, D-optimality, E-optimality, Wald criteria when fitted the PL model on synthetic data with $K = 6$ and $p = 4$

The three criteria which consist of D-optimality, E-optimality, and Wald criteria are considered. Random subsets are also generated in order to compare performance with the other criteria. We find the next 50 subsets ($n_{\text{extra}} = 50$) and evaluate each new ranking subset by comparing Kendall tau correlation, MSE, and logarithm of determinant of the observed information matrix. The

y-axis in Figure 4.10 is the average of these criteria after repeating the Algorithm 5 500 times.

The last criterion, $\log(\det(\mathbf{J}))$, is shown in Figure 4.10c. This criterion shows that the D-optimality and E-optimality criteria give the same performance as random selection. These two perform the best in $\log(\det(\mathbf{J}))$ criterion. The Wald criterion performs slightly worse than them.

Overall the E-optimality and Wald criteria are comparable with the random selection.

Simulation Study: Larger Number of Items

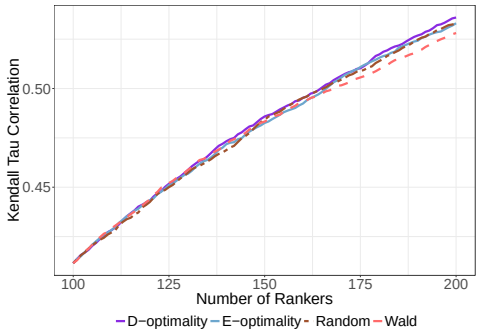
In order to compare the performances of the different criteria used to find the next subset when K is large, we explore one specific case when $K = 100$ and $p = 10$. Therefore, there are 1.731×10^{13} possible subsets and each subset has 3,628,800 possible orderings. These number of subsets and orderings are too large to compute for all possible situations. Instead, we have to select a random sample of subsets and orderings as shown in Algorithm 6.

Algorithm 6 D-optimality and E-optimality criteria

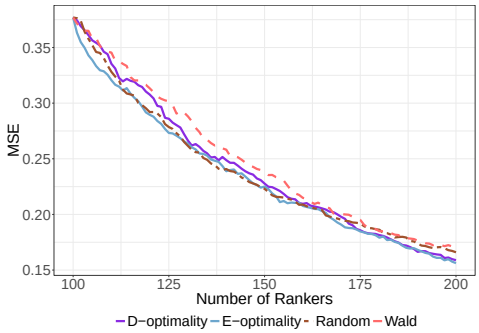
1: Given an initial dataset with an appropriate number of individual (n_{init})
2: Estimate the preference parameters ($\hat{\boldsymbol{\lambda}}^{(0)}$) and compute $\mathbf{J}^{(0)}$ based on the initial dataset
3: **for** $t = 1$ to n_{extra} **do**
4: **for** $i = 1$ to n_C **do**
5: Randomly select p items
6: **for** $j = 1$ to n_P **do**
7: Generate a ordering under the PL model and update $\mathbf{I}_{\text{new}}$ to include the ordering weighted by probability of it which based on the current estimates, $\hat{\boldsymbol{\lambda}}^{(t-1)}$
8: **end for**
9: **end for**
10: Choose the subset that fulfils the criterion based on $\mathbf{I}_{\text{new}}$
11: Generate a ranking set of chosen subset based on the true values then add it to the current dataset and calculate $\hat{\boldsymbol{\lambda}}^{(t)}$ and $\mathbf{J}^{(t)}$
12: **end for**

We set n_C and n_P equal to 200 subsets and 100 orderings, respectively.

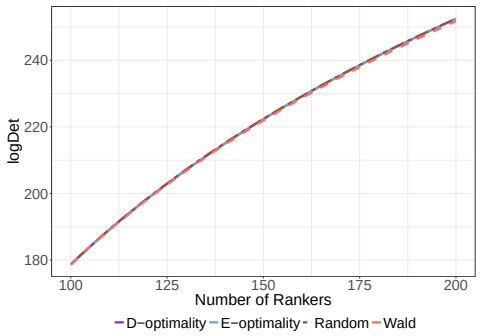
The true values are generated from a uniform distribution between 0 and 1. A sample of 100 rankings ($n_{\text{init}} = 100$) are generated under the PL model and used as an initial dataset. We repeat the Algorithm 6 200 times with different initial datasets. To evaluate the three criteria, the same evaluating criteria are considered as in the previous section. The results are illustrated in Figure 4.11.



(a) Kendall Tau Correlation



(b) MSE



(c) Logarithm of the Determinant of the Observed Information matrix

Figure 4.11: The average of Kendall tau correlation, MSE, and logarithm of the determinant of the observed information matrix, for four criteria: D-optimality, E-optimality, Wald, and random selection, on synthetic data for $K = 100$ and $p = 10$

Figure 4.11a shows the average of Kendall tau correlations between the estimates and true values. All methods perform slightly better than random selection from 100 to 150 rankers. After that the Wald gives worse performance when compared with random selection. Figure 4.11b, the MSE gives a clearer conclusion. The E-optimality performs the best at the beginning. Afterwards, it is comparable with random, while the D-optimality and the Wald

criteria perform worse than random selection. Figure 4.11c shows that there is no obvious conclusion, since all criteria give almost the same results as random selection. The E-optimality seems to work slightly better than random selection among these three criteria.

4.5.2 Evaluation of the Proposed Methods

In order to evaluate the three proposed methods, we compare them with random selection. Figure 4.12a, Figure 4.12b, and Figure 4.12c are to show how the Method I, Method II, and Method III perform, respectively. These flowcharts show the calculated estimated parameters by using the Total Data from the proposed methods to measure the performances. The measuring criteria are Kendall tau correlation, MSE, and the number of rankers where the PL model starts to converge. These are computed in Evaluate(Total Data) in the flowcharts. For convergence criterion, we consider two things which are the maximum number of iterations and the Assumption 2 in Chapter 3. If the PL model runs out of iterations, we consider that the PL model does not converge properly. Here, the maximum number of iterations is 1000. These criteria are used to measure the performance of the proposed methods.

All the proposed methods require two inputs which are starting data and n_{extra}. Method I needs one extra input, r, and Method II needs two extra inputs, r and s. Input n_{extra} is the number of repeated iterations in which the number of rankers at i^{th} iteration is $p + (i \times p)$ for each method. Input r is the number of times that the algorithm randomly reassigns all items into subsets without replacement when the process runs up to a multiple of r. Moreover, r can be any number within the range of 2 to s.

The experiments are conducted on synthetic data by using the algorithms in Figure 4.12. We set $n_{\text{extra}} = 100$, $r = 10$, $s = 40$, $K = 100$, $p = 10$, and true parameter values are the same as previous as for the larger number of items

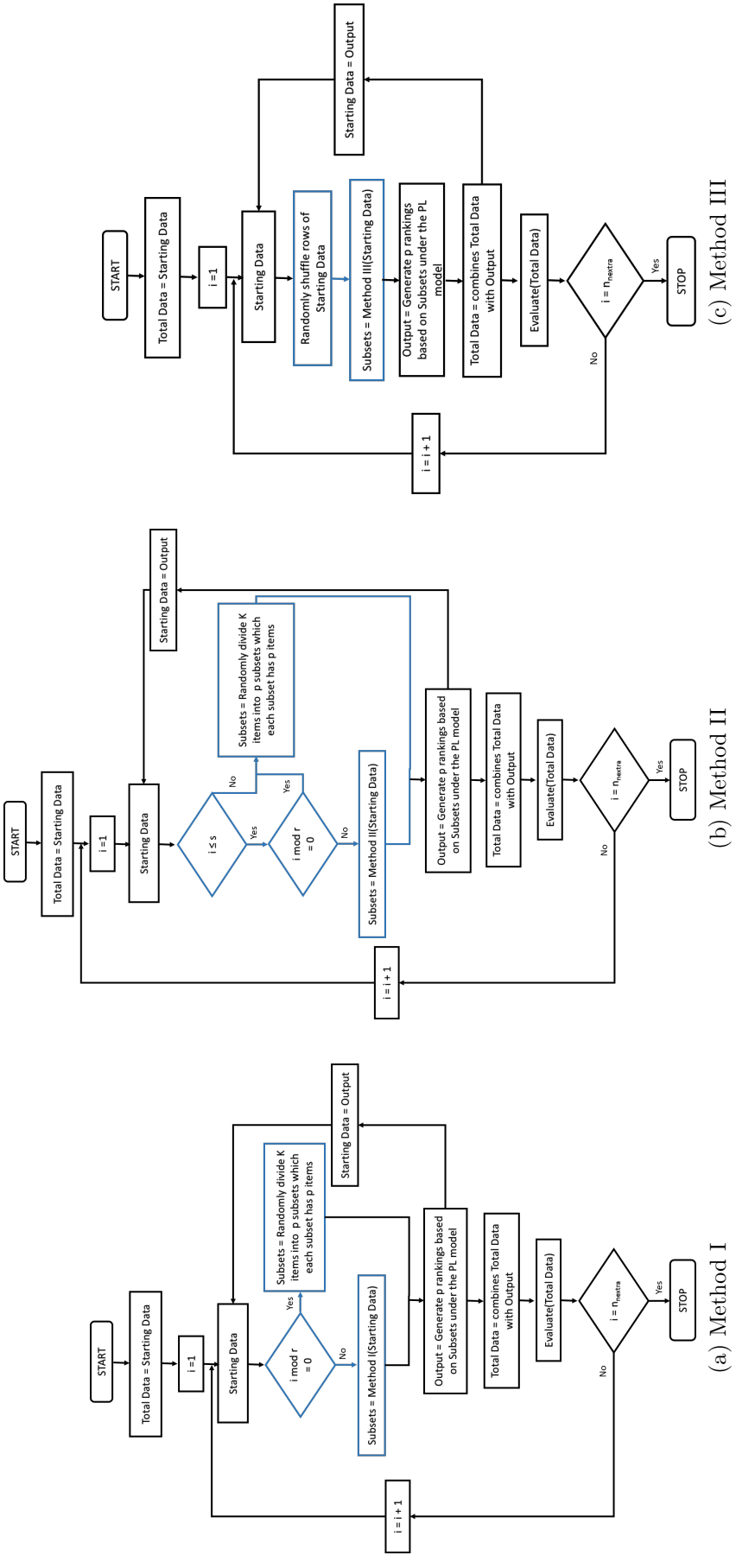
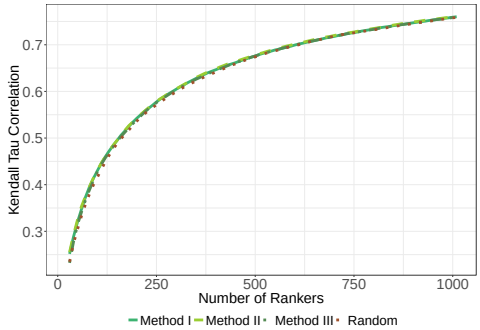
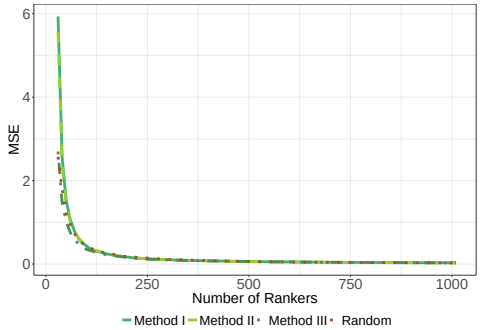


Figure 4.12: Flowchart for evaluating the proposed methods

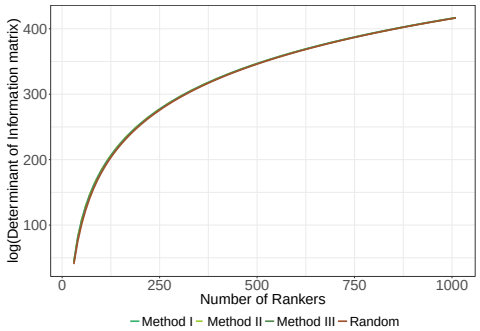
when $K = 100$ and $p = 10$. The algorithms are repeated 500 times and then the average of results from the Evaluate(Total Data) process in Figure 4.12 are plotted in Figure 4.13.



(a) Kendall Tau Correlation



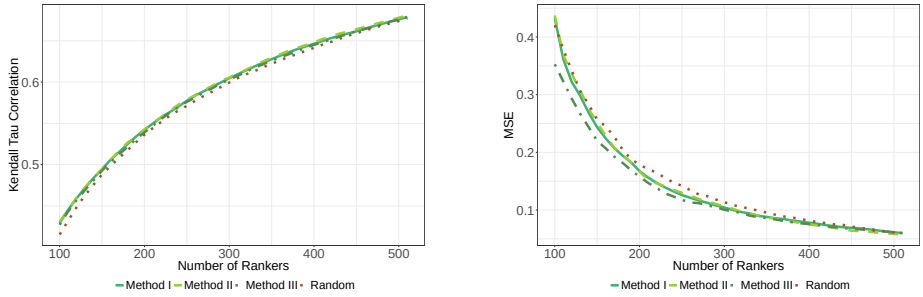
(b) MSE



(c) Logarithm of the Determinant of the Observed Information matrix

Figure 4.13: The average of Kendall tau correlation, MSE, and logarithm of the determinant of the observed information matrix criteria for the proposed methods and random selection when fitted the PL model on synthetic data with $K = 100$ and $p = 10$

The results are not clear for the Kendall tau correlation and the logarithm of the observed information matrix criteria. All the methods have almost the same performance as shown in Figure 4.13a and Figure 4.13c. We look closer at Figure 4.14a. It shows that all proposed methods perform slightly better than random selection. The MSE criterion in Figure 4.13b gives a clearer idea. At the beginning from 30 to 90 rankers, the Method III performs the best followed by random, and the Method I and Method II which give similar results. We plot another figure from 100 to 500 rankers to have a closer look as shown in Figure 4.14b. The Method III still gives a better performance



(a) Kendall Tau Correlation (b) MSE

Figure 4.14: The average of Kendall tau correlation and MSE criteria for the proposed methods and random selection when fitted the PL model on synthetic data with $K = 100$ and $p = 10$ from 100 to 500 rankers

than other methods followed by Method I, Method II, and random when the number of rankers is between 100 and 500. After 500 ranks, all methods are comparable.

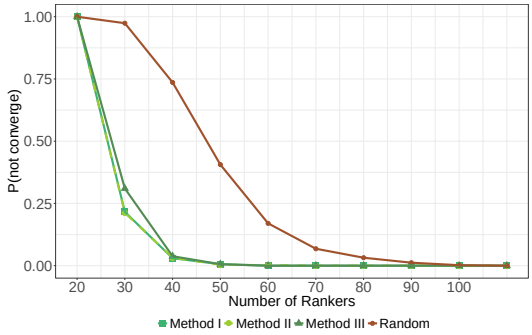


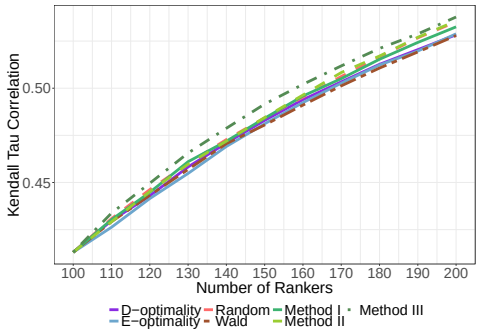
Figure 4.15: Convergence rate for the proposed methods and random selection

The last criterion is convergence rate as shown in Figure 4.15. At 20 rankers, none of the methods converge. All methods start to converge after 30 rankers. The proposed methods converge faster than random method. At 30 rankers, around 70% of datasets generated from the proposed methods converge, while datasets from random selection rarely converge.

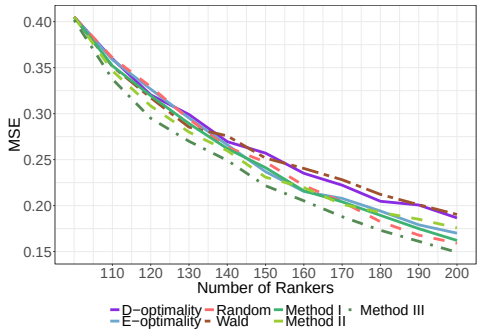
Focusing on convergence rate, the proposed methods outperform random selection. Method III is slightly better than the others in term of MSE criterion.

4.5.3 Comparisons for D-, E-optimality, Wald Criteria and Proposed Methods

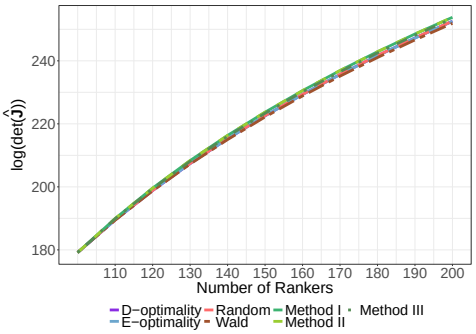
We compare all the methods which are D-optimality, E-optimality, Wald, Method I, Method II, and Method III. We use the same setting as Section 4.5.1: Larger Number of Items for D-optimality, E-optimality, and Wald criteria. We generate 100 rankings under the PL model to be starting data. The first 90 rankings are randomly generated and the last 10 rankings are generated based on the data structure that is required for the proposed methods. We randomly divide K items into p subsets with p items in each subset and then each subset is assigned an ordering under the PL model.



(a) Kendall Tau Correlation



(b) MSE



(c) Logarithm of the Determinant of the Observed Information matrix

Figure 4.16: The average of Kendall tau correlation, MSE, and logarithm of the determinant of the observed information matrix criteria for the D-optimality, E-optimality, Wald criteria, the proposed methods, and random selection when fitted the PL model on synthetic data with $K = 100$ and $p = 10$

The three evaluating criteria as before are performed in order to compare

results. We repeat this process 200 times and the average of each criterion is shown in Figure 4.16. Figure 4.16a shows average of the Kendall tau correlation and reveals that the Method III outperforms. The other methods give almost the same performance. Figure 4.16b confirms that the Method III has best performance under the MSE criterion when compared with other methods. Moreover, the proposed methods perform better than the D-optimality, E-optimality, and Wald criteria when the number of rankers is less than 150. After that E-optimality is comparable with Method I and Method II while the Wald performs better than Method I and Method II at the end of the figure. All methods perform better than random selection when number of rankers is greater than 155. The average of the logarithm of the determinant of the observed information matrix is shown in Figure 4.16c. The proposed methods give slightly better result than the D-optimality, E-optimality, Wald, and random.

4.6 Conclusions

In this chapter, we explored whether existing ranking sets contain useful information that can be used to improve the selection of subsets to be ranked in comparison to random selection. We explained this idea by giving small examples. One of these examples, gave us the idea that it might be possible to develop useful predictive models of the logarithm of the determinant of the information matrix. We used a simple model, multiple regression model, to estimate the logarithm of the determinant of the observed information matrix. The multiple regression with K, p, $\log\binom{p}{2}$, $\log\frac{p}{K}$, $K \times p$, and $K \times \log\binom{p}{2}$ included in the model provided good estimates of the logarithm of the determinant of the observed information matrix.

We compared the existing criteria, D-optimality, E-optimality, and Wald, for eliciting preference data. The first two criteria are from the framework

of experimental design. Another criterion, Wald, is based on the idea of the t-test. We investigated the performance of these criteria on simulated data. The E-optimality and Wald criteria improve the precision of estimation when compared with random selection in MSE when $K = 6$ and $p = 4$. The D-optimality criterion performs the worst in MSE. Wald does not perform well in correlation. Therefore, the E-optimality criterion provides good performances in overall criteria, Kendall tau correlation, MSE, and $\log(\det(\mathbf{J}))$.

We increase the number of items and the same criteria are applied to simulated data with $K = 100$ and $p = 10$. There are too many possible subsets to explore when $K = 100$. We randomly choose 200 possible subsets for further investigation. With $p = 10$, again we cannot test all the possible permutations/rankings due to computational time. The PL model is used to generate 100 rankings. We explore large K with these conditions. The results show that the D-optimality and E-optimality criteria are comparable with random selection while the Wald criterion performs slightly worse than random selection under the MSE criterion. The Kendall tau correlation does not give any obvious conclusion. As before, the E-optimality performs the best among the existing criteria and in overall evaluating criteria when $K = 100$ and $p = 10$.

The D-optimality, E-optimality and Wald criteria have little improvement over random selection when number of rankers is less than 150 in the Kendall tau correlation criterion. Our results are different from Soufiani et al. (2013b). Soufiani et al. (2013b) concluded that the Wald criterion can significantly improve the precision of estimation in comparison with random selection. However, we use larger number of rankers to ensure that the PL model converges and fits the data properly. Moreover, the data in Soufiani et al.'s (2013b) paper are not available. We cannot reproduce their works.

We propose systematic methods that can select the next p subsets. The simulation experiments show that the proposed methods slightly improve the

performance in both the Kendall tau correlation and MSE compared to random selection. Moreover, the convergence rate also reveals the same conclusion that the PL model can be fitted to the data from the proposed methods with less number of rankers when compared with random selection.

Finally, we compare performances of the three statistical methods and the three proposed methods. We use the same setting as before where $K = 100$ and $p = 10$. The Kendall tau correlation and MSE show that the Method III performs best. Most of the proposed methods perform better than the D-optimality, E-optimality, Wald, and random. All the methods give almost the same results of the logarithm of the determinant of the observed information matrix.

The idea of selecting informative subset is a good idea since it can quickly lead to good estimates of the parameters. However, the D-optimality, E-optimality and Wald criteria do not have outstanding results in our practical experiments. The computational costs of implementing the D-optimality, E-optimality, and Wald criteria imply that it is not worthwhile to use these criteria in practice. While the proposed methods performs better in the convergence rate criterion and they do not need much computational time.

Chapter 5

Extensions of the Plackett-Luce Model

In this chapter we describe two extensions of the Plackett-Luce (PL) model, the Rank-Ordered Logit (ROL) model and the Benter model. These models provide different types of extension. The ROL model allows the inclusion of covariates, while the Benter model allows preferences for higher-ranked items to be stronger than lower-ranked items. We explain the ROL model in Section 5.1. The Minorization-Maimization (MM) algorithm from Hunter (2004) is extended for using with the ROL model. The `ROLmm` and the `ROLinfm` algorithms are implemented in R to find estimated parameters and compute the observed information matrix of the ROL model, respectively. We compare results from our algorithms with the `optim` function. Section 5.2 provides details of the Benter model. We follow the work from Gormley and Murphy (2008) to fit the model. We implement two algorithms, `BMmm` and `BMinfm` algorithms, in R in order to fit the Benter model and to calculate an observed information matrix. We apply the `BMmm` algorithm and the `optim` function to the Group I data from the Animal dataset. The results and computational times are compared. The `BMinfm` algorithm is also compared with the Hessian matrix from the `optim` function in order to confirm that our algorithm works properly. We combine

these two extensions to give a model that incorporates covariates and allows for a dampening effect where details are given in Section 5.3. Section 5.4 briefly describes the Likelihood Ratio (LR) test. All three models are applied to the Animal dataset, as discussed in Section 5.5, in order to compare among the models and to find significant covariate(s) that affect the preferences. The LR test is used to compare the models when we add covariates to the models. The bootstrap goodness-of-fit test is applied to assess whether the Benter and the combined models provide good fits to the Group I data.

5.1 Rank-Ordered Logit Model

The PL model can be extended to incorporate covariates into the model (Alvo and Yu, 2014). This is a generalization of the conditional logit regression model introduced by McFadden (1974). This model was proposed by Beggs et al. (1981) and later Hausman and Ruud (1987) developed it further under the name *rank-ordered logit model* in the field of econometrics. Moreover, this model was also independently proposed in the marketing literature under the name *exploded logit model* (Punj and Staelin, 1978; Chapman and Staelin, 1982). We call this model the ROL model in this chapter. The ROL model, which included covariates, gives more information on each covariate about how the specific information from rankers and items affects rankings.

The most general form of the model contains three kinds of covariates describing item characteristics, ranker characteristics, and ranker-item characteristics (Alvo and Yu, 2014). Let $\mu_{\rho_{ij}}$ be the corresponding ROL parameters then a general parametric form is

$$\mu_{\rho_{ij}} = \exp\left(\lambda_{\rho_{ij}} + \sum_{l=1}^{L} \beta_l z_{l,\rho_{ij}} + \sum_{r=1}^{R} \gamma_{r,\rho_{ij}} x_{r,i} + \sum_{q=1}^{Q} \theta_q w_{q,\rho_{ij}}\right), \tag{5.1}$$

where $z_{l,\rho_{ij}}$ is a covariate that depends on the item ρ_{ij} e.g. cost, colour and β_l

is a parameter specific to items. The covariate $x_{r,i}$ describes a characteristic of the rankers e.g. the age or gender of the ranker, but does not vary over items, and the coefficient $\gamma_{r,\rho_{ij}}$ is a ranker-specific parameter. Finally, $w_{q,\rho_{ij}}$ is a covariate that describes a relation between item ρ_{ij} and ranker i and θ_q is a ranker-item specific parameter e.g. ownership of items, previous knowledge about the items. The number of ranker-specific parameters, R, must be less than or equal to $K-1$ in order to avoid linear dependence. The simplest case is to set $\gamma_{r,1}=0$. This model is not reversible; the coefficients of model for ranking from worst to best are not the negatives of the coefficients from a model of ranking from best to worst.

The ROL model specifies that the probability of the ranking ρ_i is the same as the PL model in Equation (3.6) where $\mu_{\rho_{ij}}$ in Equation (5.1) is substituted for $\lambda_{\rho_{ij}}$. A special case is when β_l and θ_q are zero, and $\gamma_{r,\rho_{ij}}$ is non-zero. This model is called the multinomial logit model (Allison and Christakis, 1994). The model with β_l and $\gamma_{r,\cdot}$ is McFadden's conditional logit model (McFadden, 1976).

The Thurstonian model is similar to the ROL model. Equation (5.1) is the same and

$$U_{\rho_{ij}} = \mu_{\rho_{ij}} + \epsilon_{\rho_{ij}},$$

where $\epsilon_{\rho_{ij}}$ assumes a normal distribution rather than an extreme value distribution. However, it is computationally demanding to fit this model (Allison and Christakis, 1994).

The ROL model has been applied in the economics and marketing area (Lareau and Rae, 1989; Moore, 1990; Katahira, 1990; Kamakura and Mazzon, 1991; Koop and Poirier, 1994; Ahn et al., 2006; Kumar and Kant, 2007). Moreover, the ROL model has also been employed in the sociology field, beginning with Allison and Christakis (1994). Commonly used covariates are

- ranker-specific covariates: gender, age, marital status, income

- item-specific covariates: price of the item, time of day of a concert
- ranker-item-specific covariates: used the item before, ownership

5.1.1 Maximum Likelihood Estimator

The log-likelihood function of the ROL model can be written like the PL model by using $\mu_{\rho_{ij}}$ from Equation (5.1) instead of $\lambda_{\rho_{ij}}$, giving

$$\ell(\boldsymbol{\zeta}) = \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \left[\log \mu_{\rho_{ij}} - \log \left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right) \right], \tag{5.2}$$

where $\boldsymbol{\zeta} = (\lambda_1, \ldots, \lambda_K, \beta_1, \ldots, \beta_L, \gamma_{1,1}, \ldots, \gamma_{1,K}, \ldots, \gamma_{R,1}, \ldots, \gamma_{R,K}, \theta_1, \ldots, \theta_Q)$ denotes the full set of parameters. We substitute $\mu_{\rho_{ij}}$ from Equation (5.1) with Equation (3.6) which can then be maximized with respect to each parameter coefficient vector. The likelihood is known to be globally concave (Beggs et al., 1981). This means that if a maximum is found, it is guaranteed to be a global rather than a local maximum.

Most of the papers in this area used numerical optimization algorithms such as Newton-Raphson algorithm (Beggs et al., 1981; Hausman and Ruud, 1987; Kamakura and Mazzon, 1991). Allison and Christakis (1994); Kumar and Kant (2007) used the Cox regression model in SAS (PHREG procedure). The resulting log-likelihood function in Equation (5.2) is equivalent to the Cox proportional hazards model. The Cox proportional hazards model calculates estimates from the rank ordering of survival times among observations (Cox, 1972).

We adopt instead an extension of the Minorization-Maximization (MM) algorithm that was proposed for the PL model by Hunter (2004) to fit this model. Moreover, the standard methods such as Newton Raphson method, for finding the estimated parameters can be applied. The Newton-Raphson algorithm is used in order to find estimates of item-specific covariates and

ranker-item specific covariates. The Newton-Raphson algorithm does not behave well for estimating $\boldsymbol{\lambda}$ when there are too many parameters. However, it can be used to find estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

5.1.2 MM Algorithm

We use the MM algorithm from Hunter (2004), as in Chapter 3, to estimate the parameters. Our log-likelihood function in Equation (5.2) is awkward to maximize because of the second term. We exploit the supporting hyperplane property of convex functions as shown in Equation (3.2) and Equation (3.3).

In Equation (3.3), let $x = \sum_{m=j}^{p_i} \mu_{\rho_{im}}$ and $y = \sum_{m=j}^{p_i} \mu^*_{\rho_{im}}$ where $\mu^*_{\rho_{im}}$ denotes the estimate of $\mu_{\rho_{im}}$ from the previous iteration. The inequality becomes

$$-\log\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right) \geq 1 - \log\left(\sum_{m=j}^{p_i} \mu^*_{\rho_{im}}\right) - \frac{\sum_{m=j}^{p_i} \mu_{\rho_{im}}}{\sum_{m=j}^{p_i} \mu^*_{\rho_{im}}}$$

and the Q function becomes

$$Q\left(\boldsymbol{\zeta}, \boldsymbol{\zeta}^*\right) = \sum_{i=1}^{n} \sum_{j=1}^{p_i-1} \left[\log\left(\mu_{\rho_{ij}}\right) + 1 - \log\left(\sum_{m=j}^{p_i} \mu^*_{\rho_{im}}\right) - \frac{\sum_{m=j}^{p_i} \mu_{\rho_{im}}}{\sum_{m=j}^{p_i} \mu^*_{\rho_{im}}}\right]. \quad (5.3)$$

By the construction of the Q function, $Q\left(\boldsymbol{\zeta}, \boldsymbol{\zeta}^*\right) \leq \ell\left(\boldsymbol{\zeta}\right)$, with equality if and only if $\boldsymbol{\zeta} = \boldsymbol{\zeta}^*$. The MM algorithm involves finding $\boldsymbol{\zeta}^{(*+1)}$ which maximizes $Q\left(\boldsymbol{\zeta}, \boldsymbol{\zeta}^*\right)$ with respect to $\boldsymbol{\zeta}$. Then

$$\ell\left(\boldsymbol{\zeta}^{(*+1)}\right) \geq Q\left(\boldsymbol{\zeta}^{(*+1)}, \boldsymbol{\zeta}^*\right) \geq Q\left(\boldsymbol{\zeta}^*, \boldsymbol{\zeta}^*\right) = \ell\left(\boldsymbol{\zeta}^*\right),$$

with equality only if $\boldsymbol{\zeta} = \boldsymbol{\zeta}^*$ and this sequence of $\boldsymbol{\zeta}^*$ values is guaranteed to increase the likelihood.

In the original MM algorithm for the PL model, the maximization $Q\left(\boldsymbol{\zeta}, \boldsymbol{\zeta}^*\right)$ with respect to $\boldsymbol{\zeta}$ can be done explicitly. Once the regression model is involved, the maximization step at each iteration will itself be iterative. Therefore, the

algorithm will inevitably be slower.

For the maximization step, it is convenient to separate the full set of parameters into $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\theta}$ parameters. The $\boldsymbol{\lambda}$ parameters and $\boldsymbol{\gamma}$ parameters can be optimized explicitly when fixing the other parameters. However, the $\boldsymbol{\beta}$ parameters and $\boldsymbol{\theta}$ parameters cannot be done without iteration. Thus, it is better to estimate each parameter separately.

The Q function from the Equation (5.3) can be simplified, for optimization, by omitting terms that do not depend on $\boldsymbol{\mu}$. The Q function becomes

$$\begin{aligned} Q\left(\boldsymbol{\zeta}, \boldsymbol{\zeta}^*\right) &= \sum_{i=1}^{n} \sum_{j=1}^{p_i-1} \left[\log \mu_{\rho_{ij}} - \frac{\sum_{m=j}^{p_i} \mu_{\rho_{im}}}{\sum_{m=j}^{p_i} \mu^*_{\rho_{im}}} \right] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{p_i-1} \left[\log \mu_{\rho_{ij}} - c^*_{ij} \sum_{m=j}^{p_i} \mu_{\rho_{im}} \right], \end{aligned}$$

where

$$c^*_{ij} = \frac{1}{\sum_{m=j}^{p_i} \mu^*_{\rho_{im}}}.$$

Optimization of $\boldsymbol{\lambda}$

We consider only $\boldsymbol{\lambda}$ parameters while the other parameters are fixed. Let $\bar{\boldsymbol{\beta}}$, $\bar{\boldsymbol{\gamma}}$, and $\bar{\boldsymbol{\theta}}$ denote these fixed parameters and let

$$\bar{\mu}_{\lambda,\rho_{ij}} = \exp\left(\lambda_{\rho_{ij}} + \sum_{l=1}^{L} \bar{\beta}_l z_{l,\rho_{ij}} + \sum_{r=1}^{R} \bar{\gamma}_{r,\rho_{ij}} x_{r,i} + \sum_{q=1}^{Q} \bar{\theta}_q w_{q,\rho_{ij}} \right).$$

The Q function with $\bar{\mu}_{\lambda,\rho_{ij}}$ is

$$Q\left(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*\right) = \sum_{i=1}^{n} \sum_{j=1}^{p_i-1} \left[\log\left(\bar{\mu}_{\lambda,\rho_{ij}}\right) - c^*_{ij} \sum_{m=j}^{p_i} \bar{\mu}_{\lambda,\rho_{im}} \right].$$

Differentiating $Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*)$ with respect to λ_k gives

$$\frac{\partial Q}{\partial \lambda_k} = \sum_{i=1}^{n} \sum_{j=1}^{p_i-1} \left[\eta_{ijk} - c^*_{ij} \sum_{m=j}^{p_i} \delta_{imk} \bar{\mu}_{\lambda,\rho_{im}} \right],$$

where η_{ijk} and δ_{imk} are indicator functions such that

$$\eta_{ijk} = \begin{cases} 1, & \text{if } \rho_{ij} = k, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\delta_{imk} = \begin{cases} 1, & k \in \{\rho_{im}, \ldots, \rho_{ip_i}\}, \\ 0, & \text{otherwise.} \end{cases}$$

Setting the log-likelihood derivative to zero yields

$$\hat{\lambda}_k = \log \left[\frac{\sum_{i=1}^{n} \sum_{j=1}^{p_i-1} \eta_{ijk}}{\sum_{i=1}^{n} \sum_{j=1}^{p_i-1} c_{ij}^{*} \sum_{m=j}^{p_i} \delta_{imk} \mu'_{\lambda,\rho_{im}}} \right],$$

where $\mu'_{\lambda,\rho_{im}} = \exp\left(\sum_{l=1}^{L} \bar{\beta}_l z_{l,\rho_{im}} + \sum_{r=1}^{R} \bar{\gamma}_{r,\rho_{im}} x_{r,i} + \sum_{q=1}^{Q} \bar{\theta}_q w_{q,\rho_{im}}\right)$.

Optimization of $\boldsymbol{\beta}$

The Newton-Raphson method is used in order to optimize the Q function for finding the effect of item-specific covariates. Therefore, the first and second derivative are required. The Q function with fixed $\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\gamma}}$, and $\bar{\boldsymbol{\theta}}$ parameters is used to find the first and second derivatives. Suppose

$$\bar{\mu}_{\beta,\rho_{im}} = \exp\left(\bar{\lambda}_{\rho_{im}} + \sum_{l=1}^{L} \beta_l z_{l,\rho_{im}} + \sum_{r=1}^{R} \bar{\gamma}_{r,\rho_{im}} x_{r,i} + \sum_{q=1}^{Q} \bar{\theta}_q w_{q,\rho_{im}} \right),$$

then the first and second derivative with respect to β_l are

$$Q'(\beta_l) = \frac{\partial Q}{\partial \beta_l} = \sum_{i=1}^{n} \sum_{j=1}^{p_i-1} \left[z_{l,\rho_{ij}} - c_{ij}^{*} \sum_{m=j}^{p_i} z_{l,\rho_{im}} \bar{\mu}_{\beta,\rho_{im}} \right],$$

and

$$Q''(\beta_l) = \frac{\partial^2 Q}{\partial \beta_l^2} = -\sum_{i=1}^{n} \sum_{j=1}^{p_i-1} c_{ij}^{*} \sum_{m=j}^{p_i} z_{l,\rho_{im}}^2 \bar{\mu}_{\beta,\rho_{im}},$$

respectively. The iteration in the Newton-Raphson method is given by

$$\beta_l = \beta_l^* - \frac{Q'(\beta_l^*)}{Q''(\beta_l^*)}. \tag{5.4}$$

The advantage of using the Newton-Raphson algorithm is that it converges quickly in low-dimensional problems. The equation above is for one item-specific covariate. If we introduce more than one item-specific covariate into the model then Equation (5.4) becomes

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* - \mathbf{H}^{-1} Q'(\boldsymbol{\beta}^*),$$

where $\mathbf{H}$ is the Hessian matrix. The way to calculate Hessian matrix is shown in Section 5.1.4.

Optimization of $\boldsymbol{\gamma}$

The optimization of $\boldsymbol{\gamma}$ parameters, the ranker-specific covariates, can be done in the same way as $\boldsymbol{\lambda}$ parameters. Let

$$\bar{\mu}_{\gamma,\rho_{im}} = \exp\left(\bar{\lambda}_{\rho_{im}} + \sum_{l=1}^{L} \bar{\beta}_l z_{l,\rho_{im}} + \sum_{r=1}^{R} \gamma_{r,\rho_{im}} x_{r,i} + \sum_{q=1}^{Q} \bar{\theta}_q w_{q,\rho_{im}}\right),$$

and the $Q(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$ function with fixed $\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\beta}}$, and $\bar{\boldsymbol{\theta}}$ parameters is

$$Q(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) = \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \left[\log\left(\bar{\mu}_{\gamma,\rho_{ij}}\right) - c_{ij}^* \sum_{m=j}^{p_i} \bar{\mu}_{\gamma,\rho_{im}}\right].$$

We differentiate $Q(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*)$ with respect to $\gamma_{r,k}$ giving

$$\frac{\partial Q}{\partial \gamma_{r,k}} = \sum_{i=1}^{n} \sum_{j=1}^{p_i - 1} \left[\eta_{ijk} x_{r,i} - c_{ij}^* \sum_{m=j}^{p_i} x_{r,i} \delta_{imk} \bar{\mu}_{\gamma,\rho_{im}}\right].$$

Setting $\frac{\partial Q}{\partial \gamma_{r,k}}$ equal to zero gives

$$\gamma_{r,k} = \log \left[\frac{\sum_{i=1}^{n} \sum_{j=1}^{p_i-1} \eta_{ijk} x_{r,i}}{\sum_{i=1}^{n} \sum_{j=1}^{p_i-1} c_{ij}^{*} \sum_{m=j}^{p_i} x_{r,i} \delta_{imk} \mu'_{\gamma,\rho_{im}}} \right],$$

where $\mu'_{\gamma,\rho_{im}} = \exp\left(\bar{\lambda}_{\rho_{im}} + \sum_{l=1}^{L} \bar{\beta}_l z_{l,\rho_{im}} + \sum_{q=1}^{Q} \bar{\theta}_q w_{q,\rho_{im}}\right)$. The total number of γ parameters is $R \times K$ and one of $1, \ldots, K$ in γ_r must be set equal to 0 in order to achieve identifiability. This algorithm works when $x_{r,\cdot}$ is a dummy variable.

Optimization of $\boldsymbol{\theta}$

The parameters associated with ranker-item-specific covariates can be estimated by using the same method used for item-specific covariates. Suppose

$$\bar{\mu}_{\theta,\rho_{im}} = \exp\left(\bar{\lambda}_{\rho_{im}} + \sum_{l=1}^{L} \bar{\beta}_l z_{l,\rho_{im}} + \sum_{r=1}^{R} \bar{\gamma}_{r,\rho_{im}} x_{r,i} + \sum_{q=1}^{Q} \theta_q w_{q,\rho_{im}} \right).$$

The first and second derivatives with respect to θ_q become

$$Q'(\theta_q) = \frac{\partial Q}{\partial \theta_q} = \sum_{i=1}^{n} \sum_{j=1}^{p_i-1} \left[w_{q,\rho_{ij}} - c_{ij}^{*} \sum_{m=j}^{p_i} w_{q,\rho_{im}} \bar{\mu}_{\theta,\rho_{im}} \right],$$

and

$$Q''(\theta_q) = \frac{\partial^2 Q}{\partial \theta_q^2} = -\sum_{i=1}^{n} \sum_{j=1}^{p_i-1} c_{ij}^{*} \sum_{m=j}^{p_i} w_{q,\rho_{im}}^2 \bar{\mu}_{\theta,\rho_{im}},$$

respectively. The Newton-Raphson algorithm for a single ranker-item-specific covariate is given as

$$\theta_q = \theta_q^{*} - \frac{Q'\left(\theta_q^{*}\right)}{Q''\left(\theta_q^{*}\right)}.$$

For more than one ranker-item-specific covariates, the estimation is

$$\boldsymbol{\theta} = \boldsymbol{\theta}^{*} - \mathbf{H}^{-1} Q'(\boldsymbol{\theta}^{*}).$$

5.1.3 Existing Package in R for the ROL Model

The software R programming is considered. The `mlogit` package, which is implemented by Croissant (2013), enables the estimation of the ROL model. However, this package is only for full rankings, where each ranker indicates their preference for all of the alternatives. We implement the `ROLmm` based on our algorithm in Section 5.1.2. The `ROLmm` can estimate partial ranked data.

The Game data which is included in this package is used in order to compare results from our algorithm and `mlogit`. In this dataset, there are 6 gaming platforms ($K = 6$) and rankers are asked to rank all of them. We consider two covariates here. First, a ranker-item-specific covariate, own is a dummy variable where 1 if the ranker owns the platform. Second, a ranker-specific covariate, hours is the number of hours spent on gaming per week. We group this covariate to make a new dummy covariate which is equal to 1 if the time spent is more than 3.5 hours per week and 0 if less than 3.5 hours. The `ROLmm` algorithm and `mlogit` package are used to fit the model to the Game data including own and hours as a categorical covariate (hoursInd).

There are two data formats, which are wide format and long format, provided in the package. We consider Game2 dataset with long format since this is same format as we use in the `ROLmm`. The `mlogit` function requires data format from `mlogit.data` function. That means we need to transform Game2 dataset into another format in order to use `mlogit` function. The log-likelihood values are the same which are -522.74. Using PC as the reference platform, the estimates are shown in Figure 5.1. We get same results from both algorithms. This experiment is conducted on a Toshiba notebook with Intel Core i5-3210M and 8 GB RAM. The `mlogit` computational times with and without the `mlogit.data` to transform data format are 2.55 seconds and 0.17 seconds, respectively. The `ROLmm` used 8.52 seconds to analyze the data. Thus, the `mlogit` algorithm is faster than the `ROLmm`.

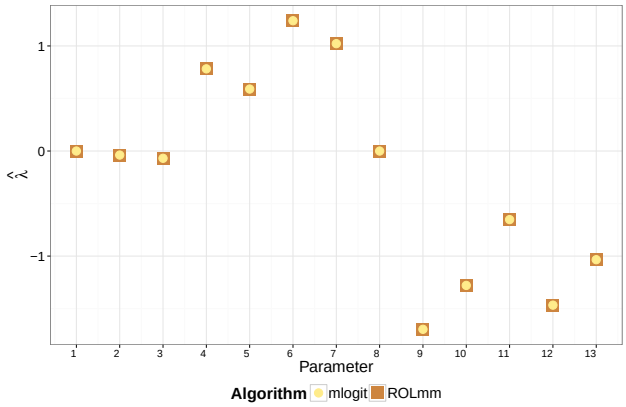


Figure 5.1: Parameter estimates for the `ROLmm` and the `mlogit` algorithms when own and hoursInd are included in the ROL model

We remove the 6^{th} position from all records in Game2 dataset in order to obtain partial ranking data. We attempted to apply the `mlogit.data` function to this data; however, it does not work. We do not investigate this function any further. We conclude that the `mlogit` package cannot analyze partial data because the `mlogit` function requires the data format from the `mlogit.data` function.

Next, we compare the `ROLmm` with the `optim` function in R. We compare three optimization methods in this function. The three methods are Nelder Mead (NM), Broyden-Fletcher-Goldfard-Shanno (BFGS), and the limited-memory modification of the BFGS (L-BFGS-B). Due to computational time, we compare these methods by fitting the PL model. The data are generated under the PL model with $K = 50$, $p = 10$, and $n = 200$ and the `optim` function is used to fit the model. This process is repeated for 50 times and then reported mean computational time and MSE for each method. The computational times are 8.02, 45.80, and 29.78 seconds from NM, BFGS, and L-BFGS-B methods, respectively. Based on the computational times, the NM method is the fastest and follows by the L-BFGS-B and the BFGS methods. The MSE are 5.858, 0.128, and 0.143 for NM, BFGS and L-BFGS-B methods. The NM method has not converged. One possible reason is that the NM method does

not require gradient but it relies only on evaluations of the objective function. In this case, there are too many parameters for such a simple optimizer. The MSE shows that the BFGS method is the best method. Therefore, the BFGS method is chosen because it gives the lowest MSE and the computational time is not much slower than the L-BFGS-B method. Moreover, both BFGS and L-BFGS-B methods are quasi-Newton method and require gradients. The BFGS method uses an approximation of the inverse Hessian matrix while the L-BFGS-B method approximates the BFGS method.

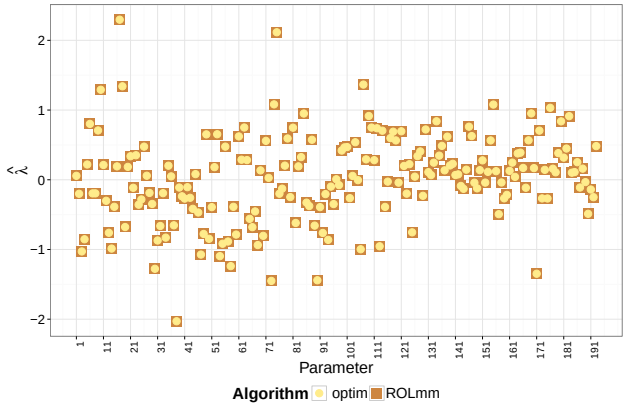


Figure 5.2: Parameter estimates for the `ROLmm` and the `optim` function when Familiarity and Gender are included in the ROL model

We apply the `ROLmm` and the `optim` function to the Group I data from the Animal dataset. Here we consider two covariates, Familiarity and Gender, where Familiarity is a ranker-item-specific covariate and Gender is a ranker-specific covariate. We use the same initial values and Hessian option in `optim` is set to FALSE in order to compare computational times. We provide only the log-likelihood function to the `optim` function. The computation times are 820.43 seconds and 112.49 seconds from the `optim` function and the `ROLmm` algorithm, respectively. Both algorithms give the same log-likelihood value which is -6156.55. The estimates are the same as presented in Figure 5.2.

Our algorithm, `ROLmm`, performs faster than the `optim` function and both algorithms give the same results. Thus, we use the `ROLmm` algorithm in later analysis.

5.1.4 Observed Information Matrix for the ROL Model

The observed information matrix can be calculated as the negative of Hessian. The Hessian matrix is the matrix of second derivatives of the log-likelihood function. The log-likelihood function is

$$\ell(\boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{j=1}^{p_i} \left[\log(\mu_{\rho_{ij}}) - \log \left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right) \right] \tag{5.5}$$

where $\mu_{\rho_{ij}}$ is from Equation (5.1). We use the log-likelihood function in Equation (5.5) to find first and second derivatives. The Hessian matrix is

$$\mathbf{H} = \begin{bmatrix} ① \ \frac{\partial^2 \ell}{\partial \lambda^2} & ⑤ \ \frac{\partial^2 \ell}{\partial \lambda \partial \beta} & ⑥ \ \frac{\partial^2 \ell}{\partial \lambda \partial \gamma} & ⑦ \ \frac{\partial^2 \ell}{\partial \lambda \partial \theta} \\ & ② \ \frac{\partial^2 \ell}{\partial \beta^2} & ⑧ \ \frac{\partial^2 \ell}{\partial \beta \partial \gamma} & ⑨ \ \frac{\partial^2 \ell}{\partial \beta \partial \theta} \\ & & ③ \ \frac{\partial^2 \ell}{\partial \gamma^2} & ⑩ \ \frac{\partial^2 \ell}{\partial \gamma \partial \theta} \\ & & & ④ \ \frac{\partial^2 \ell}{\partial \theta^2} \end{bmatrix}$$

and it can be found separately for each ranker and then summed over rankers, because rankings by different rankers are independent. Thus, a single ranker is considered in order to obtain the first and second derivatives. Expressions for the numbered components ① to ⑩ are given below.

① $\frac{\partial^2 \ell}{\partial \lambda^2}$

$$\frac{\partial \ell}{\partial \lambda_{\rho_{ir}}} = 1 - \sum_{j=1}^{r} \frac{\mu_{\rho_{ir}}}{\sum_{m=j}^{p_i} \mu_{\rho_{im}}}, \quad r \in 1, \ldots, p_i$$

$$\frac{\partial^2 \ell}{\partial \lambda_{\rho_{ir}}^2} = - \sum_{j=1}^{r} \frac{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right) \mu_{\rho_{ir}} - (\mu_{\rho_{ir}})^2}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right)^2}$$

$$\frac{\partial^2 \ell}{\partial \lambda_{\rho_{ir}} \partial \lambda_{\rho_{it}}} = \sum_{j=1}^{r} \frac{\mu_{\rho_{ir}} \mu_{\rho_{it}}}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right)^2}, \quad r < t.$$

② $\frac{\partial^2 \ell}{\partial \beta^2}$

$$\frac{\partial \ell}{\partial \beta_s} = \sum_{j=1}^{p_i} \left[z_{s,\rho_{ij}} - \frac{\sum_{m=j}^{p_i} z_{s,\rho_{im}} \mu_{\rho_{im}}}{\sum_{m=j}^{p_i} \mu_{\rho_{im}}} \right]$$

$$\frac{\partial^2 \ell}{\partial \beta_s^2} = -\sum_{m=j}^{p_i} \left[\frac{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right) \left(\sum_{m=j}^{p_i} z_{s,\rho_{im}}^2 \mu_{\rho_{im}}\right) - \left(\sum_{m=j}^{p_i} z_{s,\rho_{im}} \mu_{\rho_{im}}\right)^2}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right)^2} \right]$$

$$\frac{\partial^2 \ell}{\partial \beta_s \partial \beta_t} = -\sum_{j=1}^{p_i} \frac{1}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right)^2} \left[\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right) \left(\sum_{m=j}^{p_i} z_{s,\rho_{im}} z_{t,\rho_{im}} \mu_{\rho_{im}}\right) - \left(\sum_{m=j}^{p_i} z_{s,\rho_{im}} \mu_{\rho_{im}}\right) \left(\sum_{m=j}^{p_i} z_{t,\rho_{im}} \mu_{\rho_{im}}\right) \right], \quad s < t$$

③ $\frac{\partial^2 \ell}{\partial \gamma^2}$

$$\frac{\partial \ell}{\partial \gamma_{a,\rho_{ir}}} = x_{a,i} - \sum_{j=1}^{r} \frac{x_{a,i} \mu_{\rho_{ir}}}{\sum_{m=j}^{p_i} \mu_{\rho_{im}}}$$

$$\frac{\partial^2 \ell}{\partial \gamma_{a,\rho_{ir}}^2} = -\sum_{j=1}^{r} \left[\frac{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right) \left(x_{a,i}^2 \mu_{\rho_{ir}}\right) - \left(x_{a,i} \mu_{\rho_{ir}}\right)^2}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right)^2} \right]$$

$$\frac{\partial^2 \ell}{\partial \gamma_{a,\rho_{ir}} \partial \gamma_{a,\rho_{it}}} = \sum_{j=1}^{r} \frac{x_i \mu_{\rho_{ir}} \mu_{\rho_{it}}}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right)^2}, \quad r < t$$

④ $\frac{\partial^2 \ell}{\partial \theta^2}$

$$\frac{\partial \ell}{\partial \theta_g} = \sum_{j=1}^{p_i} \left[w_{g,\rho_{ij}} - \frac{\sum_{m=j}^{p_i} w_{g,\rho_{im}} \mu_{\rho_{im}}}{\sum_{m=j}^{p_i} \mu_{\rho_{im}}} \right]$$

$$\frac{\partial^2 \ell}{\partial \theta_g^2} = -\sum_{j=1}^{p_i} \left[\frac{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right) \left(\sum_{m=j}^{p_i} w_{g,\rho_{im}}^2 \mu_{\rho_{im}}\right) - \left(\sum_{m=j}^{p_i} w_{g,\rho_{im}} \mu_{\rho_{im}}\right)^2}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right)^2} \right]$$

$$\frac{\partial^2 \ell}{\partial \theta_g \partial \theta_h} = -\sum_{j=1}^{p_i} \frac{1}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right)^2} \left[\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}}\right) \left(\sum_{m=j}^{p_i} w_{g,\rho_{im}} w_{h,\rho_{im}} \mu_{\rho_{im}}\right) - \left(\sum_{m=j}^{p_i} w_{g,\rho_{im}} \mu_{\rho_{im}}\right) \left(\sum_{m=j}^{p_i} w_{h,\rho_{im}} \mu_{\rho_{im}}\right) \right], \quad g < h$$

⑤ $\frac{\partial^2 \ell}{\partial \lambda \partial \beta}$

$$\frac{\partial^2 \ell}{\partial \lambda_{\rho_{ir}} \partial \beta_s} = -\sum_{j=1}^{r} \left[\frac{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right) \left(z_{s,\rho_{ir}} \mu_{\rho_{ir}} \right) - \left(\mu_{\rho_{ir}} \right) \left(\sum_{m=j}^{p_i} z_{s,\rho_{im}} \mu_{\rho_{im}} \right)}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right)^2} \right]$$

⑥ $\frac{\partial^2 \ell}{\partial \lambda \partial \gamma}$

$$\frac{\partial^2 \ell}{\partial \lambda_{\rho_{ir}} \partial \gamma_{\rho_{ir}}} = -\sum_{j=1}^{r} \left[\frac{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right) \left(x_i \mu_{\rho_{ir}} \right) - \left(\mu_{\rho_{ir}} \sum_{m=j}^{p_i} x_i \mu_{\rho_{im}} \right)}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right)^2} \right]$$

⑦ $\frac{\partial^2 \ell}{\partial \lambda \partial \theta}$

$$\frac{\partial^2 \ell}{\partial \lambda_{\rho_{ir}} \partial \theta_g} = -\sum_{j=1}^{r} \left[\frac{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right) \left(w_{g,\rho_{ir}} \mu_{\rho_{ir}} \right) - \left(\mu_{\rho_{ir}} \right) \left(\sum_{m=j}^{p_i} w_{g,\rho_{im}} \mu_{\rho_{im}} \right)}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right)^2} \right]$$

⑧ $\frac{\partial^2 \ell}{\partial \beta \partial \gamma}$

$$\frac{\partial^2 \ell}{\partial \gamma_{s,\rho_{ir}} \partial \beta_l} = -\sum_{j=1}^{r} \left[\frac{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right) \left(x_{s,i} z_{l,\rho_{ir}} \mu_{\rho_{ir}} \right) - \left(x_{s,i} \mu_{\rho_{ir}} \right) \left(\sum_{m=j}^{p_i} z_{l,\rho_{im}} \mu_{\rho_{im}} \right)}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right)^2} \right]$$

⑨ $\frac{\partial^2 \ell}{\partial \beta \partial \theta}$

$$\frac{\partial^2 \ell}{\partial \theta_g \partial \beta_l} = -\sum_{j=1}^{p_i} \frac{1}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right)^2} \left[\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right) \left(\sum_{m=j}^{p_i} w_{g,\rho_{im}} \mu_{\rho_{im}} \right) - \left(\sum_{m=j}^{p_i} w_{g,\rho_{im}} \mu_{\rho_{im}} \right) \left(\sum_{m=j}^{p_i} z_{l,\rho_{im}} \mu_{\rho_{im}} \right) \right]$$

⑩ $\frac{\partial^2 \ell}{\partial \gamma \partial \theta}$

$$\frac{\partial^2 \ell}{\partial \gamma_{s,\rho_{ir}} \partial \theta_g} = -\sum_{j=1}^{r} \frac{1}{\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right)^2} \left[\left(\sum_{m=j}^{p_i} \mu_{\rho_{im}} \right) \left(x_{s,i} w_{g,\rho_{ir}} \mu_{\rho_{ir}} \right) \right.$$

$$- (x_{s,i}\mu_{\rho_{ir}}) \left(\sum_{m=j}^{p_i} w_{g,\rho_{im}} \mu_{\rho_{im}} \right) \Bigg]$$

5.1.5 The `optim` Function versus the `ROLinfm` Algorithm for the Observed Information Matrix for the ROL Model

The `ROLinfm` algorithm is implemented in order to calculate the observed information matrix for the ROL model. The `optim` function is considered where the Hessian option is set to TRUE. The `ROLinfm` needs two inputs which are a dataset and estimates from the ROL model. We apply the `optim` function to the Group I data from the Animal dataset with one covariate, Familiarity. We get the estimates and the Hessian matrix. These estimates are used as input in the `ROLinfm` algorithm. The computational time is 1.55 seconds from the `ROLinfm` algorithm. The computational time from the `ROLmm` algorithm is 232.27 seconds then the total computation time for our algorithm is 233.82 seconds. The computational time is 1790.16 seconds from the `optim` function. Our algorithms are faster than the `optim` function.

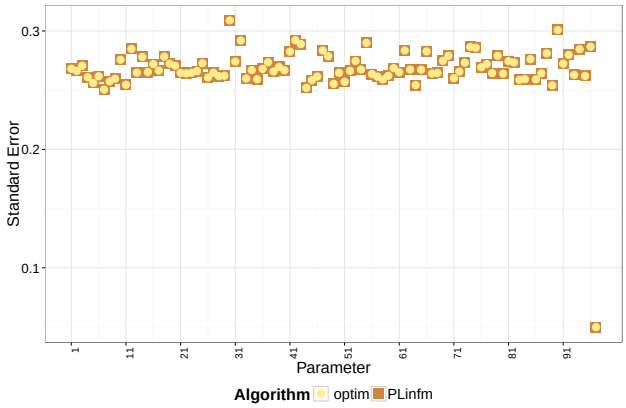


Figure 5.3: Standard errors from the `ROLinfm` and the `optim` function when fitting the Group I data with Familiarity from the Animal dataset with the ROL model where the 97^{th} parameter is for Familiarity

We compute standard errors from the observed information matrix from the `ROLinfm` algorithm and from the Hessian matrix from the `optim` function

in order to compare results. The results from the `ROLinfm` and the `optim` are shown in Figure 5.3. Figure 5.3 shows that both of them give the same results.

5.2 Benter Model

Benter (1994) proposed a model which is another type of extension of the PL model. Hausman and Ruud (1987) suggested, based on their experience and study, that rankers chose their higher preferences more carefully than the lower preferences. The additional parameters introduced by the Benter model can express this effect. The Benter model has two kinds of parameters which are item preference parameters ($\lambda_{\rho_{ij}}$) and dampening parameters (α_j). Each ranker receives the same number of items to rank, p. In the Benter model, the probability of the ranking ρ_i is

$$\begin{aligned} P\left(\rho_i; \boldsymbol{\mu}\right) &= \frac{\mu_{\rho_{i1}}^{\alpha_1}}{\mu_{\rho_{i1}}^{\alpha_1} + \cdots + \mu_{\rho_{ip}}^{\alpha_1}} \times \cdots \times \frac{\mu_{\rho_{i(p-1)}}^{\alpha_{p-1}}}{\mu_{\rho_{i(p-1)}}^{\alpha_{p-1}} + \mu_{\rho_{ip}}^{\alpha_{p-1}}} \times \frac{\mu_{\rho_{ip}}^{\alpha_p}}{\mu_{\rho_{ip}}^{\alpha_p}} \\ &= \prod_{j=1}^{p-1} \frac{\mu_{\rho_{ij}}^{\alpha_j}}{\sum_{m=j}^{p} \mu_{\rho_{im}}^{\alpha_j}}, \end{aligned} \tag{5.6}$$

where $\mu_{\rho_{ij}} = \exp\left(\lambda_{\rho_{ij}}\right)$. The Benter model is characterized by the parameters α_j and constrained to α_j that satisfy $0 \leq \alpha_j \leq 1$ for all $j = 1, \ldots, p$. This ensures that preferences for lower ranked items are at least as random as higher preference ones. For example, $\alpha = 0.9$, it means that the probability is dampened to model the effect where the second preference was made less care than the first preference. To avoid over-parameterization problems, α_1 and α_p are defined to be equal to 1 and 0, respectively. The PL model is a special case of the Benter model when all α_j equal to 1 (Gormley and Murphy, 2008).

5.2.1 Maximum Likelihood Estimator

The log-likelihood function of the Benter model has the following expression

$$\ell(\boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{j=1}^{p-1} \left[\log\left(\mu_{\rho_{ij}}^{\alpha_j}\right) - \log\left(\sum_{m=j}^{p} \mu_{\rho_{im}}^{\alpha_j}\right) \right], \qquad 0 \leq \alpha_j \leq 1. \tag{5.7}$$

The ML estimates $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ from this log-likelihood function is not straightforward. We consider the MM algorithm to optimize this log-likelihood function, following Gormley and Murphy (2008).

5.2.2 MM Algorithm

The log-likelihood function in Equation (5.7) is difficult to maximize because of the second term, as in the PL model and the ROL model. However, the μ^{α} terms cause a problem which is different from the previous models. Gormley and Murphy (2008) proposed the following MM algorithm for fitting the Benter model.

Optimization of $\boldsymbol{\mu}$

The optimization of $\boldsymbol{\mu}$ where $\mu_{\rho_{ij}} = \exp(\lambda_{\rho_{ij}})$ is the same as what we have done for the $\boldsymbol{\lambda}$ parameters of the ROL model. The α_j is treated as a constant $\bar{\alpha}_j$ here. The negative inequality logarithm function in Equation (3.3) in Chapter 3 is applied to the second term of Equation (5.7) with $x = \sum_{m=j}^{p} \mu_{\rho_{im}}^{\bar{\alpha}_j}$ and $y = \sum_{m=j}^{p} \mu_{\rho_{im}}^{*\ \bar{\alpha}_j}$ where $\mu_{\rho_{im}}^{*}$ is the estimate of $\mu_{\rho_{im}}$ from the previous iteration. The inequality becomes

$$-\log\left(\sum_{m=j}^{p} \mu_{\rho_{im}}^{\bar{\alpha}_j}\right) \geq 1 - \log\left(\sum_{m=j}^{p} \mu_{\rho_{im}}^{*\ \bar{\alpha}_j}\right) - \frac{\sum_{m=j}^{p} \mu_{\rho_{im}}^{\bar{\alpha}_j}}{\sum_{m=j}^{p} \mu_{\rho_{im}}^{*\ \bar{\alpha}_j}},$$

and the Q function becomes

$$Q(\boldsymbol{\mu}, \boldsymbol{\mu}^{*}) = \sum_{i=1}^{n} \sum_{j=1}^{p-1} \left[\log\left(\mu_{\rho_{ij}}^{\bar{\alpha}_j}\right) + 1 - \log\left(\sum_{m=j}^{p} \mu_{\rho_{im}}^{*\ \bar{\alpha}_j}\right) - \frac{\sum_{m=j}^{p} \mu_{\rho_{im}}^{\bar{\alpha}_j}}{\sum_{m=j}^{p} \mu_{\rho_{im}}^{*\ \bar{\alpha}_j}} \right].$$

We can simplify the Q function by omitting the terms which do not depend on $\boldsymbol{\mu}$, then

$$Q\left(\boldsymbol{\mu}, \boldsymbol{\mu}^*\right) = \sum_{i=1}^{n} \sum_{j=1}^{p-1} \left[\log\left(\mu_{\rho_{ij}}^{\bar{\alpha}_j}\right) - \frac{\sum_{m=j}^{p} \mu_{\rho_{im}}^{\bar{\alpha}_j}}{\sum_{m=j}^{p} \mu_{\rho_{im}}^{*\ \bar{\alpha}_j}} \right].$$

Let

$$c_{ij}^* = \frac{1}{\sum_{m=j}^{p} \mu_{\rho_{im}}^{*\ \bar{\alpha}_j}},$$

then the Q function becomes

$$Q\left(\boldsymbol{\mu}, \boldsymbol{\mu}^*\right) = \sum_{i=1}^{n} \sum_{j=1}^{p-1} \left[\log\left(\mu_{\rho_{ij}}^{\bar{\alpha}_j}\right) - c_{ij}^* \sum_{m=j}^{p} \mu_{\rho_{im}}^{\bar{\alpha}_j} \right].$$

We solve the maximization problem with respect to $\mu_{\rho_{ij}}$ by modifying the Q function once again. Letting $f(\mu) = -\mu^{\bar{\alpha}}$ and $f(\mu^*) = -\mu^{*\bar{\alpha}}$ since μ is an exponential function and by the Equation 3.2 then the inequality becomes

$$-\mu^{\bar{\alpha}} \geq -\mu^{*\bar{\alpha}} - \bar{\alpha}(\mu^*)^{\bar{\alpha}-1}\left(\mu - \mu^*\right).$$

The equation above is applied to the second term in the Q function then the Q function becomes

$$\begin{aligned} Q\left(\boldsymbol{\mu}, \boldsymbol{\mu}^*\right) = \sum_{i=1}^{n} \sum_{j=1}^{p-1} \Bigg[& \log\left(\mu_{\rho_{ij}}^{\bar{\alpha}_j}\right) - c_{ij}^* \Bigg(\sum_{m=j}^{p} \mu_{\rho_{im}}^{*\ \bar{\alpha}_j} + \sum_{m=j}^{p} \bar{\alpha}_j \mu_{\rho_{im}} \left(\mu_{\rho_{im}}^*\right)^{\bar{\alpha}_j - 1} \\ & - \sum_{m=j}^{p} \bar{\alpha}_j \mu_{\rho_{im}}^* \left(\mu_{\rho_{im}}^*\right)^{\bar{\alpha}_j - 1} \Bigg) \Bigg] \\ \equiv \sum_{i=1}^{n} \sum_{j=1}^{p-1} & \left[\log\left(\mu_{\rho_{ij}}^{\bar{\alpha}_j}\right) - c_{ij}^* \left(\sum_{m=j}^{p} \bar{\alpha}_j \mu_{\rho_{im}} \left(\mu_{\rho_{im}}^*\right)^{\bar{\alpha}_j - 1} \right) \right]. \end{aligned}$$

The Q function above contains only the parts that depend on $\boldsymbol{\mu}$. We substitute

$\mu_{\rho_{ij}} = \exp\left(\lambda_{\rho_{ij}}\right)$ and the Q function becomes

$$Q\left(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*\right) = \sum_{i=1}^{n} \sum_{j=1}^{p-1} \left[\bar{\alpha}_j \lambda_{\rho_{ij}} - c_{ij}^* \sum_{m=j}^{p} \bar{\alpha}_j \exp\left(\lambda_{\rho_{im}}\right) \left(\exp\left(\lambda_{\rho_{im}}^*\right)\right)^{\bar{\alpha}_j - 1} \right].$$

We maximize $Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^*)$ by differentiating with respect to λ_k for estimating the $\boldsymbol{\lambda}$ parameters. The first derivative is

$$\frac{\partial Q}{\partial \lambda_k} = \sum_{i=1}^{n} \sum_{j=1}^{p-1} \left[\bar{\alpha}_j \, \eta_{ijk} - c_{ij}^* \sum_{m=j}^{p} \bar{\alpha}_j \exp(\lambda_{\rho_{im}}) \left(\exp(\lambda_{im}^*)\right)^{\bar{\alpha}_j - 1} \delta_{imk} \right],$$

where

$$\eta_{ijk} = \begin{cases} 1, & \text{if } \rho_{ij} = k \\ 0, & \text{otherwise} \end{cases}$$

and

$$\delta_{imk} = \begin{cases} 1, & \text{if } k \in \{\rho_{im}, \ldots, \rho_{ip}\} \\ 0, & \text{otherwise.} \end{cases}$$

We set the first derivative equal to zero and the estimated parameter becomes

$$\hat{\lambda}_k = \log \left[\frac{\sum_{i=1}^{n} \sum_{j=1}^{p-1} \bar{\alpha}_j \, \eta_{ijk}}{\sum_{i=1}^{n} \sum_{j=1}^{p-1} c_{ij}^* \sum_{m=j}^{p} \bar{\alpha}_j \left(\exp(\lambda_{\rho_{im}}^*)\right)^{\bar{\alpha}_j - 1} \delta_{imk}} \right]. \tag{5.8}$$

Optimization of $\boldsymbol{\alpha}$

In order to find $\hat{\boldsymbol{\alpha}}$, we treat the $\boldsymbol{\lambda}$ in Equation (5.7) as constant. As for $\boldsymbol{\lambda}$, Equation (3.3) in Chapter 3 is considered. The surrogate function is applied to the second term of Equation (5.7) with $x = \sum_{m=j}^{p} \bar{\mu}^{\alpha_j}$ and $y = \sum_{m=j}^{p} \bar{\mu}^{\alpha_j^*}$ where $\bar{\mu}^{\alpha_j^*}$ is the estimate of α_j from the previous iteration. The inequality becomes

$$-\log\left(\sum_{m=j}^{p} \bar{\mu}_{\rho_{im}}^{\alpha_j}\right) \geq 1 - \log\left(\sum_{m=j}^{p} \bar{\mu}_{\rho_{im}}^{\alpha_j^*}\right) \frac{\sum_{m=j}^{p} \bar{\mu}_{\rho_{im}}^{\alpha_j}}{\sum_{m=j}^{p} \bar{\mu}_{\rho_{im}}^{\alpha_j^*}}.$$

The Q function which omits the terms that do not depend on α_j is

$$Q\left(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*\right) = \sum_{i=1}^{n} \sum_{j=1}^{p-1} \left[\log\left(\bar{\mu}_{\rho_{ij}}^{\alpha_j}\right) - c_{ij}^* \sum_{m=j}^{p} \bar{\mu}_{\rho_{im}}^{\alpha_j} \right]$$

where

$$c_{ij}^* = \frac{1}{\sum_{m=j}^{p} \bar{\mu}_{\rho_{im}}^{\alpha_j^*}}.$$

The Q function above is still difficult to optimize and needs to be modified further. The function, $f(\alpha) = -\bar{\mu}^{\alpha}$, is a concave function. In order to find a convex function, $f(\alpha)$ around α^*, this can be done by applying a quadratic function (Lange et al., 2000) and let $x = \alpha$ and $y = \alpha^*$ then

$$f(x) \leq f(y) + f'(y)(x - y) + \frac{1}{2}(x - y)^{\intercal} B (x - y)$$

where $B - H(y) > 0$, $B > 0$ and H is the Hessian. Therefore,

$$\begin{aligned}
\bar{\mu}^{\alpha} &\leq \bar{\mu}^{\alpha^*} + \log\left(\bar{\mu}\right) \bar{\mu}^{\alpha^*} \left(\alpha - \alpha^*\right) + \frac{1}{2} \left(\alpha - \alpha^*\right)^{\intercal} B \left(\alpha - \alpha^*\right) \\
&\leq \bar{\mu}^{\alpha^*} + \log\left(\bar{\mu}\right) \bar{\mu}^{\alpha^*} \left(\alpha - \alpha^*\right) + \frac{1}{2} \left(\alpha - \alpha^*\right)^2 \left(\log(\bar{\mu})\right)^2 \\
-\bar{\mu}^{\alpha} &\geq -\bar{\mu}^{\alpha^*} - \log\left(\bar{\mu}\right) \bar{\mu}^{\alpha^*} \left(\alpha - \alpha^*\right) - \frac{1}{2} \left(\alpha - \alpha^*\right)^2 \left(\log(\bar{\mu})\right)^2
\end{aligned}$$

and $\left(\log(\bar{\mu})\right)^2 > H(\alpha^*)$. The surrogate function after applying the quadratic function to the second term of the Q function is

$$\begin{aligned}
Q\left(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*\right) = \sum_{i=1}^{n} \sum_{j=1}^{p-1} \Bigg[\log\left(\bar{\mu}_{\rho_{ij}}^{\alpha_j}\right) - c_{ij}^* \Bigg(& \sum_{m=j}^{p} \bar{\mu}_{\rho_{im}}^{\alpha_j^*} + \sum_{m=j}^{p} \log\left(\bar{\mu}_{\rho_{im}}\right) \bar{\mu}_{\rho_{im}}^{\alpha_j^*} \left(\alpha_j - \alpha_j^*\right) \\
& + \sum_{m=j}^{p} \frac{1}{2} \left(\alpha_j - \alpha_j^*\right)^2 \left(\log\left(\bar{\mu}_{\rho_{im}}\right)\right)^2 \Bigg) \Bigg].
\end{aligned}$$

We iteratively maximize the $Q(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ function by taking the derivative with

respect to α_j. Thus,

$$\frac{\partial Q}{\partial \alpha_j} = \sum_{i=1}^{n} \left[\log\left(\bar{\mu}_{\rho_{ij}}\right) - c_{ij}^{*} \sum_{m=j}^{p} \left(\log\left(\bar{\mu}_{\rho_{im}}\right) \bar{\mu}_{\rho_{im}}^{\alpha_j^*} + \left(\alpha_j - \alpha_j^*\right) \left(\log\left(\bar{\mu}_{\rho_{im}}\right)\right)^2 \right) \right],$$

which implies that

$$\begin{aligned} \hat{\alpha}_j &= \frac{\sum_{i=1}^{n} \left[\log\left(\bar{\mu}_{\rho_{ij}}\right) + c_{ij}^{*} \sum_{m=j}^{p} \left(-\log\left(\bar{\mu}_{\rho_{im}}\right) \bar{\mu}_{\rho_{im}}^{\alpha_j^*} + \alpha_j^* \left(\log\left(\bar{\mu}_{\rho_{im}}\right)\right)^2 \right) \right]}{\sum_{i=1}^{n} c_{ij}^{*} \sum_{m=j}^{p} \left(\log\left(\bar{\mu}_{\rho_{im}}\right)\right)^2} \\ &= \frac{\sum_{i=1}^{n} \left[\bar{\lambda}_{\rho_{ij}} + c_{ij}^{*} \sum_{m=j}^{p} \left(-\bar{\lambda}_{\rho_{im}} \exp\left(\bar{\lambda}_{\rho_{im}}\right)^{\alpha_j^*} + \alpha_j^* \bar{\lambda}_{\rho_{im}}^2 \right) \right]}{\sum_{i=1}^{n} c_{ij}^{*} \sum_{m=j}^{p} \bar{\lambda}_{\rho_{im}}^2}. \end{aligned} \tag{5.9}$$

The algorithm for finding the parameter estimates of the Benter model is shown in Algorithm 7.

Algorithm 7 Benter model

1: Initialize parameter estimates $\boldsymbol{\lambda}^{(0)}$, $\boldsymbol{\alpha}^{(0)}$, and $h = 0$
2: $\boldsymbol{\lambda}^{(0)}$ is the parameter estimates from the PL model
3: $\boldsymbol{\alpha}^{(0)} = 0.5$ for all $\boldsymbol{\alpha}^{(0)}$
4: repeat
5: compute $c_{ij}^{(h)}$ based on $\boldsymbol{\lambda}^{(h)}$ and $\boldsymbol{\alpha}^{(h)}$
6: increment h
7: compute
8: $\boldsymbol{\lambda}^{(h)}$ by using Equation (5.8)
9: $\boldsymbol{\alpha}^{(h)}$ by using Equation (5.9)
10: until converged

5.2.3 The `optim` Function versus the Algorithm for the Benter Model

To the best of our knowledge, there is currently no package in R for fitting the Benter model. We implemented the `BMmm` based on the algorithm in the previous section. We compare the `BMmm` with `optim` function. The BFGS optimization method is selected and the Hessian option is set to FALSE for the `optim` function. We apply the `optim` function and `BMmm` algorithm to the Group I data from the Animal dataset. The computational times are

638.03 and 821.69 seconds for the `optim` function and for the `BMmm` algorithm, respectively.

Our algorithm, `BMmm`, performs slower than the `optim` function and both algorithms give the same results. However, we still consider the `BMmm` algorithm since we are going to introduce covariates to the Benter model.

5.2.4 Observed Information Matrix for the Benter Model

The rankings from different rankers are independent. The information matrix can be found directly from the negative of the Hessian matrix. The Hessian matrix can be found by differentiating the log-likelihood function. The log-likelihood function is

$$\begin{aligned}\ell(\boldsymbol{\mu}) &= \sum_{i=1}^{n}\sum_{j=1}^{p}\left[\log(\mu_{\rho_{ij}}^{\alpha_j}) - \log\left(\sum_{m=j}^{p}\mu_{\rho_{im}}^{\alpha_j}\right)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{p}\left[\alpha_j\lambda_{\rho_{ij}} - \log\left(\sum_{m=j}^{p}\exp\left(\lambda_{\rho_{im}}\right)^{\alpha_j}\right)\right].\end{aligned}$$

The Hessian is the matrix of second partial derivatives of the log-likelihood function, specifically

$$\mathbf{H} = \begin{bmatrix} ① \ \frac{\partial^2\ell}{\partial\lambda^2} & ③ \ \frac{\partial^2\ell}{\partial\lambda\partial\alpha} \\ ③ \ \frac{\partial^2\ell}{\partial\lambda\partial\alpha} & ② \ \frac{\partial^2\ell}{\partial\alpha^2} \end{bmatrix}.$$

A single ranker is considered in order to find the first and second derivatives for the Hessian matrix. First, we consider ① in the Hessian matrix. The first derivative of the log-likelihood function is

$$\frac{\partial\ell}{\lambda_{\rho_{ir}}} = \alpha_r - \sum_{j=1}^{r}\frac{\alpha_j\exp(\lambda_{\rho_{ir}})^{\alpha_j}}{\sum_{m=j}^{p}\exp(\lambda_{\rho_{im}})^{\alpha_j}}, \qquad r = 1,\ldots,p$$

The second derivative becomes

$$\frac{\partial^2 \ell}{\partial \lambda_{\rho_{ir}}^2} = -\sum_{j=1}^{r} \frac{\left(\sum_{m=j}^{p} \exp(\lambda_{\rho_{im}})^{\alpha_j}\right)\left(\alpha_j^2 \exp(\lambda_{\rho_{ir}})^{\alpha j}\right) - \left(\alpha_j \exp(\lambda_{\rho_{ir}})^{\alpha_j}\right)^2}{\left(\sum_{m=j}^{p} \exp(\lambda_{\rho_{im}})^{\alpha_j}\right)^2},$$

and the elements of the off-diagonal matrix where $r < t$ are

$$\frac{\partial^2 \ell}{\partial \lambda_{\rho_{ir}} \partial \lambda_{\rho_{it}}} = \sum_{j=1}^{r} \frac{\alpha_j^2 \exp(\lambda_{\rho_{ir}})^{\alpha_j} \exp(\lambda_{\rho_{it}})^{\alpha_j}}{\left(\sum_{m=j}^{p} \exp(\lambda_{im})^{\alpha_j}\right)^2}.$$

Second, we find the first and second derivatives for ② . The first and second derivatives with respect to α are

$$\frac{\partial \ell}{\partial \alpha_j} = \lambda_{\rho_{ij}} - \frac{\sum_{m=j}^{p} \lambda_{\rho_{im}} \exp\left(\lambda_{\rho_{im}}\right)^{\alpha_j}}{\sum_{m=j}^{p} \exp\left(\lambda_{\rho_{im}}\right)^{\alpha_j}}$$

and

$$\frac{\partial^2 \ell}{\partial \alpha_j^2} = -\frac{1}{\left(\sum_{m=j}^{p} \exp\left(\lambda_{\rho_{im}}\right)^{\alpha_j}\right)^2} \left[\left(\sum_{m=j}^{p} \exp\left(\lambda_{\rho_{im}}\right)^{\alpha_j}\right)\left(\sum_{m=j}^{p} \lambda_{\rho_{im}}^2 \exp\left(\lambda_{\rho_{im}}\right)^{\alpha_j}\right) - \left(\sum_{m=j}^{p} \lambda_{\rho_{im}} \exp\left(\lambda_{\rho_{im}}\right)^{\alpha_j}\right)^2\right].$$

The off-diagonal elements are

$$\frac{\partial^2 \ell}{\partial \alpha_j \partial \alpha_t} = 0.$$

Final part is ③ which is the second derivative with respect to both λ and α. The second derivative with respect to α_j and $\lambda_{\rho_{ij}}$ is

$$\frac{\partial^2 \ell}{\partial \alpha_j \partial \lambda_{\rho_{ij}}} = 1 - \frac{1}{\left(\sum_{m=j}^{p} \exp\left(\lambda_{\rho_{im}}\right)^{\alpha_j}\right)^2} \left[\left(\sum_{m=j}^{p} \exp\left(\lambda_{im}\right)^{\alpha_j}\right)\left(\lambda_{\rho_{ij}} \alpha_j \exp(\lambda_{\rho_{ij}})^{\alpha_j} + \exp(\lambda_{\rho_{ij}})^{\alpha_j}\right) - \left(\sum_{m=j}^{p} \lambda_{\rho_{im}} \exp(\lambda_{\rho_{im}})^{\alpha_j}\right)\left(\alpha_j \exp(\lambda_{\rho_{ij}})^{\alpha_j}\right)\right].$$

The second derivative with respect to α_j and $\lambda_{\rho_{it}}$ when $j < t$ is

$$\frac{\partial^2 \ell}{\partial \alpha_j \partial \lambda_{\rho_{it}}} = -\frac{1}{\left(\sum_{m=j}^{p} \exp\left(\lambda_{\rho_{im}}\right)^{\alpha_j}\right)^2} \left[\left(\sum_{m=j}^{p} \exp(\lambda_{\rho_{im}})^{\alpha_j} \right) \left(\lambda_{\rho_{it}} \alpha_j \exp(\lambda_{\rho_{it}}) \right. \right.$$
$$\left. \left. + \exp(\lambda_{\rho_{it}})^{\alpha_j} \right) - \left(\sum_{m=j}^{p} \lambda_{\rho_{im}} \exp(\lambda_{\rho_{im}})^{\alpha_j} \right) \left(\alpha_j \exp\left(\lambda_{\rho_{it}}\right)^{\alpha_j} \right) \right].$$

The overall Hessian matrix is obtained by summing these terms over all rankers. This can be done because rankings are independent.

5.2.5 The `optim` Function versus the Algorithm for the Observed Information Matrix for the Benter Model

We implement the `BMinfm` to calculate the observed information matrix. The `optim` function is used in order to compare the Hessian matrix. We compute standard errors from both algorithms to show that they give the same results. Figure 5.4 shows that we get same results.

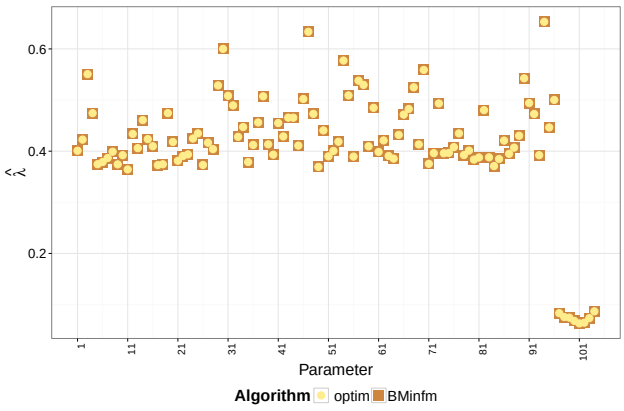


Figure 5.4: Standard errors from the `BMinfm` and the `optim` function when fitting the Group I data with the Benter model where the 97^{th} to the 104^{th} parameters are standard errors of the dampening parameters

5.3 Combining the ROL and the Benter Models

The two extensions of the PL model, the ROL model and the Benter model, may be combined to give a model that incorporates covariates and also allows for a dampening effect. We call this model the combined model for short. To fit this model, we adopt the MM algorithm that has been used for fitting the ROL model and the Benter model.

The probability of the ranking ρ_i is again given by Equation (5.6) where $\mu_{\rho_{ij}}$ is now given by Equation (5.1). The parameter estimates can be found in a similar way to the ROL and Benter models. The Q function becomes

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{p} \left[\log(\mu_{\rho_{ij}}^{\alpha_j}) - c_{ij}^{*} \left(\sum_{m=j}^{p} \alpha_j \mu_{\rho_{im}} (\mu_{\rho_{im}}^{*})^{\alpha_j - 1} \right) \right]. \tag{5.10}$$

For example, to estimate $\boldsymbol{\lambda}$ parameters, the Q function is

$$Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{*}) = \sum_{i=1}^{n} \sum_{j=1}^{p-1} \left[\log(\bar{\mu}_{\lambda,\rho_{ij}}^{\bar{\alpha}_j}) - c_{ij}^{*} \left(\sum_{m=j}^{p} \bar{\alpha}_j \bar{\mu}_{\lambda,\rho_{ij}} \left(\mu_{\lambda,\rho_{im}}^{*} \right)^{\bar{\alpha}_j - 1} \right) \right],$$

where λ^{*} is λ from the previous iteration, then

$$\mu_{\lambda,\rho_{im}}^{*} = \exp\left(\lambda_{\rho_{ij}}^{*} + \sum_{l=1}^{L} \bar{\beta}_l z_{l,\rho_{ij}} + \sum_{r=1}^{R} \bar{\gamma}_{r,\rho_{ij}} x_{r,i} + \sum_{q=1}^{Q} \bar{\theta}_q w_{q,i\rho_{ij}} \right).$$

Differentiating $Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{*})$ with respect to λ_k and setting this equals to zero gives

$$\lambda_k = \log \left[\frac{\sum_{i=1}^{n} \sum_{j=1}^{p-1} \bar{\alpha}_j \eta_{ijk}}{\sum_{i=1}^{n} \sum_{j=1}^{p-1} c_{ij}^{*} \left(\sum_{m=j}^{p} \bar{\alpha}_j \mu'_{\lambda,\rho_{im}} (\mu_{\lambda,\rho_{im}}^{*})^{\bar{\alpha}_j - 1} \delta_{imk} \right)} \right],$$

where

$$\mu'_{\lambda,\rho_{im}} = \exp\left(\sum_{l=1}^{L} \bar{\beta}_l z_{l,\rho_{ij}} + \sum_{r=1}^{R} \bar{\gamma}_{r,\rho_{ij}} x_{r,i} + \sum_{q=1}^{Q} \bar{\theta}_q w_{q,i\rho_{ij}} \right).$$

The other parameters can be estimated in the same way as the ROL model

with the Q function in Equation (5.10).

5.3.1 The `optim` Function versus the Algorithm for the Combined Model

We compare our algorithm, `CMmm`, with the `optim` function. The BFGS optimization method is selected and the Hessian option is set to FALSE for the `optim` function in order to compare computational times and estimates. We apply the `optim` function and the `CMmm` algorithm to the Group I data from the Animal dataset. One covariate, Familiarity, is included in the model then the computational times for fitting this model by using the `optim` function and the `CMmm` algorithm are 2465.87 seconds and 1536.36 seconds, respectively. We also get the same estimates and log-likelihood values.

Our `CMmm` algorithm performs faster than the `optim` function even though the `BMmm` algorithm is slower when there is no covariate in the model.

5.4 Likelihood Ratio Test

The Likelihood Ratio (LR) test is used for comparing nested models, where the simple model is a special case of the alternative, more general model. The test uses the likelihood function through the ratio of two maximizations. First, the maximum under the null hypothesis (H_0, simple model) and second, the maximum over the larger set of parameters permitting H_0 or an alternative (H_1, more general model) to be true.

Let $L(\boldsymbol{\zeta}; X)$ represent the likelihood function, and let Ω be the parameter space for the more general model and ω be the null hypothesis space, for the simple model. Let L_0 denote the maximized value of the likelihood function under H_0 and L_1 denote the maximized value over Ω, $H_0 \cup H_1$. Therefore, the

likelihood ratio is

$$\Lambda = \frac{\sup\limits_{\zeta \in \omega} L_0(\zeta; X)}{\sup\limits_{\zeta \in \Omega} L_1(\zeta; X)},$$

and the test statistic for the LR test, which is denoted by G^2, is

$$G^2 = -2 \log \Lambda = -2 \left(\log L_0 - \log L_1 \right).$$

The test statistic G^2 is distributed approximately as χ^2_{q-p} when H_0 is true and number of rankers (n) is large where p and q are the number of parameters in the restricted model specified by H_0 and the full model under H_1, respectively. The null hypothesis is rejected at the $100\alpha\%$ level if $G^2 > \chi^2_{(1-\alpha;q-p)}$.

Due to the boundary of a parameter space problem, the LR test is not a proper test for testing the Benter model and the combined model. However, we use the LR test in order to compare the PL model with the Benter model and the ROL model with the combined model. This is because there is no convenience solution for this problem.

5.5 Application to Animal Dataset

In this section, we investigate whether the addition of covariates and/or dampening parameters improves the fit of the PL model and how it effects the estimated preferences. We present results from fitting the ROL, the Benter, and the combined models with various different covariates of the Animal dataset. The Animal dataset includes item-specific covariates (Animal's type), ranker-specific covariates (Nationality, Gender, and Age) and ranker-item-specific covariates (Familiarity and Start Position).

Table 5.1: LR statistics for the ROL model with one covariate when fitted to the Animal dataset

	Group I			Group II			Group III			Group IV		
	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value
Familiarity	91.81	1	<0.001	98.22	1	<0.001	157.65	1	<0.001	96.80	1	<0.001
Start Position	17.85	1	<0.001	65.76	1	<0.001	52.01	1	<0.001	58.91	1	<0.001
Gender	115.72	96	0.083	128.48	87	0.003	124.38	95	0.023	161.77	103	<0.001
Age (2-level)	115.87	96	0.082	115.56	87	0.022	170.33	95	<0.001	119.55	103	0.127
Age (continuous)	130.89	96	0.010	132.81	87	0.001	179.49	95	<0.001	112.19	103	0.252
North America	123.97	96	0.029	100.28	87	0.156	103.75	95	0.253	100.21	103	0.559
Europe	115.89	96	0.082	92.26	87	0.330	86.04	95	0.733	113.99	103	0.216
North America and Europe	139.60	96	0.002	81.72	87	0.640	148.89	95	<0.001	137.78	103	0.013
Animal Type	No Information			No Information			< 0.001	1	0.987	< 0.001	1	0.977

5.5.1 ROL Model: Animal Dataset

The ROL model is applied to all groups of the Animal dataset. First, we fitted the model with one covariate at a time and applied the LR test in order to compare the ROL model with one covariate to the PL model.

One covariate at a time

• Familiarity and Start Position

LR tests for the covariates are displayed in Table 5.1. The LR statistics indicate that Familiarity is the most significant covariate in all four groups. Another strongly significant covariate is Start Position. This indicates that we should include both of these ranker-item-specific covariates in the ROL model.

Table 5.2: Parameter estimates for the ROL model when each ranker-item-specific covariate is included in the model (SE in brackets).

Covariates	Group I	Group II	Group III	Group IV
Familiarity	0.474(0.050)	0.501(0.051)	0.732(0.060)	0.626(0.065)
Start Position	0.152(0.036)	0.293(0.036)	0.238(0.033)	0.277(0.036)

Parameter estimates are shown in Table 5.2. Both Familiarity and Start Position have a positive effect. The positive effect of Familiarity means that rankers tend to rank the species that they are familiar with higher than unfamiliar species. Based on the estimates and their standard errors, it appears that Familiarity has stronger effect in Group III and Group IV when compared with Group I and Group II where species in Group I and Group II are provided from EDGE and the others are from WWF organization.

As mentioned in Chapter 2, there are 85 same species in the Group I data and the Group II data. We plot the $\hat{\mu}_{\text{Familiarity}}$ for these two groups as shown in Figure 5.5. Figure 5.5 shows that most of the estimates are below the reference line. This means the drawings (Group I) are more preferred to the

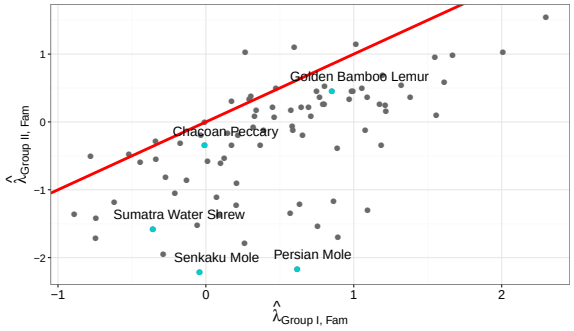


Figure 5.5: The $\hat{\mu}_{\text{Familiarity}}$ for the Group I data against the $\hat{\mu}_{\text{Familiarity}}$ for the Group II data from the Animal Dataset with the 1:1 reference line

photos (Group II) for the same species. The highlighted species show the species that have the differences of proportion of familiarity from the Group I and the Group II data higher than 95^{th} percentile.

For Start Position where 1 if top row and 0 if bottom row, the rankers are more likely to rank species in the upper row higher than species in the lower row. This may indicate a reluctance amongst rankers to move photographs between rows during the ranking process. Start Position has almost the same estimated effect in all groups, except for Group I where it is slightly lower.

• **Gender**

We investigate heterogeneity across individuals by including ranker-specific covariates in the ROL model. We begin with the Gender covariate which consists of two groups. Gender is found as moderately significant in all of the groups except Group I in which Gender is significant at the 10% level but not at the 5% significance level. Thus, we can conclude that males and females do differ in their preferences of animal species.

Considering the Group I data from the Animal dataset, the estimated parameters for Gender when this is the only covariate in the ROL model are shown in Table 5.3. Females have a stronger preference for the top 5 species than males, except for Giant Panda. The differences in coefficients for males and females have an odds interpretation and note that the coefficients compare with the reference species, Baiji. For example, since Asian Elephant

Table 5.3: Top 5 and bottom 5 parameter estimates, according to the PL model, when there is only Gender in the ROL model for the Group I data from the Animal dataset (SE in brackets)

Animal Species	PL	Male	Female	Difference
Red Panda	1.954	1.427	2.208	- 0.781(0.619)
Giant Panda	1.739	2.481	1.578	0.902(0.585)
African Elephant	1.359	1.229	1.436	- 0.207(0.603)
Fin Whale	1.288	1.125	1.427	- 0.303(0.615)
Asian Elephant	1.240	0.399	1.747	- 1.348(0.532)
⋮				
Mindanao Gymnure	- 1.052	- 1.037	- 1.116	0.079(0.587)
Eastern Sucker-footed Bat	- 1.182	- 1.370	- 1.114	- 0.256(0.615)
Chiapan Climbing Rat	- 1.226	- 0.914	- 1.335	0.421(0.598)
New Guinea Big-eared Bat	- 1.252	- 1.887	- 1.074	- 0.812(0.614)
Southern Marsupial Mole	- 1.359	- 1.385	- 1.376	- 0.009(0.711)

has a difference of -1.348 and $\exp(-1.348) = 0.26$, this means that the odds of preferring Asian Elephant to Baiji are 0.26 times as great for males as for females. To compare a specific species with other species rather than the reference species, the difference between the "Difference" in Table 5.3 of these species are considered. For example, if we compare Red Panda with Giant Panda then $\exp(-0.781 - 0.902) = 0.186$. This means the odds of males preferring Red Panda to Giant Panda is about 0.19 times the odds of females.

Another way to explore the differences of preferences between males and females is to apply an orthogonal regression model to the preference estimates. The dependent variable is $\hat{\boldsymbol{\lambda}}_{\text{Female}}$ and the independent variable is $\hat{\boldsymbol{\lambda}}_{\text{Male}}$ from the ROL model. The orthogonal regression is $\hat{\lambda}_{\text{Female-Reg},k} = 0.214 + 0.922\ \hat{\lambda}_{\text{Male},k}$. We plot $\hat{\boldsymbol{\lambda}}_{\text{Male}}$ against $\hat{\boldsymbol{\lambda}}_{\text{Female}}$ and the orthogonal regression line. We calculate distances between points and the regression line which are residuals in the regression model and show species that have distances higher than the 90^{th} percentile in Figure 5.6. Figure 5.6 shows that males prefer Giant Panda, Vaquita, Javan Rhinoceros, Saiga, and Northern Marsupial Mole more strongly than females do, while females prefer Asian

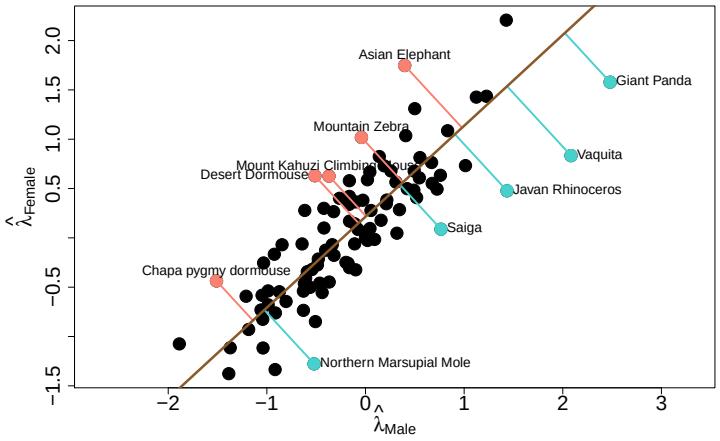


Figure 5.6: Plot $\hat{\boldsymbol{\lambda}}_{\text{Male}}$ against $\hat{\boldsymbol{\lambda}}_{\text{Female}}$ with the orthogonal regression line

Elephant, Mountain Zebra, Mount Kahuzi Climbing Mouse, Desert Dormouse, and Chapa Pygmy Dormouse.

• **Age**

The ROL model with Age as a continuous covariate is fitted by using the `optim` function since the `ROLmm` algorithm as currently implemented does not work with a continuous ranker-specific covariate. The result is that Age as a continuous covariate is significant in Group I at the 5% significance level. The age coefficients can be interpreted as $100 \times (\exp(\boldsymbol{\gamma}_{\text{Age}}) - 1)$. This is the percent change in the odds of preferring a species over Baji for each 1-year increase in Age. Moreover, Age (continuous) seems to be more significant than Age (2-level) except in Group IV where neither is significant. For Red Panda, $100 \times (\exp(-0.024) - 1)$ = -2.37%, that is with each 1 year increase in Age the odds of preferring Red Panda over Baji goes down by 2.37%. Age does not have much effect on Giant Panda (about 0.2%). Fin Whale has a positive effect when Age increases. Older people tend to prefer Fin Whale to Red Panda when they are older than 44 years old.

Age can alternatively be grouped and used as a dummy covariate similarly to Gender. We divide Age into two groups, <30 and ≥ 30 years. Age is 1 if <30 years and 0 if ≥ 30 years. The results from Age as a categorical covariate can

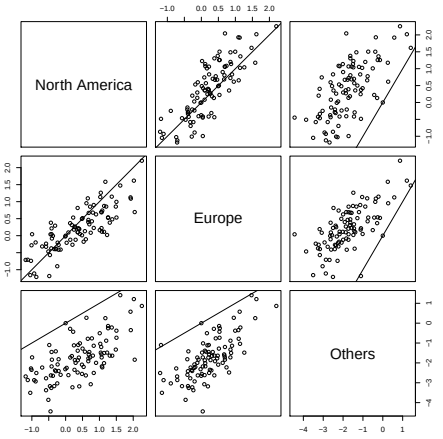
Table 5.4: Top 5 and bottom 5 parameter estimates when only Age as continuous covariate in the ROL model for the Group I data (SE in brackets)

Animal Species	PL	ROL	
		λ_{Age}	γ_{Age}
Red Panda	1.954	2.693	-0.024(0.021)
Giant Panda	1.739	1.725	-0.0002(0.021)
African Elephant	1.359	1.484	-0.005(0.021)
Fin Whale	1.288	0.467	0.026(0.021)
Asian Elephant	1.240	1.393	-0.006(0.020)
⋮			
Mindanao Gymnure	- 1.052	-0.474	-0.023(0.022)
Eastern Sucker-footed Bat	- 1.182	-0.194	-0.032(0.023)
Chiapan Climbing Rat	- 1.226	-0.747	-0.018(0.022)
New Guinea Big-eared Bat	- 1.252	-1.594	0.010(0.019)
Southern Marsupial Mole	- 1.359	-1.169	-0.009(0.021)

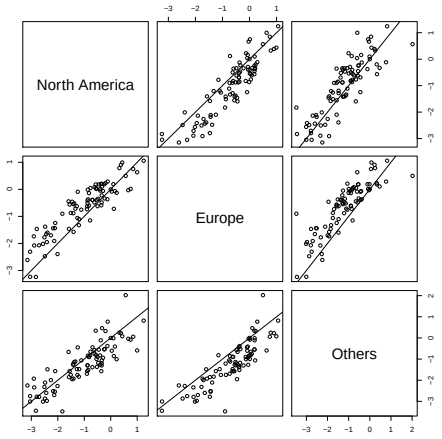
be interpreted similarly to the Gender covariate. This covariate is significant in Group II and Group III at the 5% significance level while in Group I it is significant only at the 10% significant level and it is not significant in Group IV (Table 5.1).

• **Nationality**

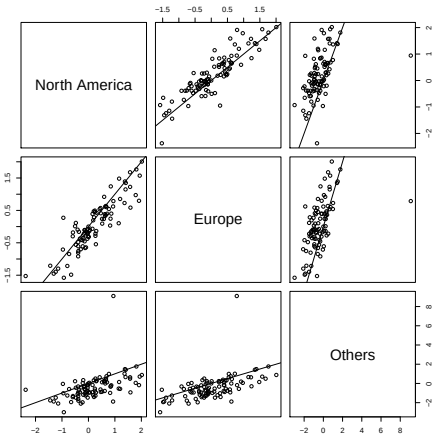
We consider three nationality groups which are North America, Europe, and other. Nationality does not have a significant effect in Groups II, III, and IV; however, it does in Group I as shown in Table 5.1. The North America nationality has a moderate significant while Europe is not significance at 5% level when compared to other nationalities. Moreover, Figure 5.7 shows that there is a correlation between North America and Europe for all groups because the plots are scattered around the reference (45°) line. Table 5.5 presents the Spearman rank correlation coefficients among pairs of nationalities. There is a strong correlation of 0.82 between North America and Europe. The plots of Americans against other and Europeans against other have an outlier as shown in Figure 5.7c. Suggestion for further investigation is to remove this outlier. However, we continue without removing the outlier and combine North



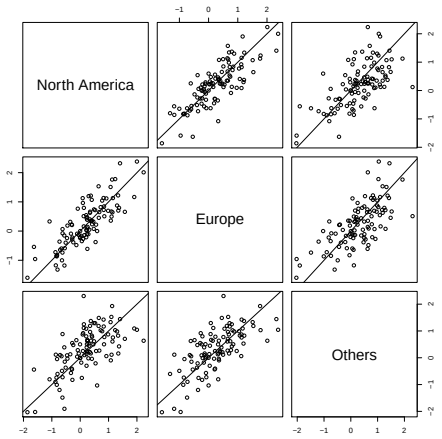
(a) Group I



(b) Group II



(c) Group III



(d) Group IV

Figure 5.7: Pairwise plots of parameter estimates for the ROL model in which the preference parameters differ between the three Nationality groups. Solid lines are the 1:1 lines.

Table 5.5: Spearman correlation between Nationalities for the Group I data from the Animal dataset

Nationality	North America	Europe	Other
North America	1	0.82	0.65
Europe		1	0.69
Others			1

America nationality with Europe nationality. The Spearman rank correlation coefficient between North America nationality and Europe nationality for the Group III is 0.87. We combine North America nationality with Europe na-

tionality and give a new group where Nationality = 1 if North America and Europe nationalities and 0 if other nationalities to the ROL model. The result is shown in Table 5.1. The rankers' nationality affects the preferences except in Group II.

• **Animal Type**

The only item-specific covariate in the Animal dataset is Animal Type where 1 if Mammal and 0 if other. We include Animal Type in the ROL model and the LR statistics when compared with the PL model are presented in Table 5.1. We compare mammals with other types. For Group III and Group IV, Animal Type is not significant.

Several covariates

The ROL models fitted so far have investigated the effect of fitting covariates individually. We now consider models that include several covariates. Table 5.6 shows the LR statistics when adding sequentially the covariates that were significant at the 5% significance level in Table 5.1. All the covariates which are significant in Table 5.1 are also significant in Table 5.6. Therefore, the ROL model for Group I contains three covariates which are Familiarity, Start Position, and Nationality. Group II has four covariates in the model which are Familiarity, Start Position, Gender, and Age. Group III has the most covariates in the model; Familiarity, Start Position, Gender, Age, and Nationality are in the ROL model for Group III. The final group, Group IV, includes Familiarity, Start Position, Gender, and Nationality in the model.

Table 5.8 presents the coefficients of Familiarity and Start Position when the ROL model includes the covariates displayed in Table 5.6. As before, Familiarity has a stronger effect in Group III and Group IV. In addition, the rankers have a stronger preference for species that they are familiar with rather than for species that are presented in the upper row of the display.

Table 5.6: LR statistics when adding Familiarity, Start Position, Gender, Age (2-level), and Nationality to the ROL model

	Group I			Group II			Group III			Group IV		
	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value
Familiarity	91.810	1	<0.001	98.216	1	<0.001	157.648	1	<0.001	96.801	1	<0.001
+ Start Position	19.428	1	<0.001	66.852	1	<0.001	51.700	1	<0.001	61.060	1	<0.001
+ Gender		-		121.499	87	0.009	127.096	95	0.016	162.901	103	<0.001
+ Age (2-level)		-		119.601	87	0.012	180.215	95	<0.001		-	
+ Nationality	142.356	96	0.002		-		153.010	95	<0.001	140.515	103	0.008

Table 5.7: LR statistics when adding Familiarity, Start Position, Gender, Age (continuous), and Nationality to the ROL model

	Group I			Group II			Group III			Group IV		
	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value
Familiarity	91.81	1	<0.001	98.22	1	<0.001	157.65	1	<0.001	96.80	1	<0.001
+ Start Position	19.43	1	<0.001	66.85	1	<0.001	51.70	1	<0.001	61.06	1	<0.001
+ Gender		-		121.50	87	0.009	127.10	95	0.016	162.90	103	<0.001
+ Age (continuous)	128.53	96	0.015	143.02	87	<0.001	186.75	95	<0.001		-	
+ Nationality	133.57	96	0.007		-		148.25	95	<0.001	140.52	103	0.008

Table 5.8: Parameter estimates of Familiarity and Start Position for the final ROL model in Table 5.6 (SE in brackets)

Covariates	Group I	Group II	Group III	Group IV
Familiarity	0.492(0.051)	0.507(0.053)	0.791(0.063)	0.687(0.069)
Start Position	0.166(0.037)	0.308(0.038)	0.264(0.035)	0.292(0.038)

Comparing the Familiarity and Start Position effects from Table 5.2 with Table 5.8, these show that both Familiarity and Start Position have a slightly stronger effect after including the covariates in Table 5.6 for each group, although the difference are generally small.

Considering Group I, the ROL model has three covariates which are Familiarity, Start Position, and Nationality. Table 5.9 shows the effects of Nationality in the model. The model allowing for Familiarity, Start Position,

Table 5.9: Top 5 and bottom 5 parameter estimates, according to the PL model, for the ROL model with Familiarity, Start Position, and Nationality covariates for the Animal dataset: Group I (SE in brackets)

Animal Species	North America and Europe $(\hat{\boldsymbol{\lambda}} + \hat{\boldsymbol{\gamma}}_{\mathbf{1}})$	Other $(\hat{\boldsymbol{\lambda}})$	Difference $(\hat{\boldsymbol{\gamma}}_{\mathbf{1}})$
Red Panda	2.026	0.662	1.363(0.828)
Giant Panda	1.556	0.844	0.712(0.744)
African Elephant	1.316	0.309	1.007(0.826)
Fin Whale	1.278	1.290	-0.012(1.105)
Asian Elephant	1.199	-0.057	1.256(0.792)
⋮	⋮	⋮	⋮
Mindanao Gymnure	-0.879	-3.062	2.182(0.772)
Eastern Sucker-footed Bat	-1.128	-2.910	1.782(0.792)
Chiapan Climbing Rat	-1.113	-2.467	1.353(0.806)
New Guinea Big-eared Bat	-1.156	-2.988	1.832(0.745)
Southern Marsupial Mole	-0.965	-3.654	2.688(1.049)

and Nationality can be written as $\log(\mu_{ij}) = \lambda_{ij} + \theta_1 w_{1,\rho_{ij}} + \theta_2 w_{2,\rho_{ij}} + \gamma_{1,\rho_{ij}} x_{1,i}$ where $x_1 = 1$ if North American or European, 0 if other. Table 5.9 shows that the five most preferred species and the five least preferred species are higher for North America and Europe nationalities as compared with other. Among

the top five preferred species, Fin Whale has the least difference. For the five least preferred species, the Southern Marsupial Mole has the largest difference and the odds of preferring Southern Marsupial Mole to Baji are 14.70 times as great for North Americans and Europeans as for the other nationalities.

5.5.2 ROL Model for Pairwise Comparisons: Animal Dataset

In Chapter 3, we consider the full rank-breaking method with different weightings. We extend this idea by including a covariate in the BT model with different weightings. The BTw and BTw-Sqrt weightings perform the best among them; therefore, we focus on these two weightings. The full rank-breaking method is applied to the Group I data from Animal dataset. The ROL model is applied to the Group I data with Familiarity and the BT model with the BTw and BTw-Sqrt weightings are applied to the paired data from the full rank-breaking method with Familiarity.

Table 5.10: The $\hat{\theta}$ of Familiarity for the ROL model and the BT model with the equal, BTw and BTw-Sqrt weightings and the Kendall tau correlation and MSE of the $\hat{\boldsymbol{\lambda}}$ and $\hat{\theta}$ when compared with the results from the ROL model

	$\hat{\theta}_{\text{Fam}}$	$\hat{\boldsymbol{\lambda}}$ and $\hat{\theta}$	
		Correlation	MSE
ROL	0.474	-	-
BT	0.579	0.914	0.020
BTw	0.693	0.907	0.023
BTw-Sqrt	0.629	0.944	0.003

Table 5.10 shows the BT model with equal, BTw, and BTw-Sqrt weightings give higher estimates of the Familiarity covariate. We compute the Kendall tau correlation and MSE of the estimates from the BT model with three different weightings with the estimates from the ROL model. The BTw-Sqrt weighting performs best and is followed equal (BT) and BTw weightings.

The BT model with the BTw-Sqrt weighting is an alternative way for analyzing the partial ranking data with covariates. The BT model with covariates is less complicated than the ROL model.

5.5.3 Benter Model: Animal Dataset

The Benter model is fitted to the animal dataset. The LR statistics show that the Benter model fits significantly better than the PL model for all groups, as shown in Table 5.11. The Benter model has 8 parameters more than the PL model. We consider the result from Group I in more detail, for illustration. First, the preference parameters are different between the PL model and the Benter model. Table 5.12 presents the top five and bottom five preference parameters for the PL model and the Benter model where the Benter dampening parameter estimates are $\hat{\boldsymbol{\alpha}} = (1, 0.882, 0.702, 0.641, 0.483, 0.354, 0.319, 0.262, 0.321, 0)$.

Table 5.11: LR statistics when compared the PL model with the Benter model for the Animal dataset

Group	G^2	df	p-value
I	109.43	8	<0.001
II	143.74	8	<0.001
III	41.26	8	<0.001
IV	67.78	8	<0.001

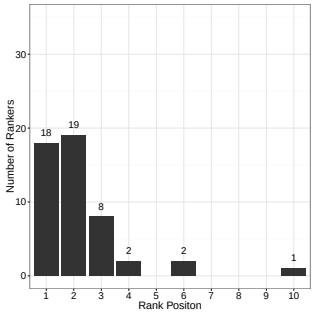
The top 5 preference species rankings according to the Benter model are the same as the PL model except Asian Elephant. The results from the Benter model show that Asian Elephant is given the third ranking while the PL model puts this species in fifth place. The $\hat{\lambda}_{\text{Giant Panda}}$ is closer to the $\hat{\lambda}_{\text{Red Panda}}$ in the Benter model. The differences between Red Panda and Giant Panda are 0.215 and 0.039 in the PL model and the Benter model, respectively. This is because the participants mainly ranked Giant Panda as their first preference as shown in Figure 5.8b. The bottom 5 rankings are not the same.

Table 5.12: Top 5 and bottom 5 of the estimated preference parameters from the PL model and Benter model for the Group I data according to the results from the PL model

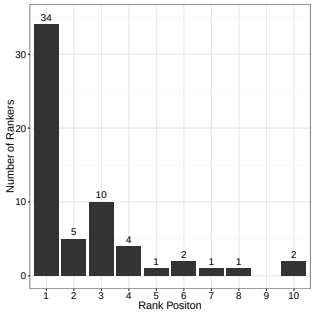
Animal Species	Average Position	First Position	PL	Benter
Red Panda	2.200	0.360	1.954	2.180
Giant Panda	2.237	0.567	1.739	2.141
African Elephant	2.973	0.297	1.359	1.561
Fin Whale	3.143	0.262	1.288	1.426
Asian Elephant	3.000	0.381	1.240	1.602
⋮	⋮	⋮	⋮	⋮
Mindanao Gymnure	8.000	0	- 1.052	- 2.615
Eastern Sucker-footed Bat	7.175	0.125	- 1.182	- 1.801
Chiapan Climbing Rat	8.116	0.023	- 1.226	- 2.960
New Guinea Big-eared Bat	7.609	0.063	- 1.252	- 2.375
Southern Marsupial Mole	7.641	0.013	- 1.359	- 2.349

The bottom five according to the Benter model are Chiapan Climbing Rat, Mindanao Gymnure, New Guinea Big-eared Bat, Southern Marsupial Mole, and Eastern Sucker-footed Bat, respectively. Considering the bottom five favourite species, the least preferred species is Chiapan Climbing Rat instead of Southern Marsupial Mole from the PL model. Figure 5.8 plots the rank distributions. Most of the rankers ranked Chiapan Climbing Rat in lower position than Southern Marsupial Mole. Therefore, Chiapan Climbing Rat has lower preference in the Benter model.

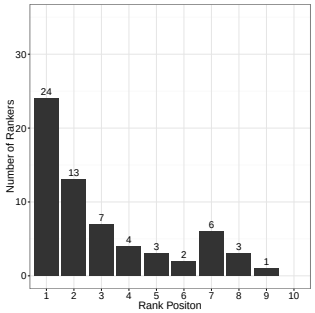
The 95% confidence interval of the preference parameters from both the PL model and the Benter model are shown in Figure 5.9. The standard errors of parameters in the Benter model are larger than those of the corresponding parameters in the PL model. The top 5 estimates from the Benter model shift to the right when compared to the PL model. Whereas, the bottom 5 shift to the left as we expect. This is because the top 5 species are tended to be ranked in the top position and reverse for the bottom 5. The 95% confidence interval of all species are shown in Figure 5.14. It can be observed that the bottom species have greater effect than the top preference species in the Benter model.



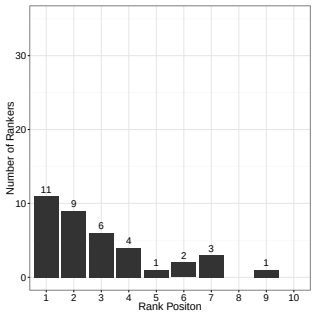
(a) Red Panda



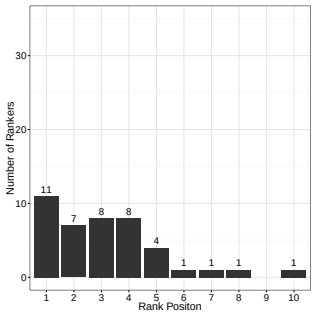
(b) Giant Panda



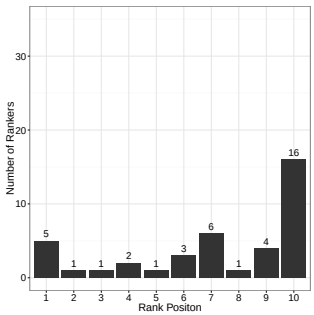
(c) Asian Elephant



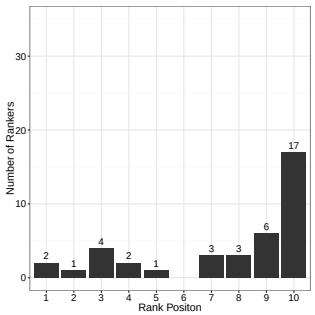
(d) African Elephant



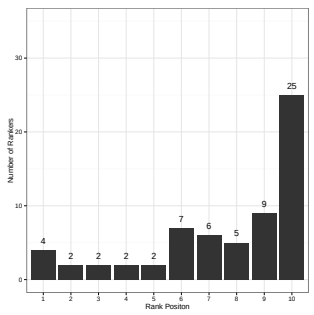
(e) Fin Whale



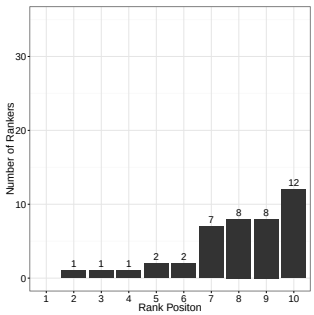
(f) Eastern Sucker-footed Bat



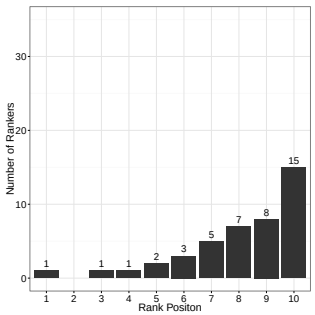
(g) Southern Marsupial Mole



(h) New Guinea Big-eared Bat



(i) Mindanao Gymnure



(j) Chiapan Climbing Rat

Figure 5.8: Rank position distributions of top 5 and bottom 5 preferred species for the Group I data from the Animal dataset

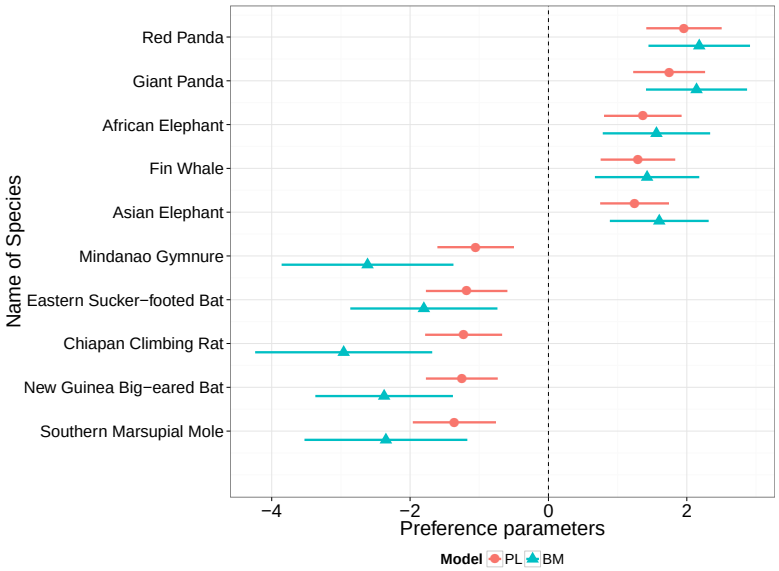


Figure 5.9: The 95% confidence interval of top 5 and bottom 5 of parameter estimates for the PL model and the Benter model for the Group I data from the Animal dataset

For more detail, the top and bottom preference species from the Benter model have stronger effect than the species which are in the middle as shown in Figure 5.14.

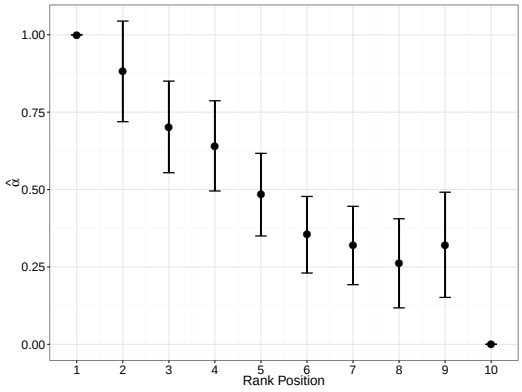


Figure 5.10: The 95% confidence interval of dampening parameter estimates for the Benter model when fitted to the Group I data from the Animal dataset

The dampening parameters are generally decreasing with rank position as shown in Figure 5.10. That means the rankers ranked their top preferences more carefully than their lower preference. The small dampening parameters values will lower the preferences of the lower-ranked species. However, α_9 is slightly higher than α_7 and α_8. The α_8 value of 0.262 suggests that the 8^{th}

place preferences are only made with one fourth of the certainty that the first preferences are.

Goodness-of-Fit for the Benter Model

A goodness of the Benter model fit is evaluated by using the bootstrap to see how well the Benter model fits the Animal dataset. We consider only the Group I data due to computational time. The number of bootstrap samples is 300 ($B = 500$) and the result is shown in Figure 5.11.

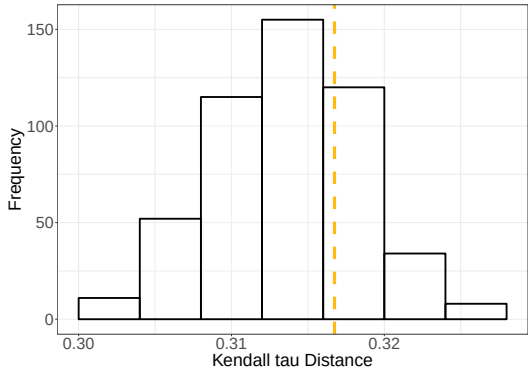


Figure 5.11: Histogram of Kendall tau distances from the bootstrapping goodness-of-fit for the Benter model where dashed line is the distance for the Group I data from the Animal dataset

Figure 5.11 shows the Kendall tau distances from the bootstrapping. The two-sided p-value is 0.572. This means the Benter model is an appropriate model for fitting the Group I data.

Next, we consider the IOS test. The IOS statistics close to zero; therefore, we compute the two-sided p-value which is 0.680. This leads to the same conclusion as the Kendall tau distance criterion.

5.5.4 Combined Model: Animal dataset

The combined ROL and Benter model has also been fitted to the Animal dataset. This model is used to find significant covariates that affect the preference of species when the dampening parameters are also in the model. The

Table 5.13: LR statistics for the combined model with one covariate when fitted to the Animal dataset

	Group I			Group II			Group III			Group IV		
	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value
Familiarity	112.91	1	<0.001	113.59	1	<0.001	174.03	1	<0.001	102.23	1	<0.001
Start Position	12.65	1	<0.001	52.29	1	<0.001	45.84	1	<0.001	59.57	1	<0.001
Gender	101.22	96	0.338	112.45	87	0.035	114.69	95	0.083	146.49	103	0.003
Age (2-level)	100.69	96	0.352	110.29	87	0.047	158.33	95	<0.001	125.81	103	0.0629
Nationality	124.55	96	0.023	70.56	87	0.900	144.26	95	<0.001	136.00	103	0.016

Table 5.14: LR statistics when adding Familiarity, Start Position, Gender, and Age to the combined model

	Group I			Group II			Group III			Group IV		
	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value
Familiarity	112.91	1	<0.001	113.59	1	<0.001	174.03	1	<0.001	102.23	1	<0.001
+ Start Position	13.18	1	<0.001	54.43	1	<0.001	43.01	1	<0.001	61.11	1	<0.001
+ Gender		-		109.09	87	0.055		-		145.20	103	0.004
+ Age (2-level)		-		112.95	87	0.032	158.50	95	<0.001		-	
+ Nationality	126.52	96	0.020		-		150.26	95	<0.001	135.07	103	0.019

LR statistics for each covariate are shown in Table 5.13. Table 5.13 shows that both Familiarity and Start Position are strongly significant among the four groups when there is only one covariate in the model. The combined models are less significant than the ROL model when including one ranker-specific covariate at a time in the models. Considering the Group I data, Gender and Age are not significant the at 5% level in either model. The North America and Europe nationalities are significant at 1% level in the ROL model while they achieve 5% significance in the combined model. In Group II, at 5% significance level, Age is significant in the ROL model, however, it is not significant in the combined model. This is the same for Gender in Group III. Group IV has the same results in both models.

Table 5.15: Parameter estimates for the combined model when including each ranker-item-specific covariate, Familiarity and Start Position, in the model for the Animal dataset (SE in brackets).

Covariates	Group I	Group II	Group III	Group IV
Familiarity	0.894(0.101)	0.910(0.100)	1.094(0.102)	1.040(0.127)
Start Position	0.201(0.057)	0.415(0.060)	0.282(0.042)	0.400(0.055)

The effects of the Familiarity and Start Position in the combined model, with only one ranker-item-specific covariate in the model, are stronger than in the ROL model with these two covariates in the ROL model as shown in Table 5.15 and Table 5.2, respectively. This is consequence of using the Benter model. However, the standard errors also increase.

We include the covariates that are significant at the 5% level in Table 5.13. The LR statistics when adding one covariate at a time to the combined model are shown in Table 5.14. All covariates that are significant in Table 5.13 are also significant at the 5% level in Table 5.14, except Gender in Group II.

The combined model allowing for Familiarity, Start Position, and Nationality for Group I can be written as $\log(\mu_{ij}) = (\lambda_{ij}+\theta_1 w_{1,\rho_{ij}}+\theta_2 w_{2,\rho_{ij}}+\gamma_{1,\rho_{ij}} x_{1,i})^{\alpha_j}$ where $x_1 = 1$ if North America or Europe, 0 if other. Table 5.16 shows that the

Table 5.16: Top 5 and bottom 5 parameter estimates for the combined model with Familiarity, Start Position, and Nationality covariates for the Group I data from the Animal dataset

Animal Species	PL	North America and Europe	Other	Difference
	$(\hat{\boldsymbol{\lambda}}_{\text{PL}})$	$(\hat{\boldsymbol{\lambda}}_{\text{Cmodel}} + \hat{\boldsymbol{\gamma}}_{\text{NorthEU}})$	$(\hat{\boldsymbol{\lambda}}_{\text{Cmodel}})$	$(\hat{\boldsymbol{\gamma}}_{\text{NorthEU}})$
Red Panda	1.954	2.105	0.697	1.408(1.028)
Giant Panda	1.739	4.105	0.478	3.627(0.913)
African Elephant	1.359	2.373	-0.192	2.565(1.034)
Fin Whale	1.288	4.761	0.834	3.927(1.274)
Asian Elephant	1.240	5.570	-0.639	6.209(1.021)
⋮	⋮	⋮	⋮	⋮
Mindanao Gymnure	- 1.052	-1.779	-5.039	3.260(1.531)
Eastern Sucker-footed Bat	- 1.182	-2.392	-4.316	1.924(1.370)
Chiapan Climbing Rat	- 1.226	-0.179	-4.197	4.018(1.598)
New Guinea Big-eared Bat	- 1.252	-0.866	-4.570	3.704(1.311)
Southern Marsupial Mole	- 1.359	-1.254	-5.262	4.008(2.138)

North American and Europe nationalities have stronger preferences than other nationalities, especially in Asian Elephant. Nationality has stronger effects in their preferences when comparing Table 5.9 with Table 5.16.

The dampening parameters are considered when adding one more covariate to the combined model. Figure 5.12 shows the estimated dampening parameters when adding more covariates to the model. The dampening parameters

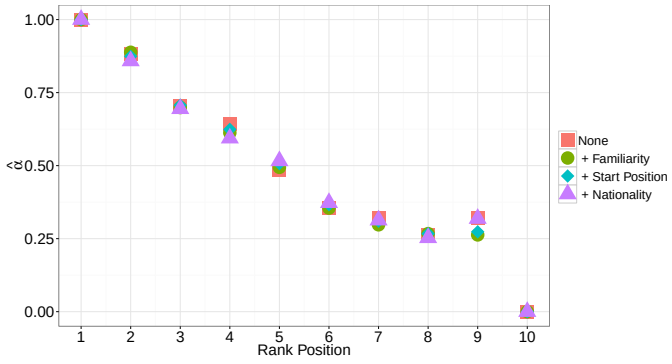


Figure 5.12: Plot of dampening parameter estimates when added one more covariate to the combined model for Group I from Animal dataset

change only slightly when adding covariates to the combined model, except the ninth rank. The $\hat{\alpha}_9$ has smaller effect when Familiarity and both Familiarity and Start Position are in the model. However, when North America and Europe nationality is added, the $\hat{\alpha}_9$ increases.

We compare the ROL model with the combined model as shown in Table 5.17. Table 5.17 shows the LR statistics when comparing the ROL model with the combined model and in brackets are p-values. The ROL model without covariate is the PL model and the combined model without covariate which is the Benter model. The combined model performs better than the ROL model since the LR statistics are strongly significant at the 0.1% level. That means the dampening parameters improve the model.

Goodness-of-Fit for the Combined Model

As previously, we perform a bootstrap to assess the goodness-of-fit statistics for the combined model. Due to the computational time, we only perform the bootstrap on the combined model with one covariate. We explore whether the Group I data is appropriately fitted by the combined model with Familiarity covariate. The bootstrap sample size is 500 ($B = 500$) and the result is shown in Figure 5.13.

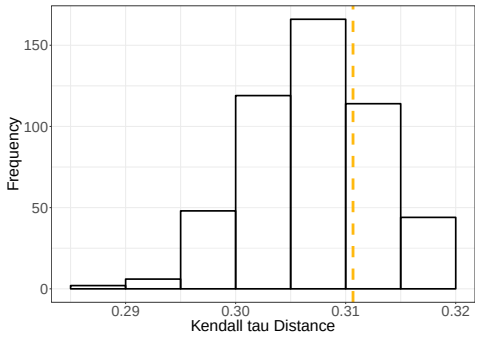


Figure 5.13: Histogram of the Kendall tau distance from the bootstrapping goodness-of-fit when $B = 150$ when fitted the combined model with the Familiarity and purple dashed line is the actual Kendall tau distance from the Group I data

Figure 5.13 shows that the Kendall tau distance from the Group I data is mostly higher than the distances from the simulated data. The 2-sided p-value is 0.560. We conclude that the combined model with the Familiarity covariate is an appropriate model for fitting the Group I data.

Another statistic is the IOS, the IOS statistics close to zero. This suggested

Table 5.17: LR statistics with 8 degree of freedom when adding covariate to the ROL model and combined model (p-value in brackets) for the Animal dataset

Covariates	ROL vs Combined			
	Group I	Group II	Group III	Group IV
None	109.43 (<0.001)	143.74 (<0.001)	41.26 (<0.001)	67.78 (<0.001)
+ Familiarity	130.53 (<0.001)	159.11 (<0.001)	57.80 (<0.001)	73.22 (<0.001)
+ Start Position	122.06 (<0.001)	146.69 (<0.001)	49.11 (<0.001)	73.25 (<0.001)
+ Gender	-	134.27 (<0.001)	-	55.56(<0.001)
+ Age	-	127.63 (<0.001)	35.63 (<0.001)	-
+ Nationality	106.22 (<0.001)	-	29.93 (<0.001)	50.12 (<0.001)

that we should compute the two-sided p-value. The two-sided p-value is 0.048. The IOS statistic is significant at 5% level but not at the 1% significance level.

5.6 Conclusions

In this chapter, we study several extensions of the PL model. First, the ROL model allows for including of covariates. There are three kinds of covariates which are item-specific, ranker-specific, and ranker-item-specific covariates. We adopt the MM algorithm from Hunter (2004) to find estimated parameters. The `ROLmm` algorithm was implemented. We compare results from the `ROLmm` with the `optim` function in order to confirm the results from the `ROLmm` algorithm and they give the same results. Moreover, the `ROLmm` algorithm performs faster than the `optim` function. Our experiments are applied on the Animal dataset. The results show that when a covariate is included, the ROL model has been shown to give an improvement when compared to the PL model. Moreover, it is easy to interpret the effects of the covariates by using an odds interpretation.

We extend the BT model with weighting, which we discussed in Chapter 3, to be able to include a covariate. Two weightings, the BTw and the BTw-Sqrt, are considered. We apply the BT model with BT, BTw, and BTw-Sqrt weightings to the Group I data from the Animal dataset. The results show that the BTw-Sqrt gives better results than the BT and BTw weightings in both Kendall tau correlation and MSE criteria.

Second, the Benter model, this model allows preferences for higher-ranked objects to be stronger than lower-ranked objects. A set of parameters, $\boldsymbol{\alpha}$, is introduced in order to take care of human ranking behaviour. We implemented the `BMmm` algorithm for fitting the Benter model. The `BMmm` gives the same result as the `optim` function; however, the `BMmm` needs more computational time. The experiment result shows that by including $\boldsymbol{\alpha}$ parameters results

in a significant improvement in the fit when compared with the PL model. The bootstrap goodness-of-fit tests with the Kendall tau distance and IOS statistics reveal the same conclusion that the Benter model is an appropriate model for fitting the Group I data from the Animal dataset.

The last presented model in this chapter is the combined model. This model combines the ROL model with the Benter model. The combined model can corporate both extensions of the PL model. It allows the inclusion of covariates and a set of parameters, $\boldsymbol{\alpha}$. We implemented the `CMmm` algorithm which is faster than the `optim` function and they give the same results. The combined model is applied to the Animal dataset. There are $p - 1$ extra parameters in the model when compared with the ROL model with the same covariates. The results from the analysis show that the combined model gives a significantly better fit. We apply the bootstrap goodness-of-fit test with the Kendall tau distance and IOS statistics to the Group I data. The Kendall tau distance statistic shows that the combined model is a suitable model; however, the IOS statistic indicates that the model is suitable at 1% significance level.

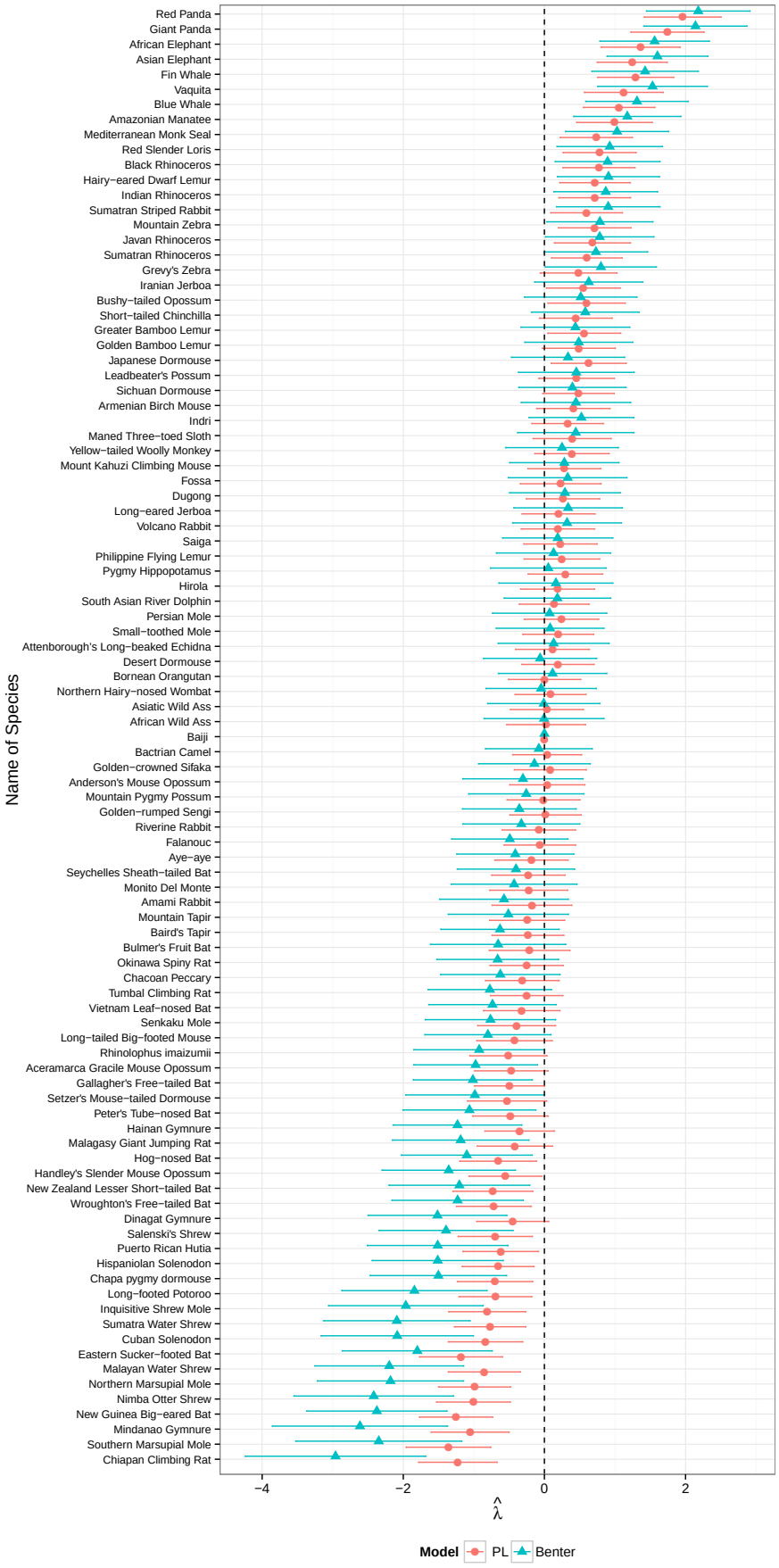


Figure 5.14: The 95% confidence interval of parameter estimates for the PL model and the Benter model for the Group I data from the Animal dataset

Chapter 6

Open Ended Rankings

In this chapter, we explore another type of partial ranking data. In the previous chapters, a set of objects is given to an individual to rank them. Now, we are interested in open ended questionnaires. For example, we ask an individual what worries them and allow the individual to mention any subjects that they are worried about at that moment. After the individual has mentioned a set of subjects, they are asked to rank these subjects according to their severity. Therefore, an open ended questionnaire allows the individuals to decide on their own what are the major issues, without bias from the researcher providing a pre-defined list of subjects to the individuals.

The purpose of this chapter is to explore this open ended ranking data. We begin with an existing method which is normally used in the sociology literature, Participatory Risk Mapping, in Section 6.1. Section 6.2 discusses tied ranking, since our real-world dataset for open ended rankings allows for ties. Two approximations, Breslow and random, are explained. The numbers of choices varies between individuals because they are allowed to mention their own choices. We discuss the number of choices in Section 6.3. In Section 6.4, we propose a new model which uses ideas from the PL model. Applications to the Sundarbans dataset are discussed in Section 6.5.

6.1 Participatory Risk Mapping

Participatory Risk Mapping (PRM) is a simple analytical tool that can be used in qualitative research in order to identify and classify risk. The PRM process comprises two stages. In the first stage, problem identification, the participant identifies problem(s), as many as he/she can think of. This is done in an open-ended fashion. The number of problems identified varies across participants. In the second stage, the participant is asked to rank order the problems he/she identified in the first step. The participant decides on their own what are the major problems rather than choosing from a given list, which may reflect the biases of the researcher.

In the analysis stage, two measures are calculated from the data, incidence and severity. The incidence of a problem is the proportion of participants who identify the problem. This proportion of incidence measure, I, shows how widespread the problem is within the population of study.

The severity measure is not straightforward to calculate because participants list different numbers of problems. For example, participant A lists two problems and ordinal rankings range from 1 to 2. Participant B lists five problems with ordinal rankings ranging from 1 to 5. The participant gives the rank 1 to the most severe problem and assigns the rank 2 to the second severe problem, etc. Inskip et al. (2013) suggested a formula for severity which is adapted from Barrett et al.'s (2001) equation. This equation can handle tied rankings. The severity index score is

$$S_{ij} = \frac{(p_{\max} + 1 - r_{ij})}{p_{\max}},$$

where S_{ij} is the severity index score for participant i and problem j, $p_{\max}$ is the maximum number of problems listed by any participant, and r_{ij} is the rank given to problem j by participant i. The value of S_{ij} ranges from 0 to 1, where

0 means the problem is not cited by participant and 1 means the problem is cited as most severe. The mean severity, S, is computed by averaging over participants who ranked that particular problem.

The PRM often includes a plot of mean severity against incidence. This plot is called risk map.

The PRM has been applied in many areas including socio-economic applications (Smith et al., 2000; Inskip et al., 2013), psychology (Chirowodza et al., 2009), public health (Fuller et al., 2014), and science (Jing et al., 2013).

6.2 Tied Data

Previously in this thesis we have assumed that rankings do not contain ties, because the datasets that we used as examples had no ties. However, in the dataset that we use in this chapter, the Sundarbans data, tied rankings of two or more objects are allowed. Tied objects, which are defined to have equal preference/severity, are assigned the same number, and any subsequently ranked objects pick up the numbering accounting for ties. For example, suppose there are four objects, A, B, C, and D, and that C is chosen to be the first, with A and B tied for second, and D the least preferred. We assign A, B, C, and D ordinal rank values of 2, 2, 1, and 4, respectively.

The likelihood for a tied ranking can be constructed by summing over all orderings that are compatible with the tied ranking. From the previous example, $A \succ B$ and $B \succ A$ are mutually exclusive events then

$$P((A \succ B) \text{ or } (B \succ A)) = P(A \succ B) + P(B \succ A).$$

Assuming a PL model, the likelihood becomes

$$L(\boldsymbol{\mu}; \rho_1) = \frac{\mu_C}{\mu_C + \mu_A + \mu_B + \mu_D} \times \left[\left(\frac{\mu_A}{\mu_A + \mu_B + \mu_D} \right) \left(\frac{\mu_B}{\mu_B + \mu_D} \right) + \right.$$

$$\left(\frac{\mu_B}{\mu_B+\mu_A+\mu_D}\right)\left(\frac{\mu_A}{\mu_A+\mu_D}\right)\Bigg] \times \frac{\mu_D}{\mu_D}.$$

It is easy to write down the likelihood in this way when the number of ties is small. However, it is difficult when the number of ties increases e.g. if we have 4 ties then there are 4! possible orderings.

Since it is problematic to find the likelihood, simplified approximations are introduced. The simplest approach to allowing for ties in the PL model is to randomly break the tied rankings. For example, since A and B are ranked second we use a random method to separate them, so that A and B are ranked respectively either second and third or third and second, with equal probability. Then, the log-likelihood function for the PL model remains the same.

Another approximation is suggested by Breslow and Crowley (1974), in which the log-likelihood function is modified as follows

$$\ell(\boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{k \in \mathcal{O}_i} \left[\log(\mu_k) - \log\left(\sum_{m=\mathcal{O}'_{i,k}}^{p_i} \mu_{\rho_{im}} \right) \right], \tag{6.1}$$

where $\mathcal{O}_i$ is a set of items that ranker i ranked and $\mathcal{O}'_{i,k}$ is a rank of item k from ranker i. We cannot cancel the last term because the summation of the second term changes in order to handle the tied rankings. For example, suppose $\mathcal{O} = \{A,\ B,\ C,\ D\}$ and $\mathcal{O}' = \{1,\ 2,\ 3,\ 3\}$ then the log-likelihood for this ranker is

$$\begin{aligned}
\ell = &(\log(\mu_A) - \log(\mu_A+\mu_B+\mu_C+\mu_D)) + \\
&(\log(\mu_B) - \log(\mu_B+\mu_C+\mu_D)) + \\
&(\log(\mu_C) - \log(\mu_C+\mu_D)) + \\
&(\log(\mu_D) - \log(\mu_C+\mu_D)).
\end{aligned}$$

Thus, the last term cannot drop out.

6.3 Number of Answers for each Ranker (p_i)

In an open-ended questionnaire, the number of objects listed varies between individuals. One possible distribution for the number of choices comes from Poisson-Binomial distribution. This distribution arises if we assume that all individuals have a potential *long list* of K objects and would report the k^{th} object in the list with $prob_k$.

Suppose $T_1, \ldots, T_K$ are independent distributed Bernoulli variables. If the probability of success is not the same for each variable then $S = T_1 + \cdots + T_K$ follows a Poisson-Binomial distribution with not-all-equal probabilities of successes, $prob_k$ where $k = 1, \ldots, K$. As a special case, S becomes a Binomial random variable when all success probabilities are equal.

Let the k^{th} ranked object be reported with probability $prob_k$. We model these probabilities through a simple logistic regression model as follows

$$\log\left(\frac{prob_k}{1 - prob_k}\right) = \alpha + \beta k, \quad k = 1, \ldots, K$$

where β is negative to give decreasing probabilities as the rankings move down. The $\hat{\alpha}$ and $\hat{\beta}$ are used to estimate $\widehat{prob}_k$ then we can use these probabilities to generate number of objects listed from the Poisson-Binomial distribution without any information of the actual number of objects listed. The results

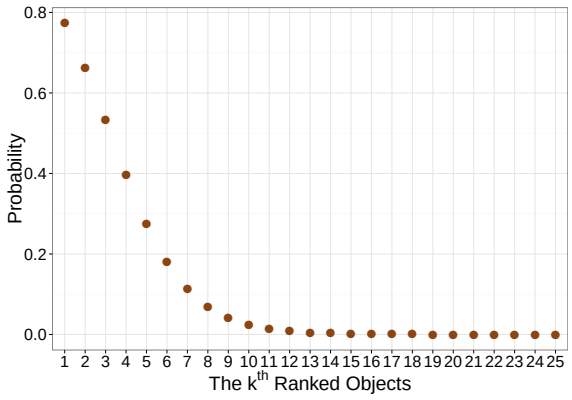


Figure 6.1: Estimated probabilities from the $\hat{\alpha}$ and $\hat{\beta}$

from fitting the logistic regression model are $\hat{\alpha} = 1.776$ and $\hat{\beta} = -0.549$. The SEs for $\hat{\alpha}$ and the $\hat{\beta}$ are 0.300 and 0.068, respectively. We calculate probabilities based on $\hat{\alpha}$ and $\hat{\beta}$. The $\widehat{prob}_k$ is shown in Figure 6.1. Figure 6.1 shows how the probability decreases when the k^{th} ranked objects increase. Only the top three objects have probability greater than 0.5.

The `truncdist` package is considered in order to generate the random number from the Poisson-Binomial distribution by setting `spec = poibin`. We use `truncdist` package because we want to generate with specific interval. We use the probabilities from the logistic regression to generate the distribution of the number of objects listed by using this package. The result is shown in Figure 6.2. Figure 6.2 shows that the estimates from the logistic regression

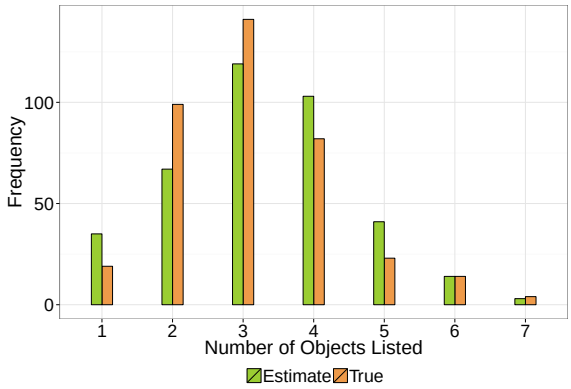


Figure 6.2: Estimated number of objects listed from the logistic regression and the true values from the Sundarbans dataset

can estimate the number of objects well. It has approximately the same shape.

6.4 Selection Preference Model

Instead of using the PL model from Chapter 3, we propose a new model, the probability of the ranking ρ_i is

$$P(\rho_i;\boldsymbol{\pi}) = \frac{\pi_{\rho_{i1}}}{1} \times \frac{\pi_{\rho_{i2}}}{1-\pi_{\rho_{i1}}} \times \frac{\pi_{\rho_{i3}}}{1-(\pi_{\rho_{i1}}+\pi_{\rho_{i2}})} \times \cdots \times \frac{\pi_{\rho_{ip_i-1}}}{1-(\pi_{\rho_{i1}}+\cdots+\pi_{\rho_{ip_i-2}})} \times \frac{\pi_{\rho_{ip_i}}}{1-(\pi_{\rho_{i1}}+\cdots+\pi_{\rho_{ip_i-1}})}$$

$$= \frac{\pi_{\rho_{i1}}}{1} \prod_{j=2}^{p_i} \frac{\pi_{\rho_{ij}}}{1 - \sum_{m=1}^{j-1} \pi_{\rho_{im}}} \tag{6.2}$$

where $\boldsymbol{\pi}$ is a vector of preference parameters and $\pi_1 + \cdots + \pi_K = 1$. We call this model as selection preference (SP) model. This model is different from the PL model. The SP model gives the probability that for a participant presented with all K and asked to pick p_i of these items in order of preference will pick $\rho_{i1}, \ldots, \rho_{ip_i}$. Therefore, the denominator of the first position is 1 because the individual picks the first item from K items and removes the picked item from the possible set of items and so on. This process is the same as the PL model. However, the PL model is the probability that a participant given these particular p_i items will put them in that order.

6.4.1 Selection Preference Model with Ranker-Specific Covariate

As for the PL model, we can extend the SP model by including covariates in the model. In particular, we introduce a ranker-specific covariate, x_i, where x_i is a dummy variable, to the SP model. The probability is

$$\begin{aligned} P(\rho_i; \boldsymbol{\pi}_\lambda, \boldsymbol{\pi}_\gamma) = & \frac{\pi_{\lambda,\rho_{i1}} + \pi_{\gamma,\rho_{i1}} x_i}{C_i} \times \frac{\pi_{\lambda,\rho_{i2}} + \pi_{\gamma,\rho_{i2}} x_i}{C_i - (\pi_{\rho_{i1}} + \pi_{\gamma,\rho_{i1}} x_i)} \times \\ & \frac{\pi_{\lambda,\rho_{i3}} + \pi_{\gamma,\rho_{i3}} x_i}{C_i - ((\pi_{\lambda,\rho_{i1}} + \pi_{\gamma,\rho_{i1}} x_i) + (\pi_{\lambda,\rho_{i2}} + \pi_{\gamma,\rho_{i2}} x_i))} \times \cdots \times \\ & \frac{\pi_{\lambda,\rho_{ip_i}} + \pi_{\gamma,\rho_{ip_i}} x_i}{C_i - \left((\pi_{\lambda,\rho_{i1}} + \pi_{\gamma,\rho_{i1}} x_i) + \cdots + (\pi_{\lambda,\rho_{ip_i-1}} + \pi_{\gamma,\rho_{ip_i-1}} x_i)\right)} \\ = & \frac{\pi_{\lambda,\rho_{i1}} + \pi_{\gamma,\rho_{i1}} x_i}{C_i} \prod_{j=2}^{p_i} \frac{\pi_{\lambda,\rho_{ij}} + \pi_{\gamma,\rho_{ij}} x_i}{C_i - \sum_{m=1}^{j-1} (\pi_{\lambda,\rho_{im}} + \pi_{\gamma,\rho_{im}} x_i)}, \end{aligned}$$

where $C_i = \sum_{k=1}^{K} (\pi_{\lambda,k} + \pi_{\gamma,k} x_i)$ and where we impose the constraint

$$\sum_{k=1}^{K} (\pi_{\lambda,k} + \pi_{\gamma,k}) = 1.$$

6.5 Application to Sundarbans Dataset

We use the Sundarbans dataset throughout this section in order to investigate the results from fitting various models. This dataset contains ties; therefore, we compare the results from the PL model, the PL model with random ties, and the PL model with Breslow approximation. We also evaluate the PL model and the SP model. The S and I from the PRM are used to compare the results from both models. After that the ROL model and the SP model with one covariate are fitted to the Sundarbans dataset to find the covariates that affect the villagers' problems. Models are fitted by using the `optim` function in R language. We chose Broyden–Fletcher–Goldfarb–Shanno (BFGS) as the optimization algorithm, which is the same as in Chapter 5.

6.5.1 Evaluation of the Breslow and Random Approaches for Tied Dataset

Three models are considered which are the PL model, the PL model with Breslow approximation, and the PL model with random method for ties. The

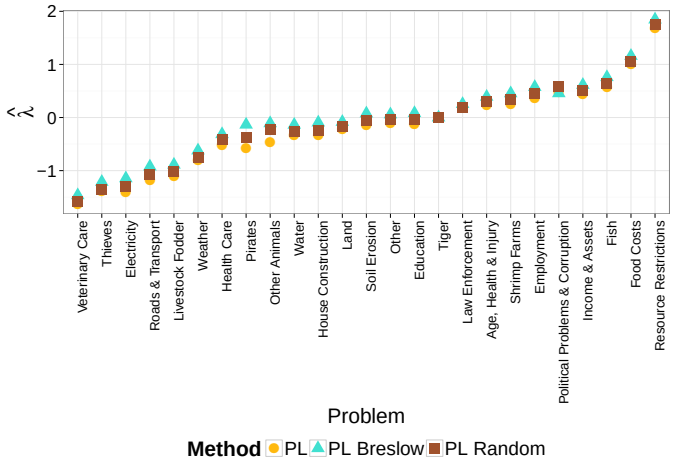


Figure 6.3: Parameter estimates for the PL model, the PL model with Breslow approximation, and the PL model with random when fitted to the Sundarbans dataset

PL model does not have any approximation for ties. It treats the ties as

if there are no ties by sorting the ties in ascending order according to the problem ID. The PL model with random method is fitted to the dataset 500 times and the average of the estimates is computed. There are 62 rankings that contain ties in the Sundarbans dataset. The estimated parameters from all three methods, where Tiger problem is taken as the reference problem, are shown in Figure 6.3. The $\hat{\boldsymbol{\lambda}}_{\text{PL-random}}$ are in between the $\hat{\boldsymbol{\lambda}}_{\text{PL-Breslow}}$ and the $\hat{\boldsymbol{\lambda}}_{\text{PL}}$. The $\hat{\boldsymbol{\lambda}}_{\text{PL-Breslow}}$ are higher than the others, except for one problem which is Political and Corruption problem.

Considering computational time, the PL model with Breslow approximation and with random method are fitted 50 times. The average computational times for the Breslow approximation and random method are 0.22 and 1.24 seconds. The Breslow approximation is faster than the random.

We use the PL model with the Breslow approximation in later experiments when there is no covariate because it can incorporate ties.

6.5.2 Evaluation of the PL Model and the SP Model

First, we compare results from the SP model with results from the PL model with the Breslow approximation as shown in Figure 6.4 where the Tiger problem is the reference problem. The figure shows that the SP model gives different results from the PL model. The Tiger problem receives the highest score from the SP model while the PL model indicates that Resource Restrictions problem is the highest. Moreover, there is no obvious pattern in this figure.

Second, we compare results from the PL model and the SP model with the severity (S) and incidence (I) scores calculated from PRM method. We plot $\hat{\boldsymbol{\pi}}_{\text{PL}}$ against S and against I in Figure 6.5. Figure 6.5a shows that $\hat{\boldsymbol{\pi}}_{\text{PL}}$ has a strong relationship with severity score. The $\hat{\boldsymbol{\pi}}_{\text{PL}}$ values capture the same trend as the severity scores do, however, it is not a linear relationship. Conversely, the plot comparing $\hat{\boldsymbol{\pi}}_{\text{PL}}$ and incidence does not show any pattern as presented

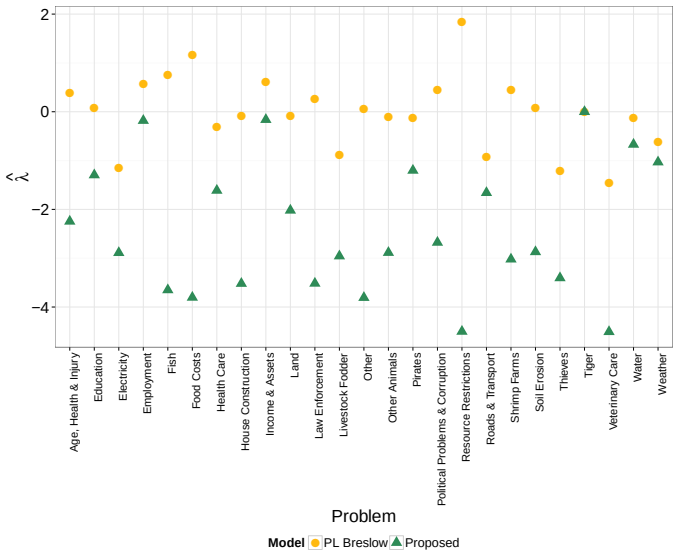
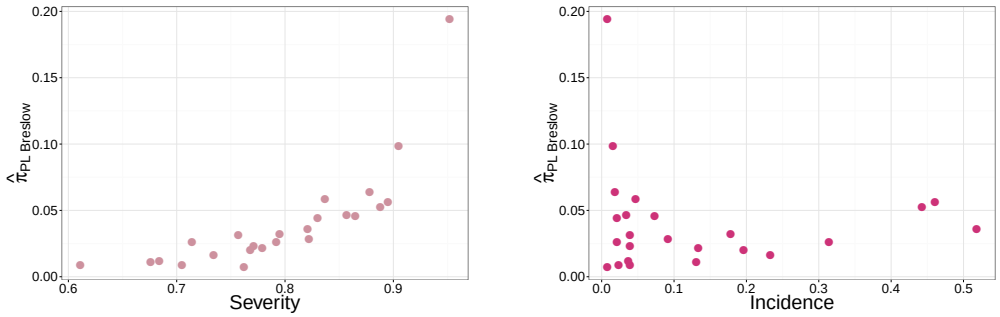


Figure 6.4: Parameter estimates for the PL model and the SP model when fitted to the Sundarbans dataset



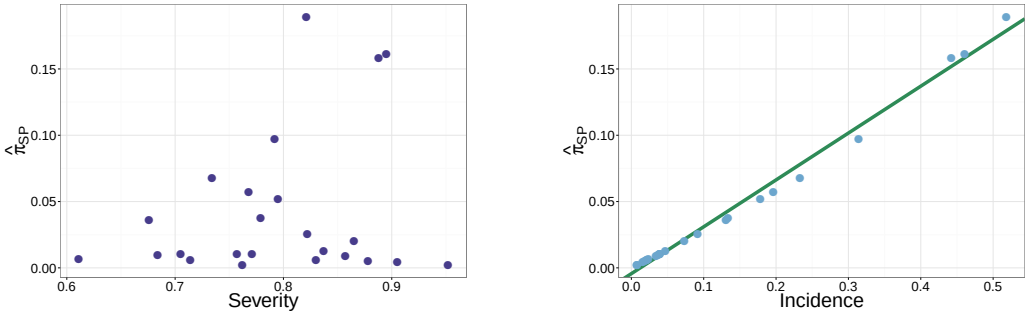
(a) $\hat{\pi}_{\text{PL-Breslow}}$ vs Severity (b) $\hat{\pi}_{\text{PL-Breslow}}$ vs Incidence

Figure 6.5: Parameter estimates from the PL model against the severity scores and incidences from the PRM

in Figure 6.5b.

We also plot the $\hat{\boldsymbol{\pi}}_{\text{Propose}}$ against severity and incidence scores. The plots give different conclusions from Figure 6.5. Figure 6.6a shows that there is no obvious relationship between the $\hat{\boldsymbol{\pi}}_{\text{SP}}$ and the severity. However, the SP model can capture the incidence as shown in Figure 6.6b. The $\hat{\boldsymbol{\pi}}_{\text{Propose}}$ have an approximately linear relationship with the incidence. The line in the figure is the simple linear regression line.

From the results above, we will use the rank-ordered logit (ROL) model to find covariates that affect severity scores.



(a) $\hat{\pi}_{\text{SP}}$ vs Severity (b) $\hat{\pi}_{\text{SP}}$ vs Incidence

Figure 6.6: Parameter estimates from the SP model against the severity scores and incidences from the PRM

6.5.3 Evaluation of the ROL Model

The previous section shows that the estimates from the PL model have a relationship with the severity scores from the PRM. The ROL model is therefore fitted to the Sundarbans dataset to find covariates that effect the severity scores.

The assumption that no item is always ranked first or last is violated when a ranker-specific covariate is introduced to the ROL model. Mainly the problems that cause this have incidence less than 2.5% (less than ten participants mentioned these problems). Eight problems have incidence less then 2.5% which are Fish, Resource Restrictions, Food Costs, Law Enforcement, Thieves, Veterinary Care, House Construction, and other. We group these problems together. Thus, there are 18 problems ($K = 18$) remaining after grouping.

However, the household type covariate, Age, Health, and Injury also violates the assumption. Therefore, we group this problem with Other and then 17 problems are considered for household type. We remove two records from the dataset because with the new grouping there is more than one problem classified as Other in the records, then 379 records remain in the dataset.

LR statistics when there is only one ranker-specific covariate in the ROL model are presented in Table 6.1. Household type is strongly significant.

Table 6.1: LR statistics for the ROL model with only one covariate when fitted to the Sundarbans dataset after grouped some problems

Covariate	G^2	df	p-value
Village Location	14.50	17	0.631
Education (years)	17.48	17	0.423
Gender	15.16	17	0.584
Age (years)	19.73	17	0.288
Household (3 categories)	77.78	32	<0.001
Interview Type	11.96	17	0.802

The model allowing for differences in values across the household types. Our objective is to examine the variations in the severity scores of different types where Tiger problem is a reference problem. The model with Household type is $\log(\mu_{\rho_{ij}}) = \lambda_{\rho_{ij}} + \gamma_{1,\rho_{ij}} x_{1,1} + \gamma_{2,\rho_{ij}} x_{2,1}$ where $x_1 = 1$ if Human Attack and $x_2 = 1$ if Livestock Attack, and 0 if No Conflict. The $\hat{\boldsymbol{\gamma}}_1$, $\hat{\boldsymbol{\gamma}}_2$, and their standard errors are shown in Table 6.2.

Table 6.2: Parameter estimates, according to γ_1, when there is only Household type in the ROL model for the Sundarbans dataset with 17 problems (SE in brackets)

Problem	λ	γ_1	γ_2
Political Problems & Corruption	-0.841(1.118)	1.864(1.313)	1.393(1.231)
Livestock Fodder	-1.618(1.082)	0.810(1.251)	0.913(1.406)
Education	-0.449(0.368)	0.406(0.504)	0.945(0.557)
Employment	0.380(0.328)	0.342(0.417)	-0.426(0.428)
Other	-0.065(0.354)	0.257(0.487)	0.384(0.553)
Health Care	-0.875(0.527)	0.121(0.660)	0.696(0.666)
Shrimp Farms	0.092(0.782)	0.084(1.106)	0.749(1.196)
Tiger	0	0	0
Roads & Transport	-0.943(0.454)	-0.101(0.660)	-0.523(0.628)
Income & Assets	0.859(0.267)	-0.231(0.387)	-1.329(0.419)
Electricity	-1.412(1.154)	-0.237(1.634)	-0.050(1.301)
Soil Erosion	0.685(0.801)	-0.241(1.037)	-1.927(1.068)
Weather	-0.639(0.417)	-0.250(0.520)	-0.275(0.536)
Other Animals	-0.011(0.795)	-0.421(0.963)	-1.585(1.348)
Pirates	-0.244(0.381)	-0.656(0.527)	-0.322(0.489)
Land	0.221(0.447)	-0.670(0.648)	-1.328(0.975)
Water	0.579(0.375)	-1.181(0.470)	-1.084(0.462)

The $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are differences when compared with the households that have no conflict with tigers. For example, Political Problems & Corruption has a difference of 1.864 for the human attack household. Then $\exp(1.864) = 6.44$, this means the odds of more severity of Political Problems & Corruption problem than Tiger problem is 6.44 times as great for human attack household as for no conflict household. We plot $\hat{\gamma}_1$ and $\hat{\gamma}_2$ for Human Attack and Livestock

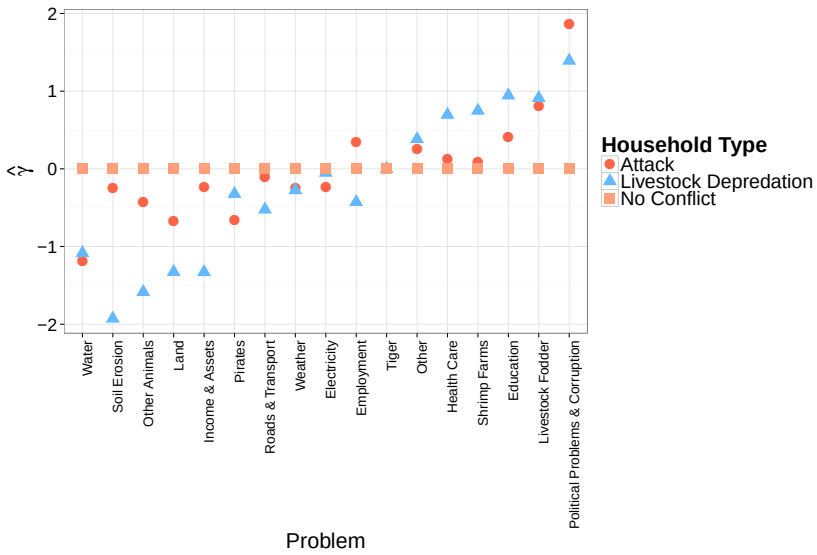


Figure 6.7: Parameter estimates when Household type covariate in the ROL model for the Sundarbans dataset with 17 problems

Attack household types are illustrated in Figure 6.7. The No Conflict is a reference household type; therefore, it is always zero. We can see the differences more easily from this figure. The severity scores for Weather and Electricity problems do not differ across the household types.

6.5.4 Evaluation of the SP Model with a Ranker-Specific Covariate

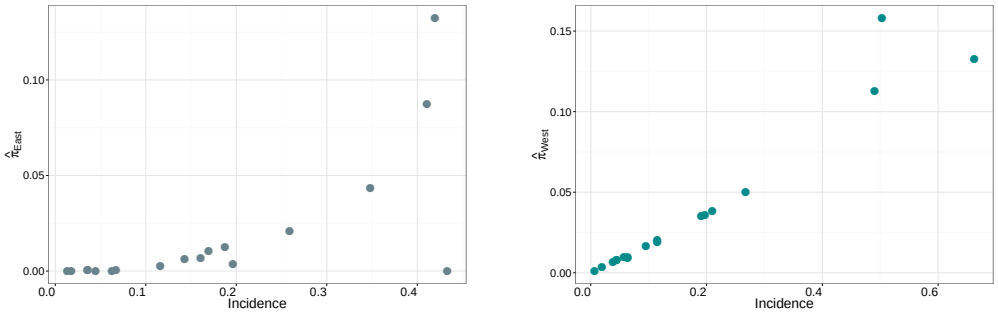
The SP model can present the incidence from the PRM method. We include a covariate in the SP model to identify the covariates that affect the incidences. We apply the SP model with one ranker-specific covariate to the Sundarbans dataset with 18 problems. LR statistics when there is only one covariate

included in the SP model are presented in Table 6.3. Most of the covariates are strongly significant at 1% significance level except Age. Age as a continuous covariate is not significant.

Table 6.3: LR statistics for the SP model when only one covariate in the model

Covariate	G^2	df	p-value
Village Location	63.02	17	<0.001
Education (years)	37.50	17	0.002
Gender	103.08	17	<0.001
Age (years)	13.96	17	0.670
Household (3 categories)	103.08	34	<0.001
Interview Type	45.86	17	<0.001

Village Location is strongly significant in the SP model with $x_i = 1$ if East and 0 if West. The log-likelihood is -2814.30. We plot $\hat{\boldsymbol{\pi}}_{\text{East}}$ and $\hat{\boldsymbol{\pi}}_{\text{West}}$ against incidence in east and west villages as shown in Figure 6.8, respectively. Figure



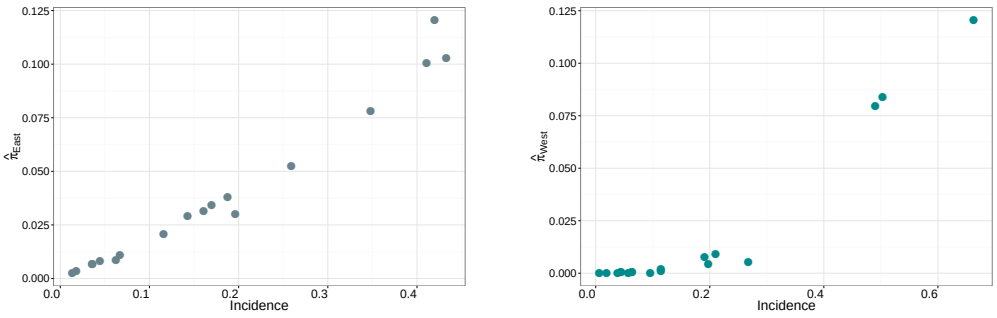
(a) East Villages (b) West Villages

Figure 6.8: Parameter estimates for the SP model with Village Location covariate where Village Location = 1 if East and and 0 if West against the incidences from the PRM

6.8a shows that results from the SP model do not have a linear relationship with I_{East} in the east villages. While in Figure 6.8b, we can observe that the estimates have approximate linear relationship with I_{West}.

However, when we swap the reference group then $x_i = 1$ if West and 0 if East, the log-likelihood is -2812.84. Ideally the reference group should not effect the model but it does in the SP model with a ranker-specific covariate.

The SP model is not symmetric. Again, we plot $\hat{\boldsymbol{\pi}}_{\text{East}}$ and $\hat{\boldsymbol{\pi}}_{\text{West}}$ against incidence in east and west villages as shown in Figure 6.9, respectively. Results



(a) East Villages (b) West Villages

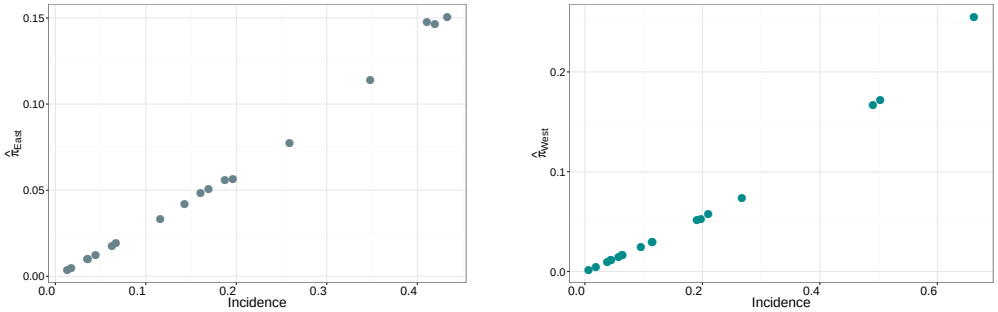
Figure 6.9: Parameter estimates for the SP model with Village Location covariate where Village Location = 1 if West and and 0 if East against the incidences from the PRM

are reversed compared to the previous figures. Figure 6.9a shows that $\hat{\boldsymbol{\pi}}_{\mathbf{East}}$ from the SP model has an approximate linear relationship with I_{East} in the east villages. While, we can observe that the estimates from the west villages do not have linear relationship with I_{West}.

We further investigate this case by splitting the original dataset into two groups which are east villages and west villages, respectively. The SP model without a covariate is applied to each group. The log-likelihood are -1864.68 and -1114.96 for east and west villages, respectively. The sum of log-likelihood is -2799.64 which is not equal to the log-likelihood from the SP model with a ranker-specific covariate. We plot the results against I in Figure 6.10. Figure 6.10 shows that the results from the SP method without covariate have an approximately linear relationship with the incidences in east and west villages, respectively.

6.6 Conclusions

In this chapter, we have explored another type of partial ranking data where individuals are allowed to mention their own choices. Therefore, the number



(a) East Villages (b) West Villages

Figure 6.10: Parameter estimates for the SP model against the incidences from the PRM

of choices varies between individuals. We suggest that the numbers of choices can be modelled using the Poisson-Binomial distribution. This is because the Poisson-Binomial distribution allows not-all-equal probabilities of successes. We attempt to estimate the number of choices by using probabilities from the logistic regression. After that we use these probabilities in order to generate the numbers of choices. Our result shows that this process can approximate the numbers of choices.

The Sundarbans dataset includes tied ranking. We consider two approximations which are the Breslow approximation and random tie breaking for the PL model. The results show that the estimates from the PL model with the Breslow approximation are higher than for random tie breaking except for one problem. We follow the idea of severity and incidence scores from the PRM. We attempt to find models that are able to capture the same trend as the severity and incidence scores. The estimates from the PL model with the Breslow approximation are compared with the severity and incidence scores calculated from the PRM method. The plots of the estimates from the PL model against severity and incidence scores reveal that the estimates from the PL model can capture the trend as the severity scores; however, there is no obvious relationship with the incidence scores.

The SP model is applied to the Sundarbans dataset then the estimates are

plotted against the severity and incidence scores. This reverses the previous result. The estimates from the SP model have an approximately linear relationship with the incidence scores while there is no pattern with the severity scores. We conclude that the PL model and the SP model can capture the same trend as the severity and incidence scores, respectively.

Since the PL model can capture the severity scores, the ROL model is considered in order to find covariates that affect the severity scores. The result shows that only the Household Type covariate is significant.

We extend the SP model to incorporate a ranker-specific covariate in order to find covariates that affect the incidence scores. However, analysis shows that the SP model is not symmetric.

Chapter 7

Discussion

The work presented in this thesis provides a better understanding of modelling partial ranking data and extending the algorithms for estimation and inference. In this chapter, we summarize contributions and possible future work.

7.1 Contributions

In the analysis of partial ranking data, the Bradley-Terry (BT) and the Plackett-Luce (PL) models are considered in Chapter 3. The existing package in R for fitting the PL model suffers from slow computational time. We provide two new R algorithms. One, `PLem`, is translated from Matlab code of Caron and Doucet (2012). We implemented `PLmm` algorithm. Both algorithms perform faster than the existing package. We also implemented the `PLinfm` algorithm to calculate the observed information matrix.

We apply rank-breaking methods, which were introduced by Soufiani and Parkes (2014), to partial rankings to break them into pairwise comparisons. The BT model with different weightings is fitted to the paired data. We consider BTw weighting which is suggested by Khetan and Oh (2016). We proposed the BTw-Sqrt weighting and it performs better than the BTw weighting when number of ranker is less than 2500. In other words, the BTw-Sqrt gives

better performance when the number of rankers is small. In real world applications, it is unusual to have data that has more than 2500 rankers. Thus, the BTw-Sqrt weighting is more practical. The rank-breaking method may useful for researcher who is not similar in statistics area since the BT model is easier to understand than the PL model.

Chapter 4, we try to find better selection methods than random selection in order to choose informative subsets. Our proposed methods perform slightly better than the existing methods, which are the D-optimality, E-optimality, Wald, and random, in terms of both the Kendall tau correlation and MSE. The Wald criterion does not perform well in our experiment. In paired comparison data, which is studied by Aftab et al. (2011) and Pfeiffer et al. (2012), their methods improve the estimates. However, the Wald criterion, which is based on the idea of the t-test, does not perform well. One possible reason is that there are 45 possible pairs when $p = 10$ and the average effect of 45 pairs is being used instead of a single pair. The proposed methods are effective to enable the PL to be fitted to data from fewer rankers than random selection. We recommend to use one of our proposed methods at the beginning of surveys. This is because the PL model converges with fewer rankers when using subsets from the proposed methods.

In Chapter 5, we introduce two extensions of the PL model. In order to include covariates in the PL model, the rank-ordered logit (ROL) model is introduced. Our main contribution here is to extend the MM algorithm of Hunter (2004) to the ROL model and implement `ROLmm` algorithm. The `ROLmm` requires less computational time when compared with the `optim` function.

Another extension is the Benter model. We follow the work of Gormley and Murphy (2008) and then implement the `BMmm` algorithm. The `BMmm` algorithm performs slower than the `optim` function. The ROL model and the Benter model are applied to the Animal dataset. The LR statistics show that both

models fit better than the PL model.

We explore the rank-breaking idea from Chapter 3 a little further by including a ranker-item-specific covariate, Familiarity, to the ROL model for paired comparison data with the BTw and BTw-Sqrt weightings. The BTw-Sqrt weighting performs better than the BTw weighing when compared with the ROL model.

We propose a new model by combining the ROL model and the Benter model in order to look for a deeper understanding of human choice preference. The combined model is also fitted to the Animal dataset. The LR statistics show that the combined model improves the fits when compared with the ROL model. The bootstrap goodness-of-fit tests with the Kendall tau distance and IOS statistics show that the combined model with Familiarity is an appropriate model to fit the Group I data from the Animal dataset at 1% significance level.

We explore another type of ranking data in Chapter 6. In general, a technique known as Participatory Risk Mapping (PRM) is used to analyze this kind of data. We would like to find appropriate models that can be used. We showed that the results from the PL model are closely related to the severity scores from the PRM. Furthermore, the ROL model incorporates covariates but the PRM method cannot include any covariates. Another result from the PRM method is incidences. We proposed a model, the Selection Preference (SP) model, to capture incidences. The estimates from the SP model have almost a linear relationship with incidence scores. We extend the SP model by including a ranker-specific covariate. However, the SP model with a ranker-specific covariate is not symmetric.

7.2 Future Work

In Chapter 2, Miller (1955) suggested that number of objects to be ranked should be no more than seven. It would be nice if we can examine the effects

of the different number of objects to be ranked on rankings. In this thesis, we assume that rankings are ranked from best to worst; however, there are other kinds of ranking behavior such as worst to best, best and worst etc. Suggested future work is to find the way to test how the participants ranked the subsets.

In Chapter 3, the BT model with different weightings is fitted to the pairwise comparison data from the full rank-breaking method. A ranker-item-specific covariate is included in this model as a pre-investigation in Chapter 5. The BT model with different weightings should explore further in this issue. It is possible to get good estimates and reduce computational time when compared with the ROL model.

In Chapter 4, the proposed methods mean that the PL model can be fitted to the smaller data than random selection; however, they have a limitation that they can perform only when $p^2 = K$. We can adapt the proposed methods in order to make them work even when $p^2 \neq K$.

In Chapter 5, we propose the `ROLmm` algorithm to fit the ROL model; however, the `ROLmm` algorithm cannot fit a continuous ranker-specific covariate. It should be possible to extend this algorithm to incorporate a continuous ranker-specific covariate. Moreover, it is interesting to explore interactions between covariates. For example, the interaction between the Familiarity and the Start Position. Another suggestion, the item-specific and ranker-item-specific covariates can may combine in the same update since these kids of covariates use the Newton-Raphson method to estimate parameters. This may reduce computational time.

Guiver and Snelson (2009) studied the PL model under the Bayesian framework. Suggested future work involves extending a Bayesian approach to the extensions of the PL model.

The combined model in Chapter 5 cannot incorporate a continuous ranker-specific covariate. When we fitted the combined model to the Group I data

from the Animal dataset with Age as continuous covariate, $\hat{\boldsymbol{\alpha}}$ was close to zero. The combined model needs to be explored further in order to include a continuous ranker-specific covariate to the model.

We use the `optim` function to cross-check with the results from our algorithms. Our algorithms, the `BMmm` and the `CMmm`, usually work well; however, they fail occasionally. This issue needs more investigation.

Another suggested future work is to develop diagnostic techniques to improve the ranking models, especially for the ROL model and the combined model since they include covariates in the models.

The LR test is not a proper test for testing the Benter model because the dampening parameter in the model is in the boundary of a parameter space. This issue has to be explored more in order to find a suitable test for comparing the PL model with the Benter model and the ROL model with the combined model.

The rank-breaking method with covariates needs further investigation. The BT model with weighting can be a good option for analyzing the partial ranking data with covariates. It is interesting because the BT model is less complicated than the ROL model.

The work in Chapter 6 explores another type of ranking data. The SP model can capture incidences from the PRM method. Our suggestion is to find a way to introduce covariates into the SP model and the model should ideally be a symmetric model.

Bibliography

Abdi, H. (2007). The Kendall Rank Correlation Coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508 – 510.

Aftab, H., Raj, N., Cuff, P., and Kulkarni, S. (2011). Mutual Information Scheduling for Ranking. In *Proceedings of the* 14^{th} *International Conference on Information Fusion*, pages 1 – 8.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons.

Ahn, J., Lee, J., Lee, J.-D., and Kim, T.-Y. (2006). An Analysis of Consumer Preferences among Wireless LAN and Mobile Internet Services. *Journal ETRI*, 28(2):205 – 215.

Ailon, N. (2010). Aggregation of Partial Rankings, p-Ratings and Top-m Lists. *Algorithmica*, 57(2):284 – 300.

Allison, P. D. and Christakis, N. A. (1994). Logit Models for Sets of Ranked Items. *Sociological Methodology*, 24:199 – 228.

Alvo, M. and Yu, P. L. (2014). *Statistical Methods for Ranking Data*. Springer.

Arrow, K. (1951). *Social Choice and Individual Values*. New York: Wiley.

Atkinson, A. C., Donev, A. N., and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*. Oxford University Press.

Babington-Smith, B. (1950). Discussion of Professor Ross's Paper. *Journal of the Royal Statistical Society*, 12:53 – 56.

Baker, R. D. and McHale, I. G. (2015). Deterministic Evolution of Strength in Multiple Comparisons Models: Who is the Greatest Golfer? *Scandinavian Journal of Statistics*, 42(1):180 – 196.

Barrett, C. B., Smith, K., and Box, P. W. (2001). Not Necessarily in the Same Boat: Heterogeneous Risk Assessment among East African Pastoralists. *The Journal of Development Studies*, 37(5):1 – 30.

Beggs, S., Cardell, S., and Hausman, J. (1981). Assessing the Potential Demand for Electric Cars. *Journal of Econometrics*, 16(1):1 – 19.

Benter, W. (1994). Computer-Based Horse Race Handicapping and Wagering Systems: A Report. *Efficiency of Racetrack Betting Markets*, pages 183 – 198.

Bonilla, E., Guo, S., and Sanner, S. (2010). Gaussian Process and Preference Elicitation. In *Proceedings of Advances in Neural Information Processing Systems 23*, pages 262 – 270.

Borda, J. C. (1781). Memoire Sur Les Elections Au Scrutin. *Histoire de l'Academie Royale des Sciences*.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambrige University Press.

Bradley, R. A. and Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons. *Biometrika*, 39(3/4):324 – 345.

Breslow, N. and Crowley, J. (1974). A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *The Annals of Statistics*, 2(3):437 – 453.

Capanu, M. and Presnell, B. (2008). Misspecification Test for Binomial and Beta-Binomial Models. *Statistics in Medicine*, 27:2536 – 2554.

Caplin, A. and Nalebuff, B. (1991). Aggregation and Social Choice: A Mean Voter Theorem. *Econometrica*, 59(1):1 – 23.

Caron, F. and Doucet, A. (2012). Efficient Bayesian Inference for Generalized Bradley-Terry Models. *Journal of Computational and Graphical Statistics*, 21(1):174 – 196.

Cattelan, M., Varin, C., and Firth, D. (2013). Dynamic Bradley-Terry Modelling of Sports Tournaments. *Journal of the Royal Statistical Society Series C - Applied Statistics*, 62(1):135 – 150.

Chapman, R. G. and Staelin, R. (1982). Exploiting Rank Ordered Choice Set Data within the Stochastic Utility Model. *Journal of Marketing Research*, 19(3):288 – 301.

Chen, W. and Soufiani, H. A. (2013). Package 'StatRank'. http://CRAN.R-project.org/package=StatRank.

Cheng, W., Dembczynski, K., and Hullermerier, E. (2010). Label Ranking Methods Based on the Plackett-Luce Model. In *Proceedings of the* 27^{th} *International Conference on Machine Learning*, pages 215 – 222, Haifa, Israel.

Chirowodza, A., van Rooyen, H., Joseph, P., Sikotoyi, S., Richter, L., and Coates, T. (2009). Using Participatory Methods and Geographic Information Systems (GIS) to Prepare for an HIV Community-Based Trial in Vulindlela, South Africa (Project Accept-HPTN 043). *Journal of Community Psychology*, 37(1):41 – 57.

Cohen, W. W., Schapire, R. E., and Singer, Y. (1999). Learning to Order Things. *Journal of Artificial Intelligence Research*, 10:243 – 270.

Condorcet, M. D. (1785). *Essai Sur L'application de L'analyse a la Probability des Decisions Rendues a la Pluralite des Voix*. Imprimerie Royal, Paris.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 34(2):187 – 220.

Cramer, J. S. (2003). *Logit Models From Economics and Other Fields*. Cambridge University Press.

Critchlow, D. E., Fligner, M. A., and Verducci, J. S. (1991). Proability Models on Rankings. *Journal of Mathematical Psychology*, 35(3):294 – 318.

Croissant, Y. (2013). Estimation of Multinomial Logit Models in R: The mlogit Packages. http://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf. Accessed: 2015-02-23.

Croux, C. and Dehon, C. (2010). Influence functions of the spearman and kendall correlation measures. *Statistical Methods and Applications*, 19:497 – 515.

Davidson, R. R. and Farquhar, P. H. (1976). A Bibliography on the Method of Paired Comparisons. *Biometrics*, 32(2):241 – 252.

Diaconis, P. (1988). Group Representations in Probability and Statistics. *Institute of Mathematical Statistics Lecture Notes*, 11. Institute of Mathematical Statistics, Hayward, CA.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank Aggregation Methods for the Web. In *Proceedings of the* 10^{th} *International Conference on World Wide Web*, pages 613 – 622.

Ebbinghaus, H.-D. (2008). *Ernst Zermelo: An Approach to His Life and Work*. Springer.

Firth, D. (2008). BradleyTerry: Bradley-Terry Models. http://CRAN.R-project.org/pack=BradleyTerry.

Fligner, M. A. and Verducci, J. S. (1986). Distance Based Ranking Models. *Journal of the Royal Statistical Society*, 48(3):359 – 369.

Fligner, M. A. and Verducci, J. S. (1988). Multistage Ranking Models. *Journal of the American Statistical Association*, 83:892 – 901.

Ford, L. R. (1957). Solution of a Ranking Problem from Binary Comparisons. *The American Mathematical Monthly*, 64(8):28 – 33.

Fuller, D. O., Troyo, A., Alimi, T. O., and Beier, J. C. (2014). Participatory Risk Mapping of Malaria Vector Exposure in Northern South America using Environmental and Population Data. *Applied Geography*, 48:1 – 7.

Gormley, I. C. and Murphy, T. B. (2008). Exploring Voting Blocs Within the Irish Electorate: A Mixture Modeling Approach. *Journal of the American Statistical Association*, 103(483):1014 – 1027.

Guiver, J. and Snelson, E. (2009). Bayesian Inference for Plackett-Luce Ranking Models. In *Proceedings of the* 26^{th} *the Annual International Conference on Machine Learning*, pages 377 – 384, ACM, New York.

Hastie, T. and Tibshirani, R. (1998). Classification by Pairwise Coupling. *The Annals of Statistics*, 26(2):451 – 471.

Hausman, J. A. and Ruud, P. A. (1987). Specifying and Testing Econometric Models for Rank-Ordered Data. *Journal of Econometrics*, 34:83 – 104.

Henery, R. J. (1981). Permutation Probabilities as Models for Horse Races. *Journal of the Royal Statistical Society*, 43:86 – 91.

Henery, R. J. (1983). Permutation Probabilities for Gamma Random Variables. *Applied Probability*, 20:822 – 834.

Hino, H., Fujimoto, Y., and Murata, N. (2010). A Grouped Ranking Model for Item Preference Parameter. *Neural Computation*, 22:2417 – 2451.

Hunter, D. R. (2004). MM Algorithms for Generalized Bradley-Terry Models. *The Annals of Statistics*, 32(1):384 – 406.

Inskip, C., Ridout, M., Fahad, Z., Tully, R., Barlow, A., Barlow, C. G., Islam, A., Roberts, T., and MacMillan, D. (2013). Human-Tiger Conflict in Context: Risks to Lives and Livelihoods in the Bangladesh Sundarbans. *Human Ecology*, 41(2):169 – 186.

Jing, L., Liu, X., and Gang, L. (2013). Public Participatory Risk Mapping for Community-Based Urban Disaster Mitigation. *Applied Mechanics and Materials*, 380 – 384:4609 – 4613.

Kamakura, W. A. and Mazzon, J. A. (1991). Value Segmentation: A Model for the Measurement of Values and Value Systems. *Journal of Consumer Research*, 18(2):208 – 218.

Kamishima, T. (2003). Nantonac Collaborative Filtering: Recommendation Based on Order Responses. In *Proceedings of the* 9^{th} *International Conference on Knowledge Discovery and Data Mining*, pages 583 – 588, Washington, DC, USA.

Katahira, H. (1990). Perceptual Mapping Using Ordered Logit Analysis. *Marketing Science*, 9(1):1 – 17.

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81 – 93.

Kendall, M. G. (1970). *Rank Correlation Methods*. Griffin.

Kendall, M. G. and Gibbons, J. D. (1990). *Rank Correlation Methods*. Griffin, London, 5 edition.

Khamis, H. (2008). Measure of Association: How to Choose? *Journal of Diagnostic Medical Sonography*, 24(3):155 – 162.

Khetan, A. and Oh, S. (2016). Data-Driven Rank Breaking for Efficient Rank Aggregation. *Journal of Machine Learning Research*, 17:1 – 45.

Koehler, J. K. and Ridpath, H. (1982). An Application of a Biased Version of the Bradley-Terry-Luce Model to Professional Basketball Results. *Journal of Mathematical Psychology*, 25(3):187 – 205.

Koop, G. and Poirier, D. J. (1994). Rank-Ordered Logit Models: An Empirical Analysis of Ontario Voter Preferences. *Journal of Applied Econometrics*, 9(4):369 – 388.

Kumar, S. and Kant, S. (2007). Exploded Logit Modeling of Stakeholders' Preferences for Multiple Forest Values. *Forest Policy and Economics*, 9(5):516 – 526.

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational Graphical Statistics*, 9(1):1 – 20.

Lareau, T. J. and Rae, D. A. (1989). Valuing WTP for Diesel Odor Reductions: An Application of Contingent Ranking Technique. *Southern Economic Journal*, 55:728 – 742.

Lee, P. H. and Yu, P. L. (2010). Distance-Based Tree Models for Ranking Data. *Computational Statistics and Data Analysis*, 54(6):1672 – 1682.

Lee, P. H. and Yu, P. L. (2013). An R Package for Analyzing and Modeling Rank Data. *BMC Medical Research Methodology*, 13:1 – 11.

Lu, T. and Boutilier, C. (2011). Learning Mallows Models with Pairwise Preferences. In *Proceedings of the* 28^{th} *Internaional Conference on Machine Learning*, pages 145 – 152, Bellevue, Washington, USA.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Luce, R. D. and Suppes, P. (1965). Preference, Utility, and Subjective Probability. *Handbook of Mathematical Psychology*, 3:249 – 410.

Mallows, C. L. (1957). Non-null Ranking Models: I. *Biometrika*, 44(1/2):114 – 130.

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*. Chapman & Hall.

Marley, A. A. J. (1968). Some Probabilistic Models of Simple Choice and Ranking. *Journal of Mathematical Psychology*, 5(2):311 – 332.

Marley, A. A. J. and Louviere, J. J. (2005). Some Probabilistic Models of Best, Worst, and Best-Worst Choices. *Journal of Mathematical Psychology*, 49(6):464 – 480.

Maydeu Olivares, A. and Bockenholt, U. (2005). Structural Equation Modeling of Paired-Comparision and Ranking Data. *Psychological Methods*, 10(3):285 – 304.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. *Frontiers in Econometrics*, pages 105 – 142.

McFadden, D. (1976). Quantal Choice Analysis: A Survey. *Annals of Economic and Social Measurement*, 5(4):363 – 390.

McLean, I., Urken, A. B., and Hewitt, F. (1995). *Classics of Social Choice*. University of Michigan Press.

Miller, G. A. (1955). The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information. *Psychological Review*, 101(2):343 – 352.

Moore, L. (1990). Segmentation of Store Choice Models Using Stated Preferences. *Papers of the Regional Science Association*, 69(1):121 – 131.

Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008). P-Values are Random Variables. *The American Statistician*, 62:242 – 245.

Peng, K., Nisbett, R. E., and Wong, N. Y. C. (1997). Validity Problems Comparing Values Across Cultures and Possible Solutions. *Psychological Methods*, 2(4):329 – 344.

Pfeiffer, T., Gao, X. A., Mao, A., Chen, Y., and Rand, D. G. (2012). Adaptive Polling for Information Aggregation. In *Proceedings of the* 26^{th} *AAAI Conference on Artificial Intelligence*, pages 122 – 128.

Plackett, R. L. (1975). The Analysis of Permutations. *Journal of the Royal Statistics Society. Series C (Applied Statistics)*, 24(2):193 – 202.

Presnell, B. and Boos, D. D. (2004). The IOS Test for Model Misspecification. *Journal of the American Statistical Association*, 99:216 – 227.

Punj, G. N. and Staelin, R. (1978). The Choice Process for Graduate Business Schools. *Journal of Marketing Research*, 15(4):588 – 598.

Rao, P. V. and Kupper, L. L. (1967). Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model. *Journal of the American Statistical Association*, 62(317):194 – 204.

Savage, I. R. (1956). Contributions to the Theory of Rank Order Statistics: The Two-Sample Case. *The Annals of Mathematical Statistics*, 27(3):590 – 615.

Savage, I. R. (1957). Contributions to the Theory of Rank Order Statistics: The "Trend" Case. *The Annals of Mathematical Statistics*, 28(4):968 – 977.

Sinsheimer, J. S., Blangero, J., and Lange, K. (2000). Gamete-Competition Models. *The Annals of Human Genetics*, 66:1168 – 1172.

Smith, K., Barrett, C. B., and Box, P. W. (2000). Participatory Risk Mapping for Targeting Research and Assistance: With an Example from East African Pastoralists. *World Development*, 28(11):1945 – 1959.

Soufiani, H. A., Chen, W., Parkes, D. C., and Xia, L. (2013a). Generalized Method-of-Moments for Rank Aggregation. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, volume 26, pages 2706 – 2714, Lake Tahoe, NV, USA.

Soufiani, H. A. and Parkes, D. C. (2014). Computing Parametric Ranking Models via Rank-Breaking. In W&CP, editor, *Proceedings of the* 31^{st} *International Conference on Machine Learning*, volume 32, Beijing, China.

Soufiani, H. A., Parkes, D. C., and Xia, L. (2013b). Preference Elicitation for General Random Utility Models. In *Proceedings of the* 29^{th} *Conference Uncertainity Artificial Intelligence (UAI)*, pages 596 – 605, Bellevue, Washington, USA.

Stern, H. (1990). Models for Distributions on Permutations. *Journal of the American Statistical Association*, 85:558 – 564.

Thurstone, L. L. (1927). A Law of Comparative Judgment. *Psychological Review*, 34(4):273 – 286.

Thurstone, L. L. (1931). Rank Order as a Psychological Method. *Journal of Experimental Psychology*, 14:187 – 201.

Train, K. (2003). *Discrete Choice Models with Simulation*. Cambridge University Press.

Tran, T., Phung, D., and Venkatesh, S. (2014). Learning Rank Functionals: An Empirical Study. *ArXiv e-prints*.

Tran, T., Phung, D., and Venkatesh, S. (2016). Modelling Human Preferences for Ranking and Collaborative Filtering: A Probabilistic Ordered Partition Approach. *Knowledge and Information Systems*, 47(1):157 – 188.

Tsukida, K. and Gupta, M. R. (2011). How to Analyze Paired Comparison Data. Technical report, University of Washington.

Turner, H. and Firth, D. (2012). Bradley-Terry Models in R: The BradleyTerry2 Package. *Journal of Statistical Software*, 48(9):1 – 21.

Tutz, G. (1986). Bradley-Terry-Luce Models with an Ordered Response. *Journal of Mathematical Psychology*, 30:306 – 316.

Tversky, A. (1972). Elimination by Aspects: A Theory of Choice. *Psychological Review*, 79:281 – 299.

UNESCO (2016). The sundarbans. http://en.wikipedia.org/wiki/Psychology. Retrieved October 20, 2016.

von Davier, M. (1997). Bootstrapping Goodness-of-Fit Statistics for Sparse Categorical Data - Results of a Monte Carlo Study -. *Methods of Psychological Research Online*, 2(2):29 – 48.

Yellott, J. (1977). The Relationship between Luce's Choice Axiom, Thurstone's Theory of Comprative Judgment, and the Double Exponential Distribution. *Journal of Mathematical Psychology*, 15(2):109 – 144.

Yu, P. L. (2000). Bayesian Analysis of Ordered-Statistics Models for Ranking Data. *Psychometrika*, 65(3):281 – 299.

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitshrift*, 29:436 – 460.