Parameter Estimation for Generalized Thurstone Choice Models

Milan Vojnovic MILANV@MICROSOFT.COM
Microsoft Research, Cambridge, UK

Se-Young Yun YUNSEYOUNG@GMAIL.COM
Microsoft Research, Cambridge, UK

Abstract

We consider the maximum likelihood parameter estimation problem for a generalized Thurstone choice model, where choices are top-1 items from comparison sets of two or more items. We provide tight characterizations of the mean square error, as well as necessary and sufficient conditions for correct classification when each item belongs to one of two classes. These results provide insights into how the estimation accuracy depends on the choice of a generalized Thurstone choice model and the structure of comparison sets. We find that for a priori unbiased structures of comparisons, e.g., when comparison sets are drawn independently and uniformly at random, the number of observations needed to achieve a prescribed estimation accuracy depends on the choice of a generalized Thurstone choice model. For a broad set of generalized Thurstone choice models, which includes all popular instances used in practice, the estimation error is shown to be largely insensitive to the cardinality of comparison sets. On the other hand, we found that there exist generalized Thurstone choice models for which the estimation error decreases much faster with the cardinality of comparison sets.

1. Introduction

We consider the problem of estimating the strengths of items based on observed choices of items, where each choice is from a subset of two or more items. This accommodates pair comparisons as a special case, where each comparison set consists of two items. In general, the outcome of each comparison is a top-1 list that singles out one item from given set of compared items. There are many applications in practice that are accommodated by this framework, e.g., single-winner contests in crowdsourcing services such as TopCoder or Taskcn, or hiring decisions where one applicant gets hired among those who applied for a job, e.g., in online labour marketplaces such as Fiverr and Upwork, as well as numerous sports competitions and online gaming platforms.

In particular, we consider the choices according to a generalized Thurstone choice model. This model accommodates several well known models, e.g. Luce's choice model, and Bradley-Terry model for pair comparisons; see discussion of related work in Section 1.1. A generalized Thurstone choice model is defined by a cumulative distribution function F and a parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbf{R}^n$, where θ_i represents the strength of item i. For every given non-empty subset of items S, the choice is assumed to be an item in S that exhibits the best performance, where the performance of each item $i \in S$ is defined as the sum of the strength parameter θ_i and an independent sample from the cumulative distribution function F. Many well known models of choice are special instances of generalized Thurstone choice models for specific choices of F; see a catalogue of examples in Section 2.3.

In this paper, our goal is to characterize the accuracy of a parameter estimator of a top-1 list generalized Thurstone choice model. In particular, we want to understand how is the estimation accuracy affected by the choice of a generalized Thurstone model, and the structure of the comparison sets. Our results show that the choice of a generalized Thurstone model can have a substantial effect on the parameter estimation accuracy.

More specifically, our main contributions in this paper can be summarized as follows.

We provide tight lower and upper bounds for the mean square error of the maximum likelihood parameter estimator (Section 3). These results provide necessary and sufficient conditions for the estimation of the parameter within a prescribed accuracy. Moreover, they reveal how the choice

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

of a generalized Thurstone choice model and the structure of comparison sets affect the estimation accuracy. In particular, we find that a key parameter is an eigenvalue gap of a pair-weight matrix. This pair-weight matrix is defined such that each element of this matrix that corresponds to a pair of items is equal to a weighted sum of the number of co-participations of the given pair of items in comparison sets of different cardinalities. The weight associated with a comparison set is a decreasing function of the cardinality of the comparison set, which depends on the choice of the generalized Thurstone choice model.

As a corollary, we derive tight characterizations of the mean square error for the case when all comparison sets are of equal cardinalities and the comparison sets are unbiased, e.g., each comparison set is sampled independently, uniformly at random without replacement from the set of all items. Such comparison sets are in spirit of tournament schedules like round-robin schedules that are common in various sports competitions. We also consider the parameter estimation problem for a generalized Thurstone choice model where each item is either of a high or a low class (Section 4). We establish necessary and sufficient conditions for correct classification of all items, when comparison sets have equal cardinalities and are drawn independently, uniformly at random without replacement from the set of all items. These conditions are shown to match those derived from the bounds for the mean square error up to constant factors.

These results provide a clear picture about the effect of a choice of a generalized Thurstone choice model and the cardinality of comparison sets. Perhaps surprisingly, we find that for a large set of special instances of generalized Thurstone choice models, which includes all popular cases used in practice, the mean square error decreases with the cardinality of comparison sets, but rather weakly. In particular, the mean square error is shown to be largely insensitive to the cardinality of comparison sets of three or more items. On the other hand, we exhibit instances of generalized Thurstone choice models for which the mean square error decreases much faster with the cardinality of comparison sets; in particular, decreasing inversely proportionally to the square of the cardinality (Section 5).

1.1. Related Work

The original Thurstone choice model was proposed by (Thurstone, 1927) as a model of comparative judgement for pair comparisons. The key property of this model is that each item is assumed to be associated with a performance random variable defined as the sum of a strength parameter and a noise random variable. Specifically, in the original Thurstone model, the noise is assumed to be a Gaussian random variable. This amounts to the winning probability of one item against another item in a pair comparison that is a cumulative Gaussian distribution function of the difference of their corresponding strength parameters. Similar model but with winning probabilities according to a logistic cumulative distribution function was originally studied by (Zermelo, 1929), and following the work by (Bradley & Terry, 1952; 1954) is often referred to as the Bradley-Terry model. A generalization of this model to comparisons of two or more items was studied by (Luce, 1959) and is referred to as the Luce's choice model (Luce, 1959). Other models of choice have also been studied, e.g., Dawkins' choice model (Dawkins, 1969). Relationships between the Luce's choice model and generalized Thurstone choice models were studied in (Yellott, 1977). Some of these models underlie the design of popular rating systems, e.g., Elo rating system (Elo, 1978) that was originally designed and has been used for rating skills of chess players but also for various other sport competitions, and TrueSkill (Graepel et al., 2006) that is used by a popular online gaming platform. All these models are instances of a generalized Thurstone model, and are special instances of generalized linear models, see, e.g., (Nelder & Wedderburn, 1972), (McCullagh & Nelder, 1989), and Chapter 9 in (Murphy, 2012). See Chapter 9 (Vojnović, 2016) for an exposition to the principles of rating systems.

Several studies argued that different models of pair comparisons yield empirically equivalent performance, e.g. (Stern, 1992), suggesting that the choice of a generalized Thurstone model does not matter much in practice. Our results show that there can be a significant fundamental difference between generalized Thurstone choice models with respect to the parameter estimation accuracy.

More recent work has focused on characterizing the parameter estimation error and deriving efficient computational methods for parameter estimation for different models of pair comparisons, e.g., (Negahban et al., 2012) and (Rajkumar & Agarwal, 2014) for pair comparisons according to Bradley-Terry model, and (Hajek et al., 2014) for full ranking outcomes according to a generalized Thurstone model with double-exponential distribution of noise. Our work is different in that we consider top-1 list models and the parameter estimation error for generalized Thurstone choice models that allow for comparisons of two or more items and different distributions of individual performances.

2. Problem Formulation and Notation

2.1. Basic Definitions

We consider a rank aggregation problem with top-1 list model. We denote with $N = \{1, 2, \ldots, n\}$ the set of all items. The input data consists of a sequence of $m \geq 1$ observations $(S_1, y_1), (S_2, y_2), \ldots, (S_m, y_m)$, where for each

observation t, $S_t \subseteq N$ is a subset of items, and y_t is the single item observed to be chosen from S_t; we refer to S_t as a *comparison set* and to y_t as a *choice*.

For every $S \subseteq N$ and $i \in S$, we denote with $w_{i,S}$ the number of observations such that the comparison set is S and the chosen item is i. In particular, for pair comparisons, we denote with $w_{i,j}$ the number of observations such that the comparison set is $\{i, j\}$ and the chosen item is i.

2.2. Generalized Thurstone Choice Model

A generalized Thurstone choice model, denoted as $\mathcal{T}_F$, is defined by a parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ that takes value in a parameter set $\Theta_n \subseteq \mathbf{R}^n$, and a cumulative distribution function F of a random variable that takes value in $\mathbf{R}$. Here θ_i represents the strength of item $i \in N$. We denote with f the density function of the cumulative distribution function F.

According to $\mathcal{T}_F$, the observations are such that for each observation, conditional on that the comparison set of this observation is S, the choice is item $i \in S$ with probability

$$p_{i,S}(\theta) = p_{|S|}(\theta_i - \theta_{S\setminus\{i\}}) \tag{1}$$

where

$$p_k(\mathbf{x}) = \int_{\mathbf{R}} f(z) \prod_{l=1}^{k-1} F(x_l + z) dz, \text{ for } \mathbf{x} \in \mathbf{R}^{k-1}. \tag{2}$$

Hereinafter, θ_A denotes the vector $\theta_A = (\theta_i, i \in A)$ for a non-empty set $A \subseteq N$, and, for brevity, with a slight abuse of notation, $a - \theta_A$ denotes the vector $(a - \theta_i, i \in A)$, for $a \in \mathbf{R}$.

A generalized Thurstone model of choice $\mathcal{T}_F$ follows from the following probabilistic generative model. For every observation with comparison set S, each item in this set is associated with independent random variables $(X_i,\ i \in S)$ that represent individual performances of these items, where each X_i is a sum of θ_i and a noise random variable ε_i with cumulative distribution function F. The choice $i \in S$ is the item that exhibits the largest performance, i.e. $p_{i,S}(\theta) = \mathbf{P}[X_i \geq \max_{j \in S} X_j]$, which corresponds to the asserted expression in (1).

Note that the probability distribution of choice depends only on the differences between the strength parameters. Hence, the probability distribution of choice for a parameter vector θ is equal to that under the parameter vector $\theta + c \cdot \mathbf{1}$, for any constant c, where $\mathbf{1}$ is the all-one vector. To allow for identifiability of the parameter vector, we admit the assumption that θ is such that $\sum_{i=1}^n \theta_i = 0$.

2.3. Special Generalized Thurstone Choice Models

Several special generalized Thurstone models of choice are given as follows.

(i) Gaussian noise with variance σ^2: $f(x) = \exp(-x^2/(2\sigma^2))/(\sqrt{2\pi}\sigma)$.

(ii) Double-exponential distribution of noise with parameter $\beta > 0$: $F(x) = \exp(-\exp(-(x + \beta\gamma)/\beta))$, where γ is the Euler-Mascheroni constant, which has variance $\sigma^2 = \pi^2\beta^2/6$.

(iii) Laplace distribution of noise with parameter β: $F(x) = \frac{1}{2}e^{\frac{x}{\beta}}$, for $x < 0$, and $F(x) = 1 - \frac{1}{2}e^{-\frac{x}{\beta}}$, for $x \geq 0$, which has variance $\sigma^2 = 2\beta^2$.

(iv) Uniform distribution of noise on $[-a, a]$: $f(x) = 1/(2a)$, for $x \in [-a, a]$, which has variance $\sigma^2 = a^2/3$.

For the special case of a generalized Thurstone model $\mathcal{T}_F$ with a double-exponential distribution of noise and a comparison set of cardinality k, we have

$$p_k(\mathbf{x}) = \frac{1}{1 + \sum_{i=1}^{k-1} e^{-x_i/\beta}}, \text{ for } \mathbf{x} \in \mathbf{R}^{k-1}.$$

Hence, for a comparison set $S \subseteq N$,

$$p_{|S|}(\theta_i - \theta_{S\setminus\{i\}}) = \frac{e^{\theta_i/\beta}}{\sum_{l \in S} e^{\theta_l/\beta}}, \text{ for } i \in S,$$

which corresponds to the well-known Luce's choice model.

In particular, for pair comparisons, we have the following two well known cases: (i) for the Gaussian distribution of noise, we have $p_2(x) = \Phi(x/(\sqrt{2}\sigma))$ where Φ is the cumulative distribution function of a standard normal random variable, and (ii) for the double-exponential distribution of noise, we have $p_2(x) = 1/(1 + e^{-x/\beta})$, which is a special case of the Luce's choice model and is commonly referred as the Bradley-Terry model.

2.4. Maximum Likelihood Estimation

For given input observations, the log-likelihood function, up to an additive constant, is equal to

$$\ell(\theta) = \sum_{S \subseteq N} \sum_{i \in S} w_{i,S} \log(p_{|S|}(\theta_i - \theta_{S\setminus\{i\}})). \tag{3}$$

The maximum likelihood estimator of the parameter vector θ is defined as a parameter vector $\hat{\theta}$ that maximizes the log-likelihood function over the set of parameters Θ_n, i.e. $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta_n} \ell(\theta)$. In particular, for pair comparisons, we can write the log-likelihood function as follows:

$$\ell(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \log\left(p_2(\theta_i - \theta_j)\right). \tag{4}$$

2.5. Some Key Definitions

We shall see that for the maximum likelihood parameter estimation problem, a special type of a matrix plays an important role. For every pair of items $\{i, j\}$ and a positive integer k, let $m_{i,j}(k)$ denote the number of observed comparison sets of cardinality k each containing the pair of items $\{i, j\}$. Let $w : \{1, 2, \dots, m\} \to \mathbf{R}_+$ be a decreasing function, we refer to as a *weight function*, which is given. We define the *pair-weight matrix* $\mathbf{M} = [m_{i,j}] \in \mathbf{R}_+^{n \times n}$ as follows:

$$m_{i,j} = \begin{cases} \frac{n}{m} \sum_{k \geq 2} w(k) m_{i,j}(k), & \text{if } i \neq j \\ 0, & \text{if } i = j. \end{cases} \tag{5}$$

Note that if all comparison sets are of cardinality k, then each non-diagonal element (i, j) of the pair-weight matrix is equal to, up to a multiplicative factor, the number of observed comparison sets that contain the pair of items $\{i, j\}$. For pair comparisons, this corresponds to the number of pair comparisons. The normalization factor n/m corresponds to a normalization with the mean number of comparison sets per item.

We say that a set of comparison sets is *unbiased*, if for each positive integer k and pair of items $\{i, j\}$, there is a common number of comparison sets of cardinality k that contain the pair of items $\{i, j\}$. An example of unbiased comparison sets is a fixture of games in some popular sport competitions that consists of games between pairs of teams such that each team plays against each other team equal number of times; e.g., fixtures of games in national football leagues like the one in Section M of the supplementary material.

Let $\mu(k)$ be the fraction of comparison sets of cardinality k. Then, for any unbiased set of comparison sets, for every positive integer k and pair of items $\{i, j\}$, it must hold

$$m_{i,j}(k) = \frac{\binom{n-2}{k-2}}{\binom{n}{k}} \mu(k) m = \frac{k(k-1)}{n(n-1)} \mu(k) m.$$

Hence, for every pair of items $\{i, j\}$, it holds that

$$m_{i,j} = \frac{1}{n-1} \sum_{k \geq 2} w(k) k(k-1) \mu(k). \tag{6}$$

We shall use the notation $\overline{\mathbf{M}}$ to denote the expected value of a pair-weight matrix $\mathbf{M}$, where the expectation is with respect to the distribution over the set of comparison sets. We say that comparison sets are *a priori unbiased* if $\overline{\mathbf{M}}$ is an unbiased matrix. For example, sampling each comparison set independently by uniform random sampling without replacement from the set of all items results in an a priori unbiased set of comparison sets. Note that any unbiased set of comparison sets is a priori unbiased.

We shall show that for the parameter estimation accuracy, the following parameters play an important role:

$$\gamma_{F,k} = \frac{1}{k^3 (k-1) (\partial p_k(\mathbf{0}) / \partial x_1)^2} \tag{7}$$

where

$$\frac{\partial p_k(\mathbf{0})}{\partial x_1} = \int_{\mathbf{R}} f(x)^2 F(x)^{k-2} dx. \tag{8}$$

We shall see that the algebraic connectivity of pair-weight matrices is a key factor that determines the estimation accuracy, for a suitable choice of the weight function that depends on the generalized Thurstone choice model $\mathcal{T}_F$. In particular, we shall see that the weight function should be set as defined by

$$w(k) = \left(k \frac{\partial p_k(\mathbf{0})}{\partial x_1} \right)^2. \tag{9}$$

For example, for the Luce's choice model this amounts to $w(k) = 1/(\beta k)^2$ and for a large class of generalized Thurstone choice models, the weight function is such that $w(k) = \Theta(1/k^2)$. Then, the mean square error of the maximum likelihood parameter estimator is decided by the pair-weight matrix $\mathbf{M}$ with the weight function (9), which is discussed in Section 3. We discuss the amount of $w(k)$ for various cumulative distributed functions F in Section 5.

2.6. Additional Notation

For a matrix $\mathbf{A}$, we denote with $\lambda_i(\mathbf{A})$ its i-th smallest eigenvalue. We denote with $\Lambda_{\mathbf{A}}$ the Laplacian matrix of matrix $\mathbf{A}$, i.e., $\Lambda_{\mathbf{A}} = \operatorname{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}$.

For any symmetric, non-negative, and irreducible matrix $\mathbf{A}$, its *Fiedler value* is defined as the smallest non-zero eigenvalue of the Laplacian matrix $\Lambda_{\mathbf{A}}$, i.e., equal to $\lambda_2(\Lambda_{\mathbf{A}})$. Please refer to Section A.5 of the supplementary material for more details about eigenvalues of $\Lambda_{\mathbf{A}}$.

3. Mean Square Error

In this section, we derive upper and lower bounds for the mean square error for the maximum likelihood parameter estimator of a generalized Thurstone choice model. For a generalized Thurstone choice model $\mathcal{T}_F$ with parameter $\theta^\star$, for any estimator $\hat{\theta}$, the mean square error $\mathrm{MSE}(\hat{\theta}, \theta^\star)$ is defined by

$$\mathrm{MSE}(\hat{\theta}, \theta^\star) = \frac{1}{n} \|\hat{\theta} - \theta^\star\|_2^2. \tag{10}$$

3.1. Pair Comparisons

In this section, we consider generalized Thurstone models $\mathcal{T}_F$ for pair comparisons, with the parameter set $\Theta_n = [-b, b]^n$, for $b \geq 0$.

We define $G_D = (N, E_D)$ to be a directed graph, with edge $(i,j) \in E_D$ if and only if $w_{i,j} > 0$; and the undirected graph $G_U = (N, E_U)$ where edge $(i,j) \in E_U$ if and only if $w_{i,j} + w_{j,i} > 0$. Let $\mathbf{M}$ be the pair-weight matrix with the weight function $w(k) = 1/k^2$. We define a condition **G** as follows:

G. G_U is connected, i.e., for every pair of vertices i and j, there exists a path that connects them.

Note that when G_U is connected, i.e., condition **G** holds true, then, $\lambda_2(\Lambda_{\mathbf{M}}) > 0$.

When $\log(p_2(x))$ is strictly concave for $x \in [-2b, 2b]$, i.e.,

$$\max_{x\in[-2b,2b]} \frac{d^2}{dx^2} \log(p_2(x)) < 0,$$

$\ell(\theta)$ is strictly concave under **G**.

Let us define $c_{F,b} = A/B$ where

$$A = \max_{x\in[-2b,2b]} \frac{d}{dx} \log(p_2(x))$$

and

$$B = \min_{x\in[-2b,2b]} \left| \frac{d^2}{dx^2} \log(p_2(x)) \right|.$$

Theorem 1. *Suppose that observations are according to a generalized Thurstone model $\mathcal{T}_F$ with parameter $\theta^\star \in [-b,b]^n$, for $n \geq 2$. If $\log(p_2(x))$ is a strictly concave function and* **G** *holds, then with probability at least $1-2/n$, the maximum likelihood estimator $\hat{\theta}$ satisfies*

$$\mathrm{MSE}(\hat{\theta}, \theta^\star) \leq c_{F,b}^2 \frac{n(\log(n)+2)}{\lambda_2(\Lambda_{\mathbf{M}})^2} \frac{1}{m}. \tag{11}$$

The result in Theorem 1 generalizes the characterization of the mean square error in (Negahban et al., 2012) and (Hajek et al., 2014) for the Bradley-Terry model to a generalized Thurstone choice model for pair comparisons.

Since the Bradley-Terry model is a generalized Thurstone choice model with noise according to the double-exponential distribution, we have $p_2(x) = 1/(1+e^{-x/\beta})$, for which we derive $A = 1/[\beta(1+e^{-2b/\beta})]$ and $B = e^{-2b/\beta}/[\beta^2(1+e^{-2b/\beta})^2]$, and hence $c_{F,b} = \beta(e^{2b/\beta}+1)$.

Condition (11) implies that for $\mathrm{MSE}(\hat{\theta}, \theta^\star) \leq \epsilon^2$ to hold for given $\epsilon > 0$, it suffices that

$$m \geq \frac{1}{\epsilon^2} c_{F,b}^2 \frac{1}{\lambda_2(\Lambda_{\mathbf{M}})^2} n(\log(n)+2). \tag{12}$$

The Fiedler value $\lambda_2(\Lambda_{\mathbf{M}})$ reflects how well is the pair-weight matrix $\mathbf{M}$ connected. If each pair is compared an equal number of times, then from (6), we have $m_{i,j} = 1/(2(n-1))$ for $i \neq j$, and in this case, $\lambda_2(\Lambda_{\mathbf{M}}) = \cdots = \lambda_n(\Lambda_{\mathbf{M}}) = n/(2(n-1))$. Hence, from the condition in (12), it suffices that

$$m \geq \frac{4}{\epsilon^2} c_{F,b}^2 \, n(\log(n)+2).$$

3.2. Arbitrary Cardinalities of Comparisons Sets

In this section, we derive upper and lower bounds for the mean square error when each comparison set consists of two or more items. Let K denote the set of distinct values of cardinalities of comparison sets observed in input data, or that can occur with a strictly positive probability if comparison sets are sampled from a distribution.

We consider a generalized Thurstone choice model $\mathcal{T}_F$ that satisfies the following assumptions:

A1 There exist $\overline{A}_{F,b} \geq \underline{A}_{F,b} > 0$ such that for all $S \subseteq N$ with $|S| \in K$ and $\{y,i,j\} \subseteq S$,

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log(p_{y,S}(\mathbf{0})) \geq 0$$

and, for all $\theta \in [-b,b]^n$, it holds

$$\underline{A}_{F,b} \leq \frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(p_{y,S}(\theta))}{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(p_{y,S}(\mathbf{0}))} \leq \overline{A}_{F,b}.$$

A2 There exist $\overline{B}_{F,b} \geq \underline{B}_{F,b} > 0$ such that for all $\theta \in [-b,b]^n$, $S \subseteq N$ with $|S| \in K$ and $y \in S$,

$$\underline{B}_{F,b} \leq \frac{p_{y,S}(\theta)}{p_{y,S}(\mathbf{0})} \leq \overline{B}_{F,b}.$$

A3 There exist $\overline{C}_{F,b} \geq \underline{C}_{F,b} > 0$ such that for all $\theta \in [-b,b]^n$, $S \subseteq N$ with $|S| \in K$ and $y \in S$,

$$\underline{C}_{F,b} \leq \frac{\|\nabla p_{y,S}(\theta)\|_2}{\|\nabla p_{y,S}(\mathbf{0})\|_2} \leq \overline{C}_{F,b}.$$

For $\theta \in [-b,b]^n$, the above conditions guarantee the local convexity of $-\log(p_{y,S}(\theta))$ (from **A1**) and bound $\lambda_2(\nabla^2 \log(p_{y,S}(\theta)))$ (from **A1**), $p_{y,S}(\theta)$ (from **A2**), and $\|\nabla p_{y,S}(\theta)\|_2$ (from **A3**).

Note that the constants $\underline{A}_{F,b}$, $\overline{A}_{F,b}$, $\underline{B}_{F,b}$, $\overline{A}_{F,b}$, $\underline{C}_{F,b}$, and $\overline{C}_{F,b}$ depend only on distribution F and parameter b. In general, in the limit as b goes to 0, all the lower- and upper-bound parameters in **A1**, **A2**, **A3** go to 1. Thus, in this limit, they are non-essential for the results presented in this section. In particular, if F is the double-exponential distribution, we can easily check that

$$\frac{\partial^2 \log(p_{y,S})}{\partial\theta_i\partial\theta_j} = \frac{p_{i,S}(\theta)p_{j,S}(\theta)}{\beta^2} \geq 0$$

and it is admissible to take $\underline{A}_{F,b} = e^{-4b/\beta}$, $\overline{A}_{F,b} = e^{4b/\beta}$, $\underline{B}_{F,b} = e^{-2b/\beta}$, $\overline{B}_{F,b} = e^{2b/\beta}$, $\underline{C}_{F,b} = e^{-4b/\beta}$, $\overline{C}_{F,b} = 4$, and $\sigma_{F,K} = 1/\beta^2$ for all $b > 0$.

The following theorem establishes an upper bound for the mean square error.

Theorem 2. *Assume* **A1**, **A2** *and* **A3**. *Let* $\overline{\mathbf{M}}_F$ *be the pair-weight matrix with the weight function (9) and* $D_{F,b} = \overline{C}_{F,b}/(\underline{A}_{F,b}\underline{B}_{F,b})$. *Suppose that*

$$m \geq 32\frac{\sigma_{F,K}}{\underline{B}_{F,b}}\frac{1}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})}n\log(n),$$

then, with probability at least $1 - 3/n$,

$$\mathrm{MSE}(\hat{\theta}, \theta^\star) \leq 32D_{F,b}^2\sigma_{F,K}\frac{n(\log(n)+2)}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})^2}\frac{1}{m}$$

where $\sigma_{F,K} = 1/\min_{k\in K}\gamma_{F,k}$.

If, in addition to the assumptions of Theorem 2, all comparison sets are of cardinality $k \geq 2$, then, the statement of the theorem holds with

$$\frac{\sigma_{F,K}}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})} = \left(1 - \frac{1}{k}\right)\frac{1}{\lambda_2(\Lambda_{\overline{\mathbf{M}}})}$$

and

$$\frac{\sigma_{F,K}}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})^2} = \left(1 - \frac{1}{k}\right)^2\gamma_{F,k}\frac{1}{\lambda_2(\Lambda_{\overline{\mathbf{M}}})^2}$$

where $\gamma_{F,k}$ is defined in (7), and $\overline{\mathbf{M}}$ is the pair-weight matrix with the weight function $w(k) = 1/k^2$.

If, in addition, each comparison set is sampled independently, uniformly at random without replacement from the set of all items, then the statement of Theorem 2 holds with

$$\mathrm{MSE}(\hat{\theta}, \theta^\star) \leq 32D_{F,b}^2\gamma_{F,K}\frac{n(\log(n)+2)}{m}\left(1 - \frac{1}{n}\right)^2 \tag{13}$$

since

$$\frac{\sigma_{F,K}}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})} = 1 - \frac{1}{n} \text{ and } \frac{\sigma_{F,k}}{\lambda_2(\Lambda_{\overline{\mathbf{M}}_F})^2} = \left(1 - \frac{1}{n}\right)^2\gamma_{F,k}.$$

In the following theorem, we establish a lower bound.

Theorem 3. *Any unbiased estimator* $\hat{\theta}$ *satisfies*

$$\mathbf{E}[\mathrm{MSE}(\hat{\theta}, \theta^\star)] \geq \frac{1}{\overline{A}_{F,b}\overline{B}_{F,b}}\left(\sum_{i=2}^{n}\frac{1}{\lambda_i(\Lambda_{\overline{\mathbf{M}}_F})}\right)\frac{1}{m}.$$

If all comparison sets are of cardinality k, then any unbiased estimator $\hat{\theta}$ satisfies the inequality in Theorem 3 with

$$\sum_{i=2}^{n}\frac{1}{\lambda_i(\Lambda_{\overline{\mathbf{M}}_F})} = \left(1 - \frac{1}{k}\right)\gamma_{F,k}\sum_{i=2}^{n}\frac{1}{\lambda_i(\Lambda_{\overline{\mathbf{M}}})}.$$

If, in addition, each comparison set is drawn independently, uniformly at random from the set of all items, then any unbiased estimator $\hat{\theta}$ satisfies the inequality in Theorem 3 with

$$\mathbf{E}[\mathrm{MSE}(\hat{\theta}, \theta^\star)] \geq \frac{1}{\overline{A}_{F,b}\overline{B}_{F,b}}\gamma_{F,k}\frac{n}{m}\left(1 - \frac{1}{n}\right)^2, \tag{14}$$

since

$$\sum_{i=2}^{n}\frac{1}{\lambda_i(\Lambda_{\overline{\mathbf{M}}_F})} = \gamma_{F,k}\left(1 - \frac{1}{n}\right)^2 n.$$

We have tight upper bound (13) and low bound (14) for the means square error for the maximum likelihood parameter estimator. The difference between (13) and (14) is $O(\log(n))$. Indeed, the $\log(n)$ gap between upper and lower bound allows us to say "with probability $1 - 3/n$". One can remove the $\log(n)$ gap by using a constant probability (e.g., "with probability $3/4$").

The tight upper and lower bound tell us that under the given assumptions, for the mean square error to be smaller than a constant, it is necessary that the number of comparisons satisfies $\frac{m}{\gamma_{F,k}} = \Omega(n)$. Therefore, for the same performance guarantee, we require more comparisons as $\gamma_{F,k}$ increases.

4. Classification of Items of Two Classes

In this section, we consider a generalized Thurstone choice model $\mathcal{T}_F$ with parameter θ that takes value in $\Theta_n = \{-b, b\}^n$, for parameter $b > 0$. This is a special case where each item is either of two classes: a low or a high class. We consider a classification problem, where the goal is to correctly classify each item as either of low or high class, based on observed input data of choices.

Suppose that $\theta_i = b$ for all $i \in N_1$ and $\theta_i = -b$ for all $i \in N_2$ where $N_1 \cup N_2 = N$ and $|N_1| = |N_2| = n/2$. Without loss of generality, assume that $N_1 = \{1, \ldots, n/2\}$ and $N_2 = \{n/2 + 1, \ldots, n\}$.

We consider a *point score ranking method* that outputs an estimate $\hat{N}_1$ of the set of items of high class and $\hat{N}_2$ that contains the remaining items, which is defined by the following algorithm:

1. Observe outcomes of m observations and associate each item with a point score defined as the number of comparison sets in which this item is the chosen item.
2. Sort items in decreasing order of the point scores.

3. Output $\hat{N}_1$ defined as the set of top $n/2$ items (with uniform random tie break) and $\hat{N}_2$ defined as the set of remaining items.

Theorem 4. *Suppose that* $b \leq 4/(k^2 \partial p_k(\mathbf{0})/\partial x_1)$ *and*

$$b \max_{\mathbf{x} \in [-2b,2b]^{k-1}} \|\nabla^2 p_k(\mathbf{x})\|_2 \leq \frac{\partial p_k(\mathbf{0})}{\partial x_1}. \qquad (15)$$

Then, for every $\delta \in (0,1]$, *if*

$$m \geq 64 \frac{1}{b^2} \left(1 - \frac{1}{k}\right) \gamma_{F,k}\, n(\log(n) + \log(1/\delta))$$

the point score ranking method correctly identifies the classes of all items with probability at least $1 - \delta$.

The bound of the theorem is tight as established in the following theorem.

Theorem 5. *Suppose that* $b \leq 1/(6k^2 \partial p_k(\mathbf{0})/\partial x_1)$ *and that condition (15) holds. Then, for every even number of items such that* $n \geq 16$, *and* $\delta \in (0, 1/4]$, *for any algorithm to correctly classify all items with probability at least* $1 - \delta$, *it is necessary that*

$$m \geq \frac{1}{62} \frac{1}{b^2} \left(1 - \frac{1}{k}\right) \gamma_{F,k}\, n(\log(n) + \log(1/\delta)).$$

Again, from Theorem 4 and 5, we can conclude that the higher $\gamma_{F,k}$ has the more error.

5. Discussion of Results

In this section, we discuss how the number of observations needed for given parameter estimation error tolerance depends on the cardinality of comparison sets. We found in Section 3.2 and Section 4 that for a priori unbiased schedules of comparisons, where each comparison set is of cardinality k and is drawn independently, uniformly at random from the set of all items, the required number of observations to bring down the mean square error or correctly classify items of two classes with high probability, the number of observations is of the order $\gamma_{F,k}$, defined in (7).

The values of parameters $\partial p_k(\mathbf{0})/\partial x_1$ and $\gamma_{F,k}$ for our example generalized Thurstone choice models $\mathcal{T}_F$ in Section 2.3 are presented in Table 1.

For every cumulative distribution F in Table 1, $\gamma_{F,k}$ is a decreasing function in k. Thus, we have a better performance by increasing the size of comparisons. The decreasing speed of $\gamma_{F,k}$ depends on the cumulative distribution F. Note that for both double-exponential and Laplace distributions of noise $\gamma_{F,k} = \Theta(1)$, and for Gaussian distribution of noise $\gamma_{F,k} = O(1/k^{\epsilon})$. On the other hand, for uniform distribution of noise, $\gamma_{F,k} = \Theta(1/k^2)$.

Table 1. The values of parameters for our examples of $\mathcal{T}_F$.

F	$\frac{\partial p_k(\mathbf{0})}{\partial x_1}$	$\gamma_{F,k}$
Gaussian	$O(\frac{1}{k^{2-\epsilon}})$	$\Omega(\frac{1}{k^{2\epsilon}})$
Double-exponential	$\frac{1}{\beta k^2}$	$\beta^2 \frac{k}{k-1}$
Laplace	$\frac{1-1/2^{k-1}}{\beta k(k-1)}$	$\beta^2 \frac{k-1}{k(1-1/2^{k-1})^2}$
Uniform	$\frac{1}{2a(k-1)}$	$4a^2 \frac{k-1}{k^3}$

In general, the value of parameter $\gamma_{F,k}$ admits the following lower and upper bounds.

Proposition 6. *For the value of parameter* $\gamma_{F,k}$, *the following two claims hold:*

1. *For every cumulative distribution function* F *with an even and continuously differentiable density function, we have* $\gamma_{F,k} = O(1)$.
2. *For every cumulative distribution function* F *with a density function such that* $f(x) \leq C$ *for all* $x \in \mathbf{R}$, *for a constant* $C > 0$, $\gamma_{F,k} = \Omega(1/k^2)$.

We observe that both double-exponential and Laplace distributions of noise are extremal in achieving the upper bound of $O(1)$ for the value of parameter $\gamma_{F,k}$, asymptotically for large k. On the other hand, a uniform distribution of noise is extremal in achieving the lower bound $\Omega(1/k^2)$ for the value of parameter $\gamma_{F,k}$. More generally, we can show that $\gamma_{F,k} = \Theta(1/k^2)$ for any cumulative distribution function F with the density function such that $f(x) \geq C$ for every point x of its support, for a constant $C > 0$.

Numerical Examples.

We consider the following simulation experiment. We fix the values of the number of items n and the number of comparisons m, and consider a choice of a generalized Thurstone model $\mathcal{T}_F$ for the value of parameter $\theta^\star = \mathbf{0}$. We consider comparison sets of the same cardinality of value k that are independent uniform random samples from the set of all items. For every fixed value of k, we run 100 repetitions to estimate the mean square error. We do this for the distribution of noise according to a double-exponential distribution (Bradley-Terry model) and according to a uniform distribution, both with unit variance.

Figure 1 shows the results for the setting of parameters $n = 10$ and $m = 100$. The results clearly demonstrate that the mean square error exhibits qualitatively different relations with the cardinality of comparison sets for the two generalized Thurstone models. Our theoretical results in Section 3.2 suggest that the mean square error should decrease with the cardinality of comparison sets as $1/(1 - 1/k)$ for the double-exponential distribution, and as $1/k^2$ for the uniform distribution of noise. Observe that

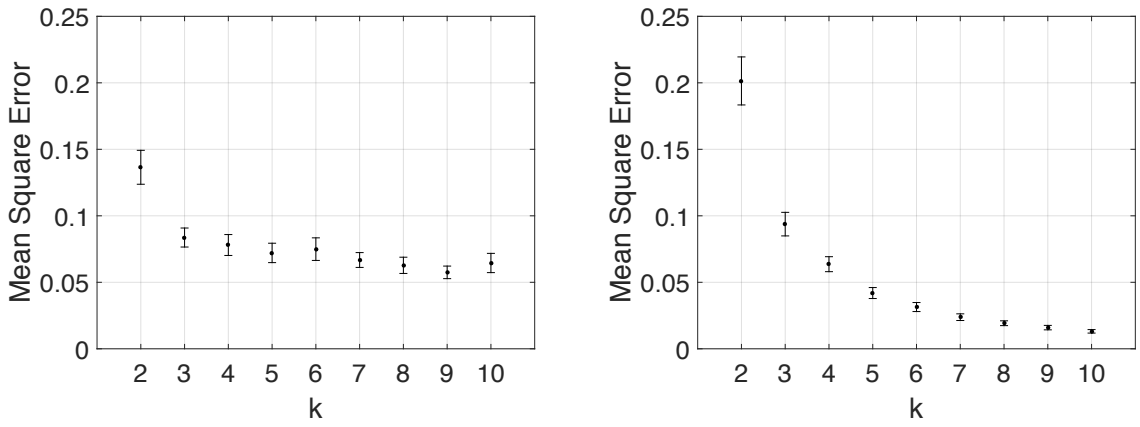

Figure 1. Mean square error for two different generalized Thurstone choice models $\mathcal{T}_F$: (left) F is a double-exponential distribution, and (right) F is a uniform distribution. The vertical bars denote 95% confidence intervals. The results confirm two qualitatively different relations with the cardinality of comparison sets as suggested by the theory.

the latter two terms decrease with k to a strictly positive value and to zero value, respectively. The empirical results in Figure 1 confirm these claims.

We found that Fiedler value of a pair-weight matrix is an important factor that determines the mean square error in Section 3.1 and Section 3.2. In Section M of the supplementary material, we evaluate Fiedler value for different pair-weight matrices of different schedules of comparisons.

6. Conclusion

The results of this paper elucidate how the parameter estimation accuracy for a generalized Thurstone choice model depends on the given model and the structure of comparison sets. They show that a key factor is an eigenvalue gap of a pair-weight matrix that reflects its algebraic connectivity, which depends in a particular way on the given model. It is shown that for a large class of generalized Thurstone choice models, including all popular instances used in practice, there is a diminishing returns decrease of the estimation error with the cardinality of comparison sets, which is rather slow for comparison sets of three of more items. This offers a guideline for the designers of schedules of competitions to ensure that the schedule has a well-connected pair-weight matrix and to expect limited gains from comparison sets of large sizes.

References

Boyd, Stephen. Convex optimization of graph Laplacian eigenvalues. In *Proceedings of the International Congress of Mathematicians*, pp. 1311–1319, 2006.

Bradley, Ralph Allan and Terry, Milton E. Rank analysis of incomplete block designs: I. method of paired comparisons. *Biometrika*, 39(3/4):324–345, Dec 1952.

Bradley, Ralph Allan and Terry, Milton E. Rank analysis of incomplete block designs: II. additional tables for the method of paired comparisons. *Biometrika*, 41(3/4): 502–537, Dec 1954.

Dawkins, Richard. A threshold model of choice behaviour. *Animal Behaviour*, 17(Part 1):120–133, February 1969.

Elo, Arpad E. *The Rating of Chessplayers*. Ishi Press International, 1978.

Graepel, Thore, Minka, Tom, and Herbrich, Ralf. Trueskill(tm): A bayesian skill rating system. In *Proc. of NIPS 2006*, volume 19, pp. 569–576, 2006.

Hajek, Bruce, Oh, Sewoong, and Xu, Jiaming. Minimax-optimal inference from partial rankings. In *Proc. of NIPS 2014*, pp. 1475–1483, 2014.

Hayes, Thomas P. A large-deviation inequality for vector-valued martingales. URL http://www.cs.unm.edu/~hayes/papers/VectorAzuma/VectorAzuma20030207.pdf.

Luce, R. Duncan. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, 1959.

McCullagh, P. and Nelder, J. A. *Generalized Linear Models*. Chapman & Hall, New York, 2 edition, 1989.

Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Negahban, Sahand, Oh, Sewoong, and Shah, Devavrat. Iterative ranking from pair-wise comparisons. In *Proc. of NIPS 2012*, pp. 2483–2491, 2012.

Nelder, J. A. and Wedderburn, R. W. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384, 1972.

Rajkumar, Arun and Agarwal, Shivani. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proc. of ICML 2014*, pp. 118–126, 2014.

Stern, Hal. Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences*, 23(1): 103–117, 1992.

Thurstone, L. L. A law of comparative judgment. *Psychological Review*, 34(2):273–286, 1927.

Tropp, Joel A. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Vojnović, Milan. *Contest Theory: Incentive Mechanisms and Ranking Methods*. Cambridge University Press, 2016.

Yellott, John I. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement and the double exponential distribution. *Journal of Mathematical Psychology*, 15:109–144, 1977.

Zermelo, E. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Math. Z.*, 29:436–460, 1929.