Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Lagrangian Inference for Ranking Problems

Yue Liu, Ethan X. Fang, Junwei Lu

To cite this article:
Yue Liu, Ethan X. Fang, Junwei Lu (2022) Lagrangian Inference for Ranking Problems. Operations Research

Published online in Articles in Advance 17 Jun 2022

. https://doi.org/10.1287/opre.2022.2313

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–22
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Lagrangian Inference for Ranking Problems

Yue Liu,[a] Ethan X. Fang,[b] Junwei Lu[c,*]

[a] Department of Statistics, Harvard University, Boston, Massachusetts 02138; [b] Department of Biostatistics & Bioinformatics, Duke University, Durham, North Carolina 27705; [c] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02130
*Corresponding author

Contact: yueliu@fas.harvard.edu (YL); xingyuan.fang@duke.edu, https://orcid.org/0000-0003-3762-9155 (EXF); junweilu@hsph.harvard.edu (JL)

Received: June 5, 2021
Revised: December 15, 2021
Accepted: May 2, 2022; May 17, 2022
Published Online in Articles in Advance: June 17, 2022
Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2022.2313

Abstract. We propose a novel combinatorial inference framework to conduct general uncertainty quantification in ranking problems. We consider the widely adopted Bradley-Terry-Luce (BTL) model, where each item is assigned a positive preference score that determines the Bernoulli distributions of pairwise comparisons' outcomes. Our proposed method aims to infer general ranking properties of the BTL model. The general ranking properties include the "local" properties such as if an item is preferred over another and the "global" properties such as if an item is among the top K-ranked items. We further generalize our inferential framework to multiple testing problems where we control the false discovery rate (FDR) and apply the method to infer the top-K ranked items. We also derive the information-theoretic lower bound to justify the minimax optimality of the proposed method. We conduct extensive numerical studies using both synthetic and real data sets to back up our theory.

Funding: E. X. Fang was partially supported by the National Science Foundation [Grants DMS-1820702, DMS-1953196, and DMS-2015539] and a Whitehead Scholarship. J. Lu was partially supported by the National Science Foundation [Grant DMS-1916211], and the National Institutes of Health [Grants R35CA220523-04, R01ES32418-01, and U01CA209414-0].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2022.2313.

Keywords: combinatorial inference • ranking • pairwise comparisons • Bradley-Terry-Luce model • minimax lower bound

1. Introduction

Ranking problems aim to study relative orderings of some set of items and find many applications such as sports competition (Pelechrinis et al. 2016, Xia et al. 2018), online gamers ranking (e.g., Microsoft TrueSkill ranking system; Minka et al. 2007, 2018), web search and information retrieval (Dwork et al. 2001, Bouadjenek et al. 2013, Guo et al. 2020), recommendation systems (Baltrunas et al. 2010, He et al. 2018, Geyik et al. 2019), crowdsourcing (Chen et al. 2013, Suh et al. 2017, Liang and de Alfaro 2020), gene ranking (Boulesteix and Slawski 2009, Kolde et al. 2012, Kim et al. 2015), assortment optimization (Aouad et al. 2018, Li et al. 2018), and healthcare (Adelman 2020), among many others. Because of the practical importance, ranking problems draw significant attention from different communities such as operations research (McFadden 1974, Mohammadi and Rezaei 2020), statistics (Hunter 2004, Chen et al. 2020), machine learning (Richardson et al. 2006, Guo et al. 2020), and sociology (Brown 2003, Subochev et al. 2018).

In ranking problems, given some comparisons among pairs of n items, we aim to infer the relative ranking of these items. Many models are proposed to study this problem, and one of the most widely used parametric models is the Bradley-Terry-Luce (BTL) model (Bradley and Terry 1952, Luce 1959). In the BTL model, each item is assigned a latent positive preference score that determines its rank, and the latent scores determine the relative preference among the n items.

Based on the BTL model, there are several tracks of works that estimate the ranks of the items by estimating the latent scores. The first track is the rank centrality method (Dwork et al. 2001, Maystre and Grossglauser 2015, Jang et al. 2016, Vigna 2016, Negahban et al. 2017), which is also known as the spectral method. This class of methods connects pairwise comparisons with random walk over the comparison graph. In particular, each node in the graph represents an item, and the probability of moving from node i to node j equals the probability that item j is preferred over item i. Based on this approach, Negahban et al. (2017) show that the preference scores of items coincide the stationary distribution under the random walk and derive a fast rate of convergence of the estimator for the scores. Chen et al. (2019) further improve the convergence rate in Negahban et al. (2017) by removing the logarithmic factor. The second track is based on considering the regularized maximum likelihood estimator (MLE) (Ford 1957, Hunter 2004, Lu and Negahban 2015). This approach estimates the latent scores by maximizing the regularized

likelihood function, and Chen et al. (2019) derive the rate of convergence of the estimator under the ℓ_2 regularization. Negahban et al. (2018) also consider nuclear norm regularization. In addition, Azari Soufiani et al. (2013) consider the method of moments for the Plackett-Luce model, and Mosteller (2006), Jiang et al. (2011), and Neudorfer and Rosset (2018) consider general least square methods, and their estimation consistency for preference scores are established by various works (Duchi et al. 2010; Rajkumar and Agarwal 2014, 2016; Chen and Suh 2015; Maystre and Grossglauser 2015; Jang et al. 2016).

Despite the aforementioned significant progress of rank estimation, the uncertainty quantification in ranking problems remains largely unexplored, which is of crucial importance in practice. For example, saying that player i is ranked higher than player j without a confidence score is not very informative in practice. In this paper, we propose a novel combinatorial inferential framework for testing ranking properties. In particular, given n items, we define a ranking list γ as a permutation over the set of n items $[n] = \{1, \ldots, n\}$, and let $\mathcal{R}$ be the set of all possible rankings (i.e., all possible permutation over $[n]$). Let ranking list γ^* be the true underlying ranking of n items. We aim to test whether γ^* satisfies certain ranking properties based on partial pairwise comparison observations. For example, let $\mathcal{R}_i$ be a subset of $\mathcal{R}$ representing the ranking property with respect to item i. We test the general ranking property for a given item i, that is, whether item i has certain properties, that is,

$$H_0: \text{ item } i \text{ does not satisfy the property} \quad \text{v.s.}$$
$$H_a: \text{ item } i \text{ satisfies the property,}$$

which is equivalent to

$$H_0: \gamma^* \notin \mathcal{R}_i \quad \text{v.s.} \quad H_a: \gamma^* \in \mathcal{R}_i.$$

1.1. Motivating Applications

The inference in ranking problems finds many applications. For instance, it is of practical interest to test whether movie A is preferred over movie B on average and test whether chess player C is stronger than player D. Such problems are pairwise ranking inference problems that fit into our framework as defined in the following example.

Example 1.1 (Pairwise Ranking Inference). Consider testing whether item i is ranked higher than item j. Let $\mathcal{R}_i$ be the set of all possible rankings that item i is ranked higher than item j. We consider the following hypothesis testing problem that

$$H_0: \text{ Item } j \text{ is ranked higher than item } i \quad \text{v.s.}$$
$$H_a: \text{ Item } i \text{ is ranked higher than item } j.$$

Another important application of inference in ranking problems is the top-K inference. For instance, in recommendation systems, one important goal is to find a few most appealing items for the users (Cremonesi et al. 2010). In biomedical studies, only a small subset of top-ranked genes is informative, and it is crucial for the investigators to identify this set of genes to perform detailed analysis (Boulesteix and Slawski 2009). In assortment optimization, the challenge is to identify a subset of items that maximize revenue based on customer preferences (Li et al. 2018, Aouad et al. 2018). We first summarize the single top-K inference problem in the following example.

Example 1.2 (Single Top-K Inference). Consider testing whether item i is among the top-K items (a special case is $K = 1$). Here $\mathcal{R}_i$ is the set of all possible rankings that item i is among the top-K items. We consider the following hypothesis testing problem that

$$H_0: \text{ Item } i \text{ is not among the top-}K \text{ items} \quad \text{v.s.}$$
$$H_a: \text{ Item } i \text{ is among the top-}K \text{ items.}$$

We then extend the problem to the multiple testing setup, where the goal is to infer the set of all top-K items.

Example 1.3 (Top-K Inference). Consider the problem of identifying the set of top-K items. Here $\mathcal{R}_i$ is the set of all possible rankings that item i is among the top-K items, $i \in [n]$. We consider the following multiple testing problem that

$$H_0: \text{ Item } i \text{ is not among the top-}K \text{ items} \quad \text{v.s.}$$
$$H_a: \text{ Item } i \text{ is among the top-}K \text{ items,} \quad \text{for } i \in [n].$$

1.2. Major Contributions

To the best of our knowledge, this paper provides the first inferential framework for ranking problems. Our proposed method can test a broad class of hypotheses for ranking problems. Theoretically, we show that the p values are valid, and our procedures are powerful. We summarize the major contributions here.

- We are among the first to study general inferential approaches for ranking problems beyond estimation, and we propose a novel general framework for inferring different ranking properties. We show that our proposed methods are asymptotically valid and powerful. Furthermore, we generalize the method to the more challenging multiple testing setup to widen the applicability.
- In our inferential framework, we propose a novel general Lagrangian debiasing procedure to handle the constrained parameter space. Our Lagrangian debiasing procedure addresses the challenge raised by the nonidentifiability of the BTL model. Most existing works on high-dimensional inference, such as Zhang and Zhang (2014), Van de Geer et al. (2014), and Ning and Liu (2017), focus on inferring the parameter under the unconstrained space and do not apply to the BTL model

because of the constraints. By considering the optimality condition of the Lagrangian dual problem, our proposed approach provides a new tool for high-dimensional inference under general constraints. We also derive the asymptotic distribution of our debiased estimator. We point out that this new Lagrangian debiasing procedure can be applied to general high-dimensional constrained inferential problems beyond ranking problems, which itself is of great interest.

- We provide a new framework to derive the minimax lower bound for multiple testing in ranking problems, which provides new theoretical insights. To the best of our knowledge, this is the first time that such a lower bound is derived. In particular, let the preference score vector be $\theta \in \mathbb{R}^n$, which represents the scores of all n items that determine the ranks of all items. We first define a new minimax risk for the multiple testing problems that

$$\mathfrak{R} = \inf_{\psi} \sup_{\theta} \mathbb{P}(\#\text{ false positives} + \#\text{ false negatives} \geq 1),$$

where the infimum is taken over all possible selection procedure ψ. Here the risk is the probability of making at least one type I or type II error. If the minimax risk $\mathfrak{R} \geq 1 - \epsilon$ for some constant $\epsilon > 0$, we say that any procedure fails for the multiple testing problem because they cannot control the type I error or type II error in the minimax sense. To derive the necessary conditions for controlling minimax risk, we further define a novel distance $\Delta(\theta)$ in (4.2) and a divider set $\mathcal{M}(\theta)$ in Definition 5.1, which capture the combinatorial structures of ranking properties. Intuitively, the distance is a signal strength for selecting the items of interest; the divider set is the set of items that are crucial for selecting the items, and the size of the divider set increases as the distance decreases. We show that the numerical signal strength $\Delta(\theta)$ and combinatorial signal strength $|\mathcal{M}(\theta)|$ together measure the difficulty in the multiple testing problems. We also give two concrete examples where $\mathfrak{R}$ is arbitrarily close to one if $\Delta(\theta) \lesssim \sqrt{\frac{\log n}{npL}}$. In addition, we show that our lower bound matches our upper bound to justify the optimality of the proposed method.

1.3. Literature Review

1.3.1. Ranking Problem. There has been a long history of works on ranking problems (Mallows 1957, Keener 1993, Altman and Tennenholtz 2005, Jiang et al. 2011, Osting et al. 2013, Vigna 2016, Ding et al. 2018, Filiberto et al. 2018, Guo et al. 2020, Pujahari and Sisodia 2020). Some ranking systems are based on explicit preference scores or ratings provided by individuals, which is closely related to the matrix completion problem (Candès and Recht 2009, Negahban and Wainwright 2012). In these problems, an individual only provides scores for a subset of items, and we estimate the individual's preference scores for other items. However, users' explicit scores can be inconsistent and noisy, or even not available in some cases. This motivates researchers to develop methods for ranking aggregations from comparison results or partial rankings provided by users (Saaty 2003; Ailon et al. 2008; Ailon 2010; Gleich and Lim 2011; Ammar and Shah 2011, 2012; Farnoud et al. 2012; Volkovs and Zemel 2012; Jang et al. 2017, 2018; Nápoles et al. 2017; Swain et al. 2017; Chen et al. 2018b; Zhang 2020).

Another important track of works on ranking problems are based on pairwise comparison data (Kendall and Smith 1940; Kendall 1955; Adler et al. 1994; Talluri and Van Ryzin 2006; Beutel et al. 2019; Jain et al. 2020; Chen et al. 2021, 2022). For instance, Lu and Boutilier (2011) study the Mallows model from pairwise comparisons. Chen et al. (2022) study the sequential design with pairwise comparisons. Chen and Suh (2015) propose a new two-step method called the spectral MLE and prove that it is minimax optimal. Jang et al. (2016) show that the spectral method itself is optimal for identifying the top-K items in the sense of achieving the minimal sample size. Chen et al. (2019) further study the sample complexity of regularized MLE and spectral method in a sparse pairwise comparison setting.

There are other general frameworks on ranking problems such as the Thurstone model (Thurstone 1927, Vojnovic and Yun 2017, Orbán-Mihálykó et al. 2019, Jin et al. 2020) and Plackett-Luce model (Guiver and Snelson 2009, Hajek et al. 2014). For instance, Jin et al. (2020) propose a heterogeneous Thurstone model capturing heterogeneity of different individuals and propose an algorithm to estimate the preference score vector and heterogeneity. Beyond parametric models, there are also nonparametric methods for ranking problems. For instance, Shah and Wainwright (2017) analyze a simple counting algorithm proposed by Copeland (1951), which counts the numbers of wins of each item, and show its optimality and robustness. Shah et al. (2016), Chen et al. (2018a), and Pananjady et al. (2017) consider the strong stochastically transitive (SST) model for pairwise comparisons. Furthermore, some other works consider ranking problems under specific settings such as active-ranking (Jamieson and Nowak 2011, Busa-Fekete et al. 2013, Heckel et al. 2019), and crowd-sourcing (Chen et al. 2013, 2016; Suh et al. 2017; Liang and de Alfaro 2020). However, we point out that all these works focus on the estimation problem and do not consider uncertainty quantification and inferential methods in ranking. One exception is Hall and Miller (2009). This work focuses on using m-out-of-n bootstrap to estimate the distribution of an empirical rank, which requires empirical choice of m and is of less practical

interest. In contrast, we provide a more general framework that solves the problems of practical interest.

1.3.2. Constrained Inference. The inference under equality or inequality constraints is of great interest in literature. The low-dimensional constrained inference dates back to Chernoff (1954), which proves that the likelihood ratio weakly converges to a weighted chi-square distribution for constrained testings. Under the low-dimensional setting, the constrained inference has been further studied in Gourieroux et al. (1982), Kodde and Palm (1986), Rogers (1986), Shapiro (1988), Wolak (1989), Molenberghs and Verbeke (2007), and Susko (2013), among many others. For the high-dimensional constrained inference, Yu et al. (2019) assume the existence of natural constraint on parameters and test whether the parameters lie on the boundary of the constraint. By applying the debiasing approach in Ning and Liu (2017), the authors study the asymptotic distribution of test statistics under constraints.

We note that the previously mentioned methods cannot be applied to solve our problem. This is mainly because of the unique challenge that, in our setting, the Fisher information matrix is singular because of the nonidentifiability issue.

During the revision of this work, we note an arXiv work (Gao et al. 2021) for ranking inference, which focuses on inferring the latent scores without debiasing. In comparison, our work focuses on inferring the combinatorial structures of the rankings.

1.3.3. Paper Organization. The rest of our paper is organized as follows. In Section 2, we introduce some preliminaries of ranking problems and some ranking properties. In Section 3, we present our debiased estimator with constraints. We then provide the general hypothesis testing. In Section 4, we extend our method to handle multiple testing problems. In Section 5, we present the lower bound theory with applications to several examples. We provide numerical results in Section 6 and some discussions in Section 7.

1.3.4. Notations. Let $|A|$ represent the cardinality of set A, and $[n]$ represent the set of $\{1,\cdots,n\}$ for $n \in \mathbb{Z}^+$. For vector $v = (v_1,\ldots,v_d)^T \in \mathbb{R}^d$, and $1 \le q \le \infty$, we define norm of v as $\|v\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$. In particular, $\|v\|_\infty = \max_{1\le i\le d} |v_i|$. For a matrix $M = [M_{ij}]$, let ℓ_1-norm $\|M\|_1 = \max_j \sum_i |M_{ij}|$, ℓ_∞-norm $\|M\|_\infty = \max_i \sum_j |M_{ij}|$, and the operator norm $\|M\|_2 = \sigma_{\max}(M)$ where $\sigma_{\max}(M)$ represents the largest singular value of matrix M. In addition, $a_n = O(b_n)$ or $a_n \lesssim b_n$ means there exists a constant $C > 0$ such that $a_n \le Cb_n$, and $a_n = o(b_n)$ means $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$. we write $a_n \asymp b_n$ if $C \le a_n/b_n \le C'$ for some $C, C' > 0$. For a sequence of random variables $\{X_n\}$, we write $X_n \xrightarrow{d} X$ if X_n converges in distribution to the random variable X. Throughout the paper, we let $C, C_1, C_2, \cdots, c, c_1, c_2, \cdots$ be generic constants which may change in different places.

2. Preliminaries and Problem Setup

In this section, we provide some preliminaries to facilitate our discussions. We first briefly review the BTL model and introduce our data generating scheme. Then, we provide the definitions of rankings and ranking properties.

2.1. BTL Model

We consider the BTL parametric model (Bradley and Terry 1952, Luce 1959). This model assumes a hidden preference score $\omega_i^* > 0$ for each item i, $1 \le i \le n$. The scores determine the ranking and the distributions of comparison results. Let $\omega^* = (\omega_1^*,\cdots,\omega_n^*)^\top \in \mathbb{R}^n$ be the true preference score vector, and its log-transformation is

$$\theta^* = (\theta_1^*,\cdots,\theta_n^*)^\top, \text{ where } \theta_i^* = \log\omega_i^*.$$

Here $\omega_j^* > \omega_i^*$ or $\theta_j^* > \theta_i^*$ means that item j is ranked higher (preferred) than item i. In this paper, for ease of presentation, we consider the case that all scores are in a bounded domain that $w_i^* \in [w_{\min}, w_{\max}]$ for all $i \in [n]$, where $w_{\min}, w_{\max} > 0$. We let $\kappa = \omega_{\max}/\omega_{\min}$ be the condition number, and κ is a constant which does not depend on n.

When we collect data, we compare the items in a pairwise fashion. To model the random pairs for comparisons, we adopt the Erdös-Rényi random graph. In particular, suppose we have an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{V} = [n]$ is the vertex set, and $\mathcal{E} \subseteq [n] \times [n]$ is the edge set. In the Erdös-Rényi random graph $\mathcal{G}(n,p)$, each edge is drawn independently from a Bernoulli distribution with probability p. Here we assume that for each pair $(i,j) \in \mathcal{E}$, we observe the comparisons L times. We assume for all pairs in $\mathcal{E}$, we have a same number of observations for ease of presentation, but our proposed method can be easily generalized to handle the general setting where we have different numbers of observations of different pairs. Denote by $y_{i,j}^{(\ell)}$ the ℓ th comparison between items i and j for some $i < j$, which depends only on the relative scores of the two items. We assume that each $y_{i,j}^{(\ell)}$ is generated independently from a Bernoulli distribution that

$$y_{i,j}^{(\ell)} \overset{\text{ind.}}{=} \begin{cases} 1, & \text{with probability } \dfrac{w_j^*}{w_i^* + w_j^*} = \dfrac{e^{\theta_j^*}}{e^{\theta_i^*} + e^{\theta_j^*}} \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

where $y_{i,j}^{(\ell)} = 1$ means item j is preferred over item i. Here we assume that all $y_{i,j}^{(\ell)}$'s are independent for all $i, j, (i < j)$, and ℓ.

We point out that the BTL model is invariant that if we multiply ω_i^*, or increase θ^*, by a constant c, the

distribution of $y_{i,j}^{(\ell)}$ does not change. That is, θ^* and $\theta^* + c = (\theta_1^* + c, \cdots, \theta_n^* + c)^\top$ are observationally equivalent. Hence, we regard a score vector $\theta^* \in \mathbb{R}^n$ as an equivalence class $[\theta^*] = \{\theta' : \theta^* + c, c \in \mathbb{R}\}$, and regard the parameter space as the set of equivalence classes of $\mathbb{R}^n$ (Hunter 2004, Negahban et al. 2017). For the identifiability of the parameters, we impose a constraint on the parameter space $\mathcal{C}$ that we let $\mathcal{C} = \{\theta : f(\theta) = 0\}$, and propose our inferential framework under this constraint, where the function f is smooth, and ensures the identifiability of the parameter θ. Specific examples of f include $\mathbf{1}^\top\theta = 0$, $\theta_1 = 1$ (θ_1 is the preference score of the first item), among others (Chen et al. 2013, Negahban et al. 2017, Jin et al. 2020).

2.2. Ranking and Ranking Property

As discussed in Section 1, our goal is to infer some general ranking properties based on the BTL model using samples of pairwise comparisons among all items. We first provide the formal definition of the ranking and its properties.

Definition 2.1 (Ranking). Assume there are n items. Let $\mathcal{R}$ be all bijections from the set $[n]$ onto itself. Then each $\gamma \in \mathcal{R}$ is a possible rank of the n items. Let γ_i be the rank of item i in ranking γ, where $\gamma_i < \gamma_j$ if item i is ranked higher (preferred) than item j. Let $\gamma^* \in \mathcal{R}$ be the true ranking of these n items. Finally, we let $\gamma(\theta)$ be the induced ranking from the underlying preference score vector θ.

When we are interested in some ranking property of a given item, we are essentially interested in testing if the ranking satisfies some properties as we discussed in the introduction. Thus, we infer if the true ranking belongs to some set of rankings. To facilitate our discussion, we define the equivalent rankings and ranking properties with respect to a single item.

Definition 2.2 (Equivalent Rankings with Respect to a Single Item). Rankings γ and $\gamma' \in \mathcal{R}$ are equivalent with respect to item i if

$$\{j \in [n] : \gamma_j < \gamma_i\} = \{j \in [n] : \gamma'_j < \gamma'_i\},$$

or equivalently,

$$\{j \in [n] : \gamma_j > \gamma_i\} = \{j \in [n] : \gamma'_j > \gamma'_i\}.$$

Furthermore, we let the equivalent class of a ranking γ with respect to item i be $\mathcal{R}_{\gamma,i} = \{\gamma' \in \mathcal{R} : \gamma'$ is equivalent to γ with respect to item $i\}$.

Definition 2.3 (Ranking Property with Respect to a Single Item). A ranking property $\mathcal{R}_i$ with respect to a single item i is a set of rankings such that $\mathcal{R}_i \subset \mathcal{R}$, and for any ranking $\gamma \in \mathcal{R}_i$, its equivalent class $\mathcal{R}_{\gamma,i}$ satisfies $\mathcal{R}_{\gamma,i} \subseteq \mathcal{R}_i$.

Essentially, the ranking property $\mathcal{R}_i$ with respect to item i is a subset of all possible rankings $\mathcal{R}$, and a collection of disjoint equivalent classes. Specific examples of ranking property with respect to item i include Examples 1.1 and 1.2 where we are interested in testing if item i is preferred over another given item, or item i is ranked within top-K.

- Example 1.1 (Pairwise Preference Between Item i and Item j). We aim to test if item i is ranked higher than item j, which means $\gamma_i^* < \gamma_j^*$ or $\theta_i^* > \theta_j^*$. Letting

$$\mathcal{R}_i = \{\gamma : \gamma_i < \gamma_j\} = \{\gamma(\theta) : \theta_i > \theta_j\},$$

we show that $\mathcal{R}_i$ is a ranking property as defined in Definition 2.3. If $\gamma \in \mathcal{R}_i$, which means $\gamma_i < \gamma_j$, then for any ranking γ' equivalent to γ (i.e., $\gamma' \in \mathcal{R}_{\gamma,i}$), we have

$$\{k \in [n] : \gamma_k < \gamma_i\} = \{k \in [n] : \gamma'_k < \gamma'_i\}.$$

Thus, $\gamma'_i < \gamma'_j$, which means that $\gamma' \in \mathcal{R}_i$ and further gives the equivalent class $\mathcal{R}_{\gamma,i} \subseteq \mathcal{R}_i$. We have that $\mathcal{R}_i$ in this example satisfies Definition 2.3.

- Example 1.2 (Top-K Test). If item i's preference score is larger than $n - K$ items, that is, $\theta_i > \theta_{(K+1)}$, where $\theta_{(K+1)}$ denotes the $(K+1)$-th largest preference score, or equivalently, $\gamma_i \le K$. We aim to test if item i is ranked among top-K items. Thus, we have that $\mathcal{R}_i$ is

$$\mathcal{R}_i = \{\gamma : \gamma_i \le K\} = \{\gamma(\theta) : \theta_i > \theta_{(K+1)}\}.$$

To see that $\mathcal{R}_i$ is a ranking property as defined in Definition 2.3, we have that, if $\gamma \in \mathcal{R}_i$, which means $\gamma_i \le K$, then for any ranking γ' equivalent to γ (i.e., $\gamma' \in \mathcal{R}_{\gamma,i}$), and we have $\{k \in [n] : \gamma_k < \gamma_i\} = \{k \in [n] : \gamma'_k < \gamma'_i\}$. Thus, the ranking of item i does not change, and we still have $\gamma'_i \le K$, which means $\gamma' \in \mathcal{R}_i$. We have that $\mathcal{R}_i$ is a ranking property satisfying Definition 2.3.

In the next section, given a ranking property $\mathcal{R}_i$, we propose a novel approach to test whether item i satisfies this property that

$$H_0 : \gamma^* \notin \mathcal{R}_i \ \text{ v.s. } \ H_a : \gamma^* \in \mathcal{R}_i.$$

3. Inference

In this section, we propose our inferential framework to test general ranking properties. The first step in our inferential framework is a novel Lagrangian debiasing method, which handles the general constrained inference with penalization, and we apply the method to infer the latent scores. We then adopt a Gaussian multiplier bootstrap approach to test general ranking properties. We conclude this section by showing that our method controls the type I error and is asymptotically powerful.

3.1. Lagrangian Debiasing Approach

We first propose a novel Lagrangian debiased estimator of the preference scores. Our proposed method is motivated from the regularized MLE approach. Assuming the BTL model, the MLE approach (Ford 1957, Hunter 2004) provides an estimator for the latent preference scores by solving the following convex optimization problem

$$\widehat{\theta} = \underset{\theta\in\mathbb{R}^n}{\operatorname{argmin}}\, \mathcal{L}_{\lambda_0}(\theta) := \mathcal{L}(\theta) + \lambda_0 \|\theta\|_2^2, \tag{3.1}$$

where $\lambda_0 > 0$ is a tuning parameter, and the negative log-likelihood function $\mathcal{L}(\theta)$ is

$$\mathcal{L}(\theta) = \sum_{(i,j)\in\mathcal{E}, i>j} \Big\{ -y_{j,i}(\theta_i - \theta_j) + \log(1 + e^{\theta_i - \theta_j}) \Big\}, \tag{3.2}$$

where $y_{j,i} = \sum_{\ell=1}^{L} y_{j,i}^{(\ell)}/L$.

The regularization guarantees that the obtained estimator satisfying $\mathbf{1}^T\widehat{\theta} = 0$ (Chen et al. 2019), and the deduction of optimal rate of the obtained estimator relies on the strong convexity of $\mathcal{L}_{\lambda_0}(\theta)$. Different works study the theoretical guarantees of the MLE approach (Shah et al. 2015, Negahban et al. 2017, Chen et al. 2019, Wang et al. 2020). In particular, Chen et al. (2019) study the convergence rate of the estimator (3.1) in terms of ℓ_∞-norm. For self-completeness, we provide the following result.

Lemma 3.1. *Under the BTL model, suppose that* $\kappa := \frac{w_{\max}}{w_{\min}} < C$ *for some constant* $C > 0$. *If the pairwise comparison probability* p *in Erdös-Rényi graph satisfies* $p \geq \frac{C_0 \log n}{n}$ *for some sufficiently large constant* $C_0 > 0$, *and the regularization parameter* $\lambda_0 = c_{\lambda_0}\sqrt{np\log n/L}$ *for some constant* $c_{\lambda_0} > 0$, *we have that the estimator* $\widehat{\theta}$ *derived from the regularized MLE achieves the optimal rate*

$$\|\widehat{\theta} - \theta^*\|_\infty \lesssim \sqrt{\frac{\log n}{npL}}$$

with probability at least $1 - \mathcal{O}(n^{-5})$.

Remark 3.1. We point out that the same rate can also be achieved by the spectral method Negahban et al. (2017), and we provide the proof in the online appendix, Section H.1.

Because $\widehat{\theta}$ is derived from a regularized MLE, conducting inference based on $\widehat{\theta}$ is challenging. Over the last few years, the debiasing approach achieves great successes for penalized regression. For examples, Zhang and Zhang (2014), Van de Geer et al. (2014), and Javanmard and Montanari (2014a, b) study the debiasing approach based on linear or generalized linear models, and Ning and Liu (2017) provide a decorrelation approach to inferring estimators derived from penalized MLE methods.

However, these existing debiasing methods cannot be directly used in our problem. This is because that the parameter of interest is nonidentifiable in the BTL model, and the Fisher information matrix is singular. To ensure the identifiability, as discussed in Section 2.1, we impose a constraint of the parameter that we let θ belongs to the set $\mathcal{C} = \{\theta : f(\theta) = 0\}$ for some smooth function f. To handle the challenges raised by the constraint, we propose a general Lagrangian debiasing method in the next part.

3.1.1. Lagrangian Debiasing Method. We propose a general Lagrangian debiasing method for inference based on penalized MLE with constraints. Our method is motivated by Ning and Liu (2017), where the authors consider a one-step estimator by solving the first-order approximation of the score function $\nabla\mathcal{L}(\widehat{\theta}) + \nabla^2\mathcal{L}(\widehat{\theta})(\theta - \widehat{\theta}) = 0$. To handle the constraint on the parameters that $f(\theta) = 0$, we consider the Lagrangian dual function. In particular, under the constraint $f(\theta) = 0$, the MLE method aims to solve the problem that

$$\min_{\theta} \mathcal{L}(\theta), \text{ subject to } f(\theta) = 0.$$

The corresponding Lagrangian dual problem is

$$\max_{\lambda}\min_{\theta} \mathcal{L}(\theta) + \lambda f(\theta),$$

where $\lambda \in \mathbb{R}$ is the Lagrangian multiplier. Considering the first-order optimality condition of the Lagrangian dual problem, we have that an optimal dual solution pair (θ, λ) satisfies

$$\nabla\mathcal{L}(\theta) + \lambda\nabla f(\theta) = 0, \text{ and } f(\theta) = 0. \tag{3.3}$$

Based on these equations, we propose our debiasing approach. In particular, given a penalized estimator $\widehat{\theta}$ from (3.1), we obtain a debiased estimator $\widehat{\theta}^d$ by solving the following system of equations of θ and λ, which are first-order approximations of (3.3),

$$\begin{cases} \nabla\mathcal{L}(\widehat{\theta}) + \nabla^2\mathcal{L}(\widehat{\theta})(\theta - \widehat{\theta}) + \lambda\nabla f(\widehat{\theta}) = 0 \\ f(\widehat{\theta}) + \nabla f(\widehat{\theta})^\top(\theta - \widehat{\theta}) = 0, \end{cases} \tag{3.4}$$

or equivalently,

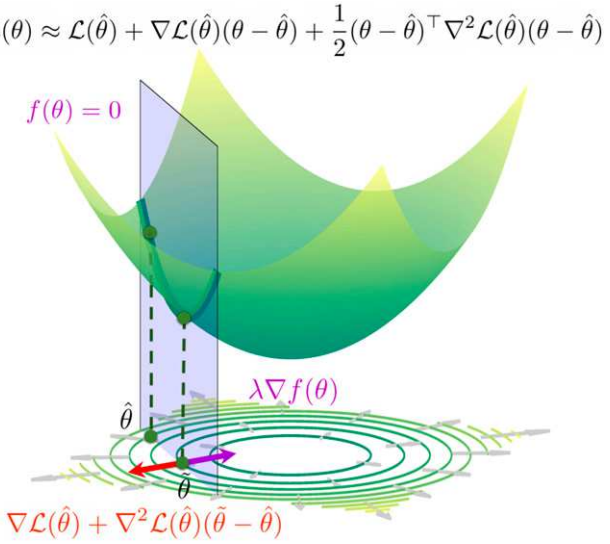
$$\begin{pmatrix} \nabla^2\mathcal{L}(\widehat{\theta}) & \nabla f(\widehat{\theta}) \\ \nabla f(\widehat{\theta})^\top & 0 \end{pmatrix} \begin{pmatrix} \theta - \widehat{\theta} \\ \lambda \end{pmatrix} = \begin{pmatrix} -\nabla\mathcal{L}(\widehat{\theta}) \\ -f(\widehat{\theta}) \end{pmatrix}. \tag{3.5}$$

See Figure 1 for illustration.

We point out that our Lagrangian debiasing method can be applied to general inference problems beyond the BTL model. In what follows, we first present the debiasing approach for inference under general constraints. Then we provide the debiasing method under the BTL model under the special case that the constraint function f is linear.

For general inferential problems under some constraint that the parameter belongs to the set $\mathcal{C} = \{\theta : f(\theta) = 0\}$,

Figure 1. (Color online) Geometric Illustration of Our Lagrangian Debiasing Method



Notes. The green surface is the approximation of the loss function $\mathcal{L}(\theta)$. The green circles are the contour lines of the surface. The transparent purple plane is the constraint $f(\theta) = 0$. The red arrow represents $\nabla\mathcal{L}(\widehat{\theta}) + \nabla^2\mathcal{L}(\widehat{\theta})(\widetilde{\theta} - \widehat{\theta})$, and the purple arrow represents $\lambda\nabla f(\theta)$.

suppose we have an initial estimator $\widehat{\theta} \in \mathcal{C}$, and let the loss function be $\mathcal{L}(\theta)$. By (3.4), if the following matrix is invertible, we define

$$\widehat{\Sigma} := \begin{pmatrix} \nabla^2\mathcal{L}(\widehat{\theta}) & \nabla f(\widehat{\theta}) \\ \nabla f(\widehat{\theta})^\top & 0 \end{pmatrix}. \tag{3.6}$$

We have that the debiased estimator $\widehat{\theta}^d$ satisfies

$$\begin{pmatrix} \widehat{\theta}^d - \widehat{\theta} \\ \lambda \end{pmatrix} = \widehat{\Sigma}^{-1} \begin{pmatrix} -\nabla\mathcal{L}(\widehat{\theta}) \\ -f(\widehat{\theta}) \end{pmatrix}. \tag{3.7}$$

When the problem is high dimensional, the matrix $\widehat{\Sigma}$ is not invertible because of the rank deficiency, and it becomes challenging to solve Problem (3.4). Motivated by (3.7), we aim to find an estimator for the inverse of the population version of $\widehat{\Sigma}$, which is

$$\Sigma^* = \begin{pmatrix} \mathbb{E}[\nabla^2\mathcal{L}(\theta^*)] & \nabla f(\theta^*) \\ \nabla f(\theta^*)^\top & 0 \end{pmatrix}.$$

We achieve this by first finding an estimator for the inverse of the population version of $\mathbb{E}[\nabla^2\mathcal{L}(\theta^*)]$, and then obtain an estimator for the inverse by block matrix inverse.

Specifically, we estimate the inverse of $\mathbb{E}[\nabla^2\mathcal{L}(\theta^*)]$ using the constrained ℓ_1-minimization for inverse matrix estimation (CLIME) method (Cai et al. 2011). Denote the estimator as $\widehat{\Omega}$. We obtain an estimator for the inverse of $\widehat{\Sigma}$ by $\widehat{\Theta} = \begin{pmatrix} \widehat{\Theta}_{11} & \widehat{\Theta}_{12} \\ \widehat{\Theta}_{12}^\top & \widehat{\Theta}_{22} \end{pmatrix}$, where

$$\widehat{\Theta}_{11} = \widehat{\Omega} - \widehat{\Omega}\nabla f(\widehat{\theta})(\nabla f(\widehat{\theta})^\top \widehat{\Omega} \nabla f(\widehat{\theta}))^{-1} \nabla f(\widehat{\theta})^\top \widehat{\Omega},$$

and

$$\widehat{\Theta}_{12} = \widehat{\Omega}\nabla f(\widehat{\theta})(\nabla f(\widehat{\theta})^\top \widehat{\Omega} \nabla f(\widehat{\theta}))^{-1}, \text{ and}$$
$$\widehat{\Theta}_{22} = -(\nabla \mathrm{f}(\widehat{\theta})^\top \widehat{\Omega} \nabla \mathrm{f}(\widehat{\theta}))^{-1}.$$

Thus, we obtain $\widehat{\theta}^d$ by plugging $\widehat{\Theta}$ into (3.7) that

$$\begin{pmatrix} \widehat{\theta}^d - \widehat{\theta} \\ \lambda \end{pmatrix} = \widehat{\Theta} \begin{pmatrix} -\nabla\mathcal{L}(\widehat{\theta}) \\ -f(\widehat{\theta}) \end{pmatrix} \tag{3.8}$$

and

$$\widehat{\theta}^d = \widehat{\theta} - \widehat{\Theta}_{11}\nabla\mathcal{L}(\widehat{\theta}). \tag{3.9}$$

Before presenting the asymptotic properties of $\widehat{\theta}^d$, we first impose some assumptions. We point out that here we purposely do not specify the convergence rates in the following assumptions because our proposed method is a general framework, and as long as the assumptions for Theorem 3.1 are satisfied, the Lagrangian debiased method achieves the asymptotic normality. We also point out that, under our scaling assumptions in the following theorems, the assumptions are indeed satisfied.

Assumption 3.1 (Consistency for Initial Estimation of Parameters). *For some rate r_1 that depends on the sample size and parameter dimension, we assume $\|\widehat{\theta} - \theta^*\|_\infty \le r_1$.*

Assumption 3.2 (Condition on Loss Function). *For some rate r_2 and constant L_1, if $\theta = \theta^* + t(\widehat{\theta} - \theta^*)$ for $t \in [0,1]$, it holds that*

$$\|\nabla\mathcal{L}(\theta^*)\|_\infty \le r_2,\ \|\nabla^2\mathcal{L}(\theta) - \nabla^2\mathcal{L}(\theta^*)\|_\infty \le L_1 \|\theta - \theta^*\|_\infty.$$

Assumption 3.3 (Condition on Constraint Function). *For some constants c_1 and L_2, if $\theta = \theta^* + t(\widehat{\theta} - \theta^*)$ for $t \in [0,1]$, it holds that*

$$\|\nabla f(\theta^*)\|_\infty \le c_1,\ \|\nabla f(\theta) - \nabla f(\theta^*)\|_\infty \le L_2 \|\theta - \theta^*\|_\infty.$$

Assumption 3.4. *For some rates r_3, r_4, r_5 and constants c_2, c_3, we assume that*

$$\|I - \widehat{\Omega}\nabla^2\mathcal{L}(\widehat{\theta})\|_\infty \le r_3,\ \|\widehat{\Omega} - \Omega^*\|_\infty \le r_4,$$
$$\|\Omega^*\|_\infty \le c_2,\ \nabla f(\theta^*)^\top \Omega^* \nabla f(\theta^*) \ge c_3,\ \|\Theta^*\|_\infty \le r_5.$$

Assumption 3.5 (Central Limit Theorem (CLT) of the Score Function). *For every $i \ne j$, if $(\Theta^*_{11}\Sigma^*_{11}\Theta^{*\top}_{11})_{jj} \ge C$ and $(\mathbf{e}_i - \mathbf{e}_j)^\top (\Theta^*_{11}\Sigma^*_{11}\Theta^{*\top}_{11})(\mathbf{e}_i - \mathbf{e}_j) \ge C$ for some constant $C > 0$, it holds that*

$$\frac{\sqrt{n}[\Theta^*_{11}\nabla\mathcal{L}(\theta^*)]_j}{\sqrt{[\Theta^*_{11}\Sigma^*_{11}\Theta^{*\top}_{11}]_{jj}}} \xrightarrow{d} N(0,1)$$

and

$$\sqrt{n}\frac{[\Theta^*_{11}\nabla\mathcal{L}(\theta^*)]_i-[\Theta^*_{11}\nabla\mathcal{L}(\theta^*)]_j}{\sqrt{(e_i-e_j)^T(\Theta^*_{11}\Sigma^*_{11}\Theta^{*\top}_{11})(e_i-e_j)}}\xrightarrow{d}N(0,1),$$

where $[\Theta^*_{11}\nabla\mathcal{L}(\theta^*)]_j$ *is the jth entry of* $\Theta^*_{11}\nabla\mathcal{L}(\theta^*)$, $[\Theta^*_{11}\Sigma^*_{11}\Theta^{*\top}_{11}]_{jj}$ *is the jth diagonal element of matrix* $\Theta^*_{11}\Sigma^*_{11}\Theta^{*\top}_{11}$, *and* Σ^*_{11}, Θ^*_{11} *is the upper left* $n\times n$ *block of* Σ^*, Θ^*, *respectively.*

By these assumptions, the following two corollaries hold, which are crucial for later proofs. Proofs of Corollaries 3.1 and 3.2 can be found in the online appendix, Section B.1 and B.2.

Corollary 3.1. *Under Assumptions 3.1–3.4, we have* $\|\widehat{\Theta}-\Theta^*\|_\infty\le r_1+r_4$.

Corollary 3.2. *Under Assumptions 3.1–3.4, we have* $\|I-\widehat{\Theta}\widehat{\Sigma}\|_\infty\le r_3$.

We then present the asymptotic distribution of the Lagrangian debiased estimator.

Theorem 3.1. *Under Assumptions 3.1–3.5, if* $c_1,c_2,c_3,L_1,L_2=\mathcal{O}(1)$ *and* $\sqrt{n}(r_1^2r_5+(r_1+r_4)(r_1^2+r_2)+r_1r_3)=o(1)$, *we have the following asymptotic distribution for the Lagrangian debiased estimator (3.9):*

$$\sqrt{n}\frac{\widehat{\theta}^d_j-\theta^*_j}{\sqrt{[\Theta^*_{11}\Sigma^*_{11}\Theta^{*\top}_{11}]_{jj}}}\xrightarrow{d}N(0,1),$$

and

$$\sqrt{n}\frac{\left(\widehat{\theta}^d_i-\theta^*_i\right)-\left(\widehat{\theta}^d_j-\theta^*_j\right)}{\sqrt{(e_i-e_j)^\top\Theta^*_{11}\Sigma^*_{11}\Theta^{*\top}_{11}(e_i-e_j)}}\xrightarrow{d}N(0,1),$$

where e_k *is the natural basis with the kth entry be one and other entries be zero.*

Proof. See the online appendix, Section A.1, for the detailed proof. □

Remark 3.2. If the constraint function f is linear, it is not difficult to say that $\widehat{\theta}^d$ satisfies the constraint. Meanwhile, under the general constraint, as the problem is nonconvex, $\widehat{\theta}^d$ may violate the constraint. However, even if the constraint is violated, the previous asymptotic results still hold.

Remark 3.3. In the Lagrangian debiasing procedure, we do not explicitly assume the loss functions and the constraints on parameters to be convex. We provide more discussions here. First, in our Assumption 3.1, we assume that the convergence rate for initial estimator $\widehat{\theta}$ satisfies $\|\widehat{\theta}-\theta^*\|_\infty\le r_1$. Recall that r_1 needs to be sufficiently small for the proposed method to work. Obtaining a sufficiently "good" initial estimator implicitly assumes some nice properties of the loss function $\mathcal{L}(\cdot)$ (Casella and Berger 2021), which is usually convex.

Second, in our debiasing step, as long as the initial estimator is good enough, we do not need convexity, because we approximate $\nabla\mathcal{L}(\theta)$ by $\nabla\mathcal{L}(\widehat{\theta})+\nabla^2\mathcal{L}(\widehat{\theta})\ (\theta-\widehat{\theta})$ and $f(\theta)$ by $f(\widehat{\theta})+\nabla f(\widehat{\theta})^\top(\theta-\widehat{\theta})$ when Assumption 3.1 (the initial estimator $\widehat{\theta}$ converges to true parameter θ^* sufficiently rapidly), Assumption 3.2 ($\nabla\mathcal{L}(\cdot)$ is smooth around θ^*), and Assumption 3.3 ($f(\cdot)$ is smooth around θ^*) are satisfied.

3.1.2. Lagrangian Debiasing for BTL Model. We present the debiasing method under the BTL model. In particular, for ranking problems, letting the constraint function be linear that $\mathbf{1}^\top\theta=0$ as in Chen et al. (2013), Negahban et al. (2017), and Jin et al. (2020), we define

$$\begin{pmatrix}\widehat{\Theta}_{11} & \frac{1}{n}\mathbf{1}\\ \frac{1}{n}\mathbf{1}^\top & 0\end{pmatrix}=\begin{pmatrix}\nabla^2\mathcal{L}(\widehat{\theta}) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}\text{ and }\begin{pmatrix}\Theta^*_{11} & \frac{1}{n}\mathbf{1}\\ \frac{1}{n}\mathbf{1}^\top & 0\end{pmatrix}=\begin{pmatrix}\nabla^2\mathcal{L}(\theta^*) & \mathbf{1}\\ \mathbf{1}^\top & 0\end{pmatrix}^{-1}.\tag{3.10}$$

Here the invertibility is provided in Remark H.2 in the online appendix, and the form of inverse is validated in Corollary H.1.

The next theorem shows that under mild scaling conditions, the Lagrangian debiasing estimator $\widehat{\theta}^d_j$ and the component-wise difference $\widehat{\theta}^d{}_i-\widehat{\theta}^d{}_j$ for any i and j $(1\le i,j\le n)$ are asymptotically normal with mean θ^*_j and $\theta^*_i-\theta^*_j$, respectively.

Theorem 3.2. *Considering the BTL model, under constraint parameter set* $\mathcal{C}=\{\theta:\mathbf{1}^\top\theta=0\}$, *if* $p\ge\frac{C_0\log n}{n}$ *for some sufficiently large constant* $C_0>0$ *and* $\frac{n\log n}{\sqrt{L}}+\frac{\log n}{\sqrt{pL}}=o(1)$, *we have that the Lagrangian debiasing estimator satisfies that, as* $n,L\to\infty$,

$$\sqrt{L}\frac{\widehat{\theta}^d_j-\theta^*_j}{\sqrt{[\Theta^*_{11}]_{jj}}}\xrightarrow{d}N(0,1),$$

and

$$\sqrt{L}\frac{\widehat{\theta}^d_i-\widehat{\theta}^d_j-(\theta^*_i-\theta^*_j)}{\sqrt{(e_i-e_j)^\top\Theta^*_{11}(e_i-e_j)}}\xrightarrow{d}N(0,1).$$

Proof. See the online appendix, Section A.2, for the detailed proof. □

Remark 3.4. By Corollary H.1 in the online appendix, we have that $[\Theta^*_{11}]_{jj}\asymp\frac{1}{np}$ for all $1\le j\le n$. Consequently,

for all $j \in [n]$, we have

$$\left|\widehat{\theta}_j^d - \theta_j^*\right| \lesssim \sqrt{\frac{1}{npL}}$$

with a probability that goes to one. This matches the ℓ_∞-norm error achieved by the spectral and regularized MLE methods as analyzed in Chen et al. (2019).

The asymptotic normality of $\widehat{\theta}^d{}_i - \widehat{\theta}^d{}_j$ is the fundamental building block for inferring the pairwise preference such as in Example 1.1, where we test if item i is preferred over item j. Basically, it is a "local" test that only involves two items, which can be done with the asymptotic distribution of $\widehat{\theta}_i^d - \widehat{\theta}_j^d$. For the more challenging "global" testing problems, such as Example 1.2, where we test if a given item is among the top-K ranked items, we need to uniformly control the quantile of maximal statistic, which will be discussed in the next section.

3.2. Hypothesis Testing

In this section, we propose our general inferential framework for ranking problems. As mentioned in Section 1, we first test whether a given item i satisfies some property that

$$H_0 : \gamma^* \notin \mathcal{R}_i \quad \text{v.s.} \quad H_a : \gamma^* \in \mathcal{R}_i.$$

To facilitate our discussion, we define the legal pair and a distance between the null and alternative, which essentially measure the signal strengths in our testing problems. In particular, when we test some property of item i, we say a pair (i, i') is a legal pair if the property is true, and if we swap the scores of item i and item i', the property no longer holds. That is, swapping the scores of item i and item i' changes the property of our interest. The distance between the null and alternative is thus defined as the minimal difference of scores among all legal pairs.

Definition 3.1 (Legal Pair). Suppose $\gamma \in \mathcal{R}_i$. After swapping the scores of item i and item i', we obtain a new rank γ'. We say that the pair of items (i, i') is legal if $\gamma' \notin \mathcal{R}_i$.

Definition 3.2 (Distance Between the Null and Alternative). We define the distance between the null and alternative as

$$\Delta(\theta, \mathcal{R}_i) = \min_{i':(i,i') \text{ is legal}} |\theta_i - \theta_{i'}| .$$

Then, we provide the specific legal pairs and distances $\Delta(\theta, \mathcal{R}_i)$ for Examples 1.1 and 1.2.

- Example 1.1 (Pairwise Preference Between Item i and Item j). We aim to test if item i is ranked higher than item j, which means $\gamma_i^* < \gamma_j^*$ or $\theta_i^* > \theta_j^*$. Let the ranking property be

$$\mathcal{R}_i = \{\gamma : \gamma_i < \gamma_j\} = \{\gamma(\theta) : \theta_i > \theta_j\}.$$

If $\gamma \in \mathcal{R}_i$ (i.e., $\theta_i > \theta_j$) and we swap scores of item i and item i' where $\theta_{i'} \le \theta_j$, the new rank does not satisfy $\mathcal{R}_i$. Meanwhile, if we swap scores of item i and item i'' where $\theta_{i''} > \theta_j$, the new rank still satisfies $\mathcal{R}_i$. So (i, i') is a legal pair if $\theta_{i'} \le \theta_j$. See Figure 2 for illustration. This observation leads to the distance

$$\Delta(\theta, \mathcal{R}_i) = \min_{i':\theta_{i'} \le \theta_j} |\theta_i - \theta_{i'}| = |\theta_i - \theta_j|.$$

Equivalently, we can test on preference scores instead of ranking, that is, testing whether item i has a larger score than item j,

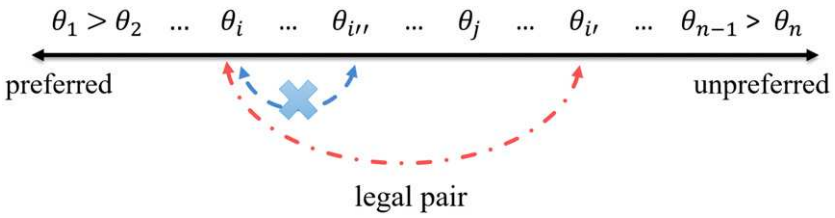
$$H_0 : \theta_i^* \le \theta_j^* \quad \text{v.s.} \quad H_a : \theta_i^* > \theta_j^*.$$

- Example 1.2: (Top-K Test). If item i's preference score is larger than $n - K$ items, that is, $\theta_i > \theta_{(K+1)}$, where $\theta_{(K+1)}$ denotes the $(K+1)$ th largest preference score, or equivalently, $\gamma_i \le K$, we say that item i is ranked among top-K items. We aim to test if item i is ranked among top-K items. Thus, we have that $\mathcal{R}_i$ is

$$\mathcal{R}_i = \{\gamma : \gamma_i \le K\} = \{\gamma(\theta) : \theta_i > \theta_{(K+1)}\}.$$

If $\gamma \in \mathcal{R}_i$ (i.e., $\theta_i > \theta_{(K+1)}$), and we swap scores of item i and item j where $\theta_j \le \theta_{(K+1)}$, the new rank does not satisfy $\mathcal{R}_i$. Meanwhile, if we swap scores of item i and item j where $\theta_j > \theta_{(K+1)}$, we have that the new rank satisfies $\mathcal{R}_i$. Thus, (i, j) is a legal pair if $\theta_j \le \theta_{(K+1)}$, and

Figure 2. (Color online) Illustration of Example 1.1



Notes. Assuming $\theta_1 > \theta_2 > \cdots > \theta_n$. Here $\gamma \in \mathcal{R}_i$ since $\theta_i > \theta_j$. If we swap scores of item i and item i', where $\theta_{i'} \le \theta_j$, the new rank does not satisfy $\mathcal{R}_i$. Meanwhile, if we swap scores of item i and item i'', where $\theta_{i''} > \theta_j$, the new rank still satisfies $\mathcal{R}_i$. Thus, (i, i') is a legal pair if $\theta_{i'} \le \theta_j$.

the distance is

$$\Delta(\theta, \mathcal{R}_i) = \min_{j:\theta_j \le \theta_{(K+1)}} |\theta_i - \theta_j| = |\theta_i - \theta_{(K+1)}|.$$

Similarly, we can transform the test on ranking into testing on preference score, our test is

$$H_0 : \theta_i^* - \theta_{(K+1)}^* \le 0 \quad \text{v.s.} \quad H_a : \theta_i^* - \theta_{(K+1)}^* > 0.$$

Next, we explain our testing procedure with the above two examples. Consider Example 1.1, where we test whether item i is ranked higher than item j, or equivalently, we test whether item i has a larger score than item j,

$$H_0 : \theta_i^* \le \theta_j^* \quad \text{v.s.} \quad H_a : \theta_i^* > \theta_j^*.$$

For this example, by Theorem 3.2, if the assumptions are satisfied, we have

$$\sqrt{L}\frac{\widehat{\theta}^d{}_i - \widehat{\theta}^d{}_j - (\theta_i^* - \theta_j^*)}{\sqrt{(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \Theta_{11}^* (\boldsymbol{e}_i - \boldsymbol{e}_j)}} \xrightarrow{d} N(0,1).$$

We thus reject H_0 if

$$\sqrt{L}\frac{\widehat{\theta}_i^d - \widehat{\theta}_j^d}{\sqrt{(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \widehat{\Theta}_{11} (\boldsymbol{e}_i - \boldsymbol{e}_j)}} > \Phi(1-\alpha),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

This example only involves two items, which is a relatively simple local test. However, for more general testing problems, we need to consider more than two items. For instance, in Example 1.2, we test if item i is ranked among the top-K items that

$$H_0 : \theta_i^* - \theta_{(K+1)}^* \le 0 \quad \text{v.s.} \quad H_a : \theta_i^* - \theta_{(K+1)}^* > 0,$$

where $\theta_{(K+1)}^*$ is the $(K+1)$ th largest score in terms of order statistic. If the score of item i is larger than the $(K+1)$ th largest score, we have that item i is ranked among top-K items. In this example and more general problems, we need to study the maximal statistic $\max_{j\neq i}(\widehat{\theta}_i^d - \theta_i^* - \widehat{\theta}_j^d + \theta_j^*)$. In what follows, we demonstrate that we can estimate the quantiles of this maximal statistic via the Gaussian multiplier bootstrap.

3.2.1. Gaussian Multiplier Bootstrap. We start from a general fixed edge set $E \subseteq \mathcal{V} \times \mathcal{V}$. The goal is to control the tail probability of the statistic that

$$\begin{aligned} T &:= \max_{(i,j)\in E} \sqrt{npL}\left(\widehat{\theta}_i^d - \theta_i^* - \widehat{\theta}_j^d + \theta_j^*\right) \\ &= -\max_{(i,j)\in E} \sqrt{\frac{1}{npL}} \sum_{\ell=1}^{L} \sum_{k>m} \mathcal{E}_{mk}\left(-y_{mk}^{(\ell)} + \frac{e^{\theta_k^*}}{e^{\theta_k^*} + e^{\theta_m^*}}\right) np \\ &\quad \times ([\Theta_{11}^*]_i - [\Theta_{11}^*]_j)(\boldsymbol{e}_k - \boldsymbol{e}_m) + \sqrt{npL}(r_i - r_j) \\ &:= \max_{(i,j)\in E} \sqrt{\frac{1}{L}} \sum_{\ell=1}^{L} x_{ij}^{(\ell)} + \sqrt{npL}(r_i - r_j), \end{aligned} \tag{3.11}$$

where $\boldsymbol{e}_k$ is the natural basis, and $[\Theta_{11}^*]_i$ is the ith row of matrix Θ_{11}^* defined in (3.10). The second equality comes from (A.13) and (A.5); $x_{ij}^{(\ell)}$ is defined as

$$\begin{aligned} x_{ij}^{(\ell)} &:= -\sqrt{np} \sum_{k>m} \mathcal{E}_{mk}\left(-y_{mk}^{(\ell)} + \frac{e^{\theta_k^*}}{e^{\theta_k^*} + e^{\theta_m^*}}\right) \\ &\quad \times ([\Theta_{11}^*]_i - [\Theta_{11}^*]_j)(\boldsymbol{e}_k - \boldsymbol{e}_m). \end{aligned} \tag{3.12}$$

which is an independent zero-mean random variable in $\mathbb{R}$ for $\ell = 1, \cdots, L$. Here $\mathcal{E}_{mk} = 1$ if $(m,k) \in \mathcal{E}$ the comparison graph, and $\mathcal{E}_{mk} = 0$ otherwise.

To estimate the quantile of T, we consider the Gaussian multiplier bootstrap in Chernozhukov et al. (2013). The main idea is to approximate the distribution of the maximum of a sum of independent random vectors with unknown covariance by the distribution of the maximum of a sum of conditional Gaussian random vectors, which is obtained by multiplying the original vectors with independently and identically distributed normal random variables. In our case, even though vector θ^* is not observable, we have some estimators $\widehat{\theta}$ are available, and we use the estimators to approximate θ^* in the bootstrap.

Hence, we define the following statistic from Gaussian multiple bootstrap

$$\begin{aligned} W &:= \max_{(i,j)\in E} \sqrt{\frac{1}{L}} \sum_{\ell=1}^{L} \left\{ -\sqrt{np} \sum_{k>m} \mathcal{E}_{mk}\left(-y_{mk}^{(\ell)} + \frac{e^{\widehat{\theta}_k}}{e^{\widehat{\theta}_k} + e^{\widehat{\theta}_m}}\right) \right. \\ &\quad \left. \times ([\widehat{\Theta}_{11}]_i - [\widehat{\Theta}_{11}]_j)(\boldsymbol{e}_k - \boldsymbol{e}_m) \right\} z_\ell, \end{aligned} \tag{3.13}$$

where z_ℓ, $\ell = 1, \cdots, L$, are i.i.d standard normal random variables.

We then estimate the conditional quantile of W given data $\boldsymbol{y} = \{y_{mk}^{(\ell)}\}_{k>m}^{\ell=1,\cdots,L}$ by

$$c_W(\alpha, E) = \inf\{t \in \mathbb{R} : \mathbb{P}(W > t \mid \boldsymbol{y}) \le \alpha\}. \tag{3.14}$$

The next theorem uniformly controls the tail probability of T by $c_W(\alpha, E)$.

Theorem 3.3. *Considering the BTL model, for any edge set $E \subseteq \mathcal{V} \times \mathcal{V}$, if $n^2 p \frac{(\log(nL))^7}{L} = o(1)$ and $\frac{n(\log n)^{3/2}}{\sqrt{L}} = o(1)$, we have*

$$\sup_{\alpha\in(0,1)} |\mathbb{P}(T > c_W(\alpha, E)) - \alpha| \to 0.$$

as $n, L \to \infty$,

Proof. We provide the proof in the online appendix, Section D. □

This theorem shows that $c_W(\alpha, E)$ obtained from the Gaussian multiplier bootstrap is a valid quantile estimator for $T = \max_{i,j\in E}\sqrt{npL}(\widehat{\theta}_i^d - \theta_i^* - \widehat{\theta}_j^d + \theta_j^*)$. In this theorem, the first scaling condition $n^2 p \frac{(\log(nL))^7}{L} = o(1)$ is from the Gaussian approximation for the maximum of a sum of

random vectors, and the second scaling condition $\frac{n(\log n)^{3/2}}{\sqrt{L}} = o(1)$ is from approximating T and W by their leading terms. Given this statistic, we are ready to present the procedure for testing general ranking properties.

3.2.2. General Testing Procedure. For general ranking property test with respect to item i

$$H_0 : \gamma^* \notin \mathcal{R}_i \quad \text{v.s.} \quad H_a : \gamma^* \in \mathcal{R}_i,$$

we perturb the preference score of every item (i.e., $\widehat{\theta}_i^d$) up to α-quantile of $\max_{j\neq i}(\widehat{\theta}_i^d - \theta_i^* - \widehat{\theta}_j^d + \theta_j^*)$, and conduct the test. Specifically, let $\widetilde{\Theta}$ be the set of all possible score vectors after perturbation that

$$\widetilde{\Theta} = \Big\{\theta : \theta_k \in \Big[\widehat{\theta}_k^d - c_W(\alpha,i)/\sqrt{npL}, \widehat{\theta}_k^d + c_W(\alpha,i)/\sqrt{npL}\Big], \\ 1 \le k \le n\Big\}, \tag{3.15}$$

where $c_W(\alpha,i) = c_W(\alpha,\{i\}\times\{j : j\neq i\})$. We reject the null hypothesis if $\gamma(\theta)\in\mathcal{R}_i$ for any $\theta\in\widetilde{\Theta}$, that is, the event $\cap_{\theta\in\widetilde{\Theta}}\{\gamma(\theta)\in\mathcal{R}_i\}$ holds.

Remark 3.5. We point out that we can simplify this general procedure for specific problems. For instance, when we test if item i is ranked within the top-K, we only need to consider the extreme point of the perturbation, where for $k = 1,\dots,n$, its kth entry is defined as

$$\theta_k = \begin{cases} \widehat{\theta}_k^d & \widehat{\theta}_k^d > \widehat{\theta}_i^d, \\ \widehat{\theta}_k^d - c_W(\alpha,i)/\sqrt{npL} & k = i, \\ \widehat{\theta}_k^d + c_W(\alpha,i)/\sqrt{npL} & \widehat{\theta}_k^d < \widehat{\theta}_i^d. \end{cases}$$

In fact, we can further simplify this procedure that we only consider θ where its kth entry is defined as

$$\theta_k = \begin{cases} \widehat{\theta}_k^d & \widehat{\theta}_k^d > \widehat{\theta}_i^d, \\ \widehat{\theta}_k^d - c_W(\alpha,i)/\sqrt{npL} & k = i, \\ \widehat{\theta}_k^d & \widehat{\theta}_k^d < \widehat{\theta}_i^d. \end{cases}$$

We justify this procedure in the online appendix, Section E.2.

We conclude this section by the following theorem that we show that the proposed procedure controls the type I error, and we provide the power analysis.

Theorem 3.4. *Under same assumptions as in Theorem 3.2 and 3.3., we have the general testing procedure satisfies that, as* $n, L \to \infty$,

$$\sup_{\gamma^*\notin\mathcal{R}_i} \mathbb{P}_0(\textit{Reject } H_0) \le \alpha,$$

and we have

$$\inf_{\gamma^*\in\mathcal{R}_i,\, \Delta(\theta^*,\mathcal{R}_i)>\delta} \mathbb{P}(\textit{Reject } H_0) \to 1,$$

where $\delta = C\sqrt{\frac{\log n}{npL}}$ *for some constant* C.

Proof. We provide the proof in the online appendix, Section E. □

4. Multiple Testing

In this section, we extend the proposed procedure to the multiple testing setting. As discussed in the introduction, the multiple testing finds important applications in our ranking inference problems such as Example 1.3, where we aim to infer the set of all top-K ranked items. In general multiple testing problems, we aim to control the family-wise type I error rate (FWER), or the false discovery rate (FDR), while achieving certain power. Widely used procedures include Bonferroni correction, Dunn-Šidák procedure (Šidák 1967), Holm procedure (Holm 1979) for controlling FWER, and Benjamini-Hochberg procedure (BH) (Benjamini and Hochberg 1995) and Benjamini-Yekutieli procedure (Benjamini and Yekutieli 2001) for controlling FDR. In addition, there are also resampling based procedures such as permutation testing and bootstrap method (Westfall and Young 1993, Ge et al. 2003). However, these methods cannot be directly applied to our problems, as their theoretical properties cannot be easily justified. This is mainly because that the test statistics for the hypotheses are clearly dependent, which makes our multiple testing problems challenging.

In general, we aim to test the following hypotheses simultaneously,

$$\begin{aligned} H_{0i} :&\ \text{item } i \text{ does not satisfy property } \mathcal{R}_i \quad \text{v.s. } H_{ai} \\ :&\ \text{item } i \text{ satisfies } \mathcal{R}_i, \text{ for } i \in [n]. \end{aligned} \tag{4.1}$$

4.1. Control FWER

We first present our procedure for controlling the FWER. Recall that the FWER is the probability of making at least one type I error that

$$\text{FWER} = \mathbb{P}(\#\text{ false positives} > 0).$$

Specifically, when we aim to control the FWER in multiple testing, we let the maximal statistic be

$$M = \max_{i\in[n]} \max_{j\in[n]} \sqrt{npL}(\widehat{\theta}_i^d - \theta_i^* - \widehat{\theta}_j^d + \theta_j^*),$$

and we estimate its $(1-\alpha)$ th percentile $C_M(\alpha,[n]\times[n])$ by Gaussian multiplier bootstrap and taking the edge set E as $[n]\times[n]$ in (3.11). Next, we reject the H_{0i} in (4.1) if item i satisfies property $\mathcal{R}_i$ for all possible perturbation for the debiased estimator $\widehat{\theta}^d$ up to $C_M(\alpha,[n]\times[n])/\sqrt{npL}$ entrywise. Equivalently, letting $\widetilde{\Theta}$ be the set of possible latent scores after perturbation that

$$\widetilde{\Theta} = \Big\{\theta : \theta_k \in \Big[\widehat{\theta}_k^d - C_M(\alpha,[n]\times[n])/\sqrt{npL}, \widehat{\theta}_k^d \\ + C_M(\alpha,[n]\times[n])/\sqrt{npL}\Big], 1 \le k \le n\Big\},$$

we reject H_{0i} in (4.1) if for any $\theta\in\widetilde{\Theta}$, we have $\gamma(\theta)\in\mathcal{R}_i$, that is, the event $\cap_{\theta\in\widetilde{\Theta}}\{\gamma(\theta)\in\mathcal{R}_i\}$ holds.

To facilitate our power analysis, we define the signal strength for multiple testing problems as

$$\Delta(\theta) = \min_{i:\gamma(\theta)\in\mathcal{R}_i} \Delta(\theta, \mathcal{R}_i) = \min_{i:\gamma(\theta)\in\mathcal{R}_i} \min_{i':(i,i') \text{ is legal}} |\theta_i - \theta_{i'}|. \tag{4.2}$$

In what follows, we use two examples to illustrate some insights of this signal strength.

• Consider the problem of selecting all top-K ranked items. Denote the ith largest order statistic as $\theta_{(i)}$. For any given $1 \le i \le n$ satisfying $\theta_i \ge \theta_{(K)}$, the pairs of items (i, i') such that $\theta_{i'} \le \theta_{(K+1)}$ are the legal pairs, and the smallest distance is $|\theta_i - \theta_{(K+1)}|$. Hence, we have

$$\Delta(\theta) = \min_{\theta_i \ge \theta_{(K)}} \min_{\theta_{i'} \le \theta_{(K+1)}} |\theta_i - \theta_{i'}| = \min_{\theta_i \ge \theta_{(K)}} |\theta_i - \theta_{(K+1)}| = |\theta_{(K)} - \theta_{(K+1)}|.$$

This is consistent with our intuition that when we aim to choose the top K items, the gap between the scores of $\theta_{(K)}$ and $\theta_{(K+1)}$ somewhat determines this problem's difficulty.

• Consider the problem of selecting all items ranked higher than item k with score θ_k. For any item i satisfying $\theta_i > \theta_k$, the pairs (i, i') where $\theta_{i'} \le \theta_k$ are legal pairs. Hence, we have that the distance is

$$\Delta(\theta) = \min_{\theta_i > \theta_k} \min_{\theta_i' \le \theta_k} |\theta_i - \theta_{i'}| = \min_{\theta_i > \theta_k} |\theta_i - \theta_k|.$$

The following theorem shows that the FWER based on our procedure above is guaranteed to be no greater than α asymptotically and is powerful.

Theorem 4.1 (Familywise Type I Error Rate). *Under the same assumptions for Theorems 3.2 and 3.3, following the previous multiple testing procedure, for any $0 < \alpha < 0.5$, we have*

$$\text{FWER} = \mathbb{P}(\textit{Making at least one Type I error}) \le \alpha + o(1).$$

Furthermore, if $\Delta(\theta^*) \gtrsim \sqrt{\frac{\log n}{npL}}$ holds, we have

$$\mathbb{P}(\textit{Making at least one Type II error}) \to 0.$$

Proof. We provide the proof in the online appendix, Section F.1. □

This theorem shows that our method controls the FWER asymptotically for the given level. Meanwhile, our procedure is asymptotically powerful if $\Delta(\theta^*) \gtrsim \sqrt{\frac{\log n}{npL}}$, which matches the lower bound we derive in Section 5.

4.2. FDR Control

We then consider the problem of controlling the FDR. FDR is the expected proportion of type I errors among all discoveries (Benjamini and Hochberg 1995), and our goal is to control the FDR under some prespecified level α that

$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{false positives}}{\#\text{discoveries}} \mathbb{I}[\#\text{ discoveries} > 0]\right] \le \alpha.$$

Consider the multiple testing problem of interest (4.1). For each hypothesis of item i, we perform our proposed single testing procedure in Section 3.2 and get the p value p_i that

$$p_i = \inf\{\alpha_0 : \cap_{\theta \in \widetilde{\Theta}(\alpha_0)} \{\gamma(\theta) \in \mathcal{R}_i\}\} \quad \text{for } 1 \le i \le n,$$

where

$$\widetilde{\Theta}(\alpha_0) = \left\{\theta : \theta_k \in \left[\widehat{\theta}_k^d - c_W(\alpha_0, i)/\sqrt{npL},\ \widehat{\theta}_k^d + c_W(\alpha_0, i)/\sqrt{npL}\right], 1 \le k \le n\right\}.$$

Because the p values p_is for different tests have complicated dependency, we consider the Benjamini-Yekutieli procedure by Benjamini and Yekutieli (2001) to control the FDR, which ranks the hypotheses according to their corresponding p values and chooses a cutoff to control the FDR. In particular, we order the n p values in the ascending order as $p_{(1)}, \cdots, p_{(n)}$ and reject the null hypothesis for all $H_{(i)}$, $i = 1, \ldots, r$, where

$$r = \max_k \left\{k : p_{(k)} \le \frac{k}{n \cdot N} \cdot \alpha\right\} \text{ and } N = \sum_{k=1}^{n} \frac{1}{k}.$$

The following theorem shows that the FDR based on our procedure achieves the desired FDR level asymptotically if $|\mathcal{H}_0|\left(\frac{1}{L^{c_3}} + \frac{2}{n^5}\right) = o(1)$, where $|\mathcal{H}_0|$ is the number of true null hypotheses.

Theorem 4.2 (FDR Control). *Suppose that the conditions in Theorems 3.2 and 3.3 hold. Following the previous multiple testing procedure, for any $0 < \alpha < 1$, we have*

$$\text{FDR} \le \frac{|\mathcal{H}_0|}{n} \cdot \alpha + C\,|\mathcal{H}_0|\left(\frac{1}{L^{c_3}} + \frac{2}{n^5}\right),$$

for some constant $C > 0$, and the constant c_3 satisfies the condition in Remark D.1.

Proof. We provide the proof in the online appendix, Section F.2. □

Remark 4.1. We point out that, as show in the previous theorem and in our simulation studies, our developed Benjamini-Yekutieli–based FDR controlling procedure is relatively conservative when the number of true nulls $|\mathcal{H}_0|$ is small compared with total number of items n. The similar conservative result also holds in Eisenach et al. (2020).

5. Lower Bound Theory

In this section, we derive a lower bound for the multiple testing problem (4.1). To the best of our knowledge, even beyond ranking problems, all existing works focus on discussing the lower bound for single hypothesis testing problems. To facilitate our discussion, we first define a novel minimax risk of multiple testing problems. In particular, we let the risk be the probability of making at least one type I or type II error that

$$\mathfrak{R} = \inf_{\psi} \sup_{\theta^* \in \Xi} \mathbb{P}(\#\text{false positives} + \#\text{false negatives} \geq 1), \tag{5.1}$$

where ψ is any selection procedure giving a vector $\psi \in \{0,1\}^n$ that $\psi_i = 1$ means that we reject H_{0i} in (4.1), and Ξ is our parameter space, which is closed under swapping scores of any two items. If $\mathfrak{R} \geq 1-\epsilon$, where $\epsilon > 0$ is a constant, we say that any procedure fails in the sense that type I or type II error cannot be controlled. Given this setup, we aim to characterize necessary conditions under which we can control the minimax risk to the desired level.

5.1. Lower Bound Theorem

Recall that we define the legal pair in Definition 3.1 and the distance $\Delta(\theta)$ in (4.2). We then define a critical subset of all legal pairs, which is crucial in deriving the information-theoretic lower bound and captures the "hardness" of multiple testing for different properties.

Definition 5.1 (Divider Set). A set $\mathcal{M}(\theta) \subseteq [n] \times [n]$ is a divider set if $\mathcal{M}(\theta)$ satisfies the following two conditions:

- For any $(i,i') \in \mathcal{M}(\theta)$, it satisfies $\gamma \in \mathcal{R}_i$, (i,i') is legal, and $|\theta_i - \theta_{i'}| = \Delta(\theta)$.
- For any two pairs $(i_1,i_1'),(i_2,i_2') \in \mathcal{M}(\theta)$, letting γ_1 be the scores by swapping items i_1 and i_1', and γ_2 be the scores by swapping items of i_2 and i_2'. They must satisfy $\{i : \gamma_1 \in \mathcal{R}_i\} \neq \{i : \gamma_2 \in \mathcal{R}_i\}$.

The following theorem derives necessary signal strengths in $\Delta(\theta)$ and $|\mathcal{M}(\theta)|$ for controlling the risk $\mathfrak{R} \leq 1-\epsilon$.

Theorem 5.1 (Necessary Signal Strength). *Considering the BTL model where the pairwise comparison probability p in Erdös-Rényi graph satisfies $p \gtrsim \frac{\log n}{n}$, if there exists $\theta \in \Xi$ such that*

$$\Delta(\theta) \lesssim \sqrt{\frac{\epsilon \log(|\mathcal{M}(\theta)|+1) - \log 2}{npL}}, \tag{5.2}$$

we have the minimax risk $\mathfrak{R} = \inf_{\psi} \sup_{\theta^* \in \Xi} \mathbb{P}(\#\text{false positives} + \#\text{false negatives} \geq 1) > 1-\epsilon$.

Proof. Our proof is based on three steps. In the first step, we reduce the problem of obtaining a lower bound of minimax risk in (5.1) for $\theta^* \in \Xi$ to the problem of deriving a lower bound for θ^* in a finite set. Next, we construct a finite set of hypotheses. Finally, we derive the minimax risk by Fano-type arguments and derive the necessary signal strength condition.

Step 1. Let $\mathcal{R}_j$ be a ranking property for a single item j as defined in Definition 2.3. We have

$$\begin{aligned} \mathfrak{R} &= \inf_{\psi} \sup_{\theta^* \in \Xi} \mathbb{P}(\#\text{false positives} + \#\text{false negatives} \geq 1) \\ &= \inf_{\psi} \sup_{\theta^* \in \Xi} \mathbb{P}(\psi \neq \psi^*) = \inf_{\psi} \sup_{\theta^* \in \Xi} \mathbb{P}(d(\psi,\psi^*) \geq 1/2), \end{aligned} \tag{5.3}$$

where $\psi^* \in \{0,1\}^n$ with $\psi_j^* = 1$ meaning $\gamma^* \in \mathcal{R}_j$ and $\psi_j^* = 0$ meaning $\gamma^* \notin \mathcal{R}_j$ for $j = 1,\cdots,n$, and $\psi \in \{0,1\}^n$ is a selection procedure that $\psi_i = 1$ means that we reject H_{0i} in (4.1), and the distance $d(\psi,\psi^*) = \|\psi - \psi^*\|_1$. By section 2.2 of Tsybakov (2008), the problem of obtaining a lower bound of minimax risk in (5.3) for $\theta^* \in \Xi$ can be reduced to the problem of deriving a lower bound for θ^* in a finite set.

In particular, suppose that we have a collection $M+1$ hypotheses $\mathcal{H}_M = \{H_i : \theta^* = \theta^{(i)}, 0 \leq i \leq M\}$. Recall that a test is any measurable function $\phi : \boldsymbol{Y} = \{\boldsymbol{Y}^{(\ell)}\}_{\ell=1}^L \mapsto \{0,1,\cdots,M\}$. Let $\mathbb{P}_{e,M}$ be the minimax probability of error

$$\begin{aligned} \mathbb{P}_{e,M} &:= \inf_{\phi} \max_{\theta^* \in \{\theta^{(0)},\theta^{(1)},\cdots,\theta^{(M)}\}} \mathbb{P}_i(\phi \neq i) \\ &= \inf_{\phi} \max_{0 \leq i \leq M} \mathbb{P}_i(\phi \neq i), \end{aligned} \tag{5.4}$$

where $\mathbb{P}_i$ denotes the probability measure under the ith hypothesis $\theta^* = \theta^{(i)}$. The following lemma in section 2.2 of Tsybakov (2008) shows that the minimax risk $\mathfrak{R}$ in (5.3) is lower bounded by $\mathbb{P}_{e,M}$ above, and we provide the proof in the online appendix, Section G.1, for self-completeness.

Lemma 5.1. *Suppose we have $M+1$ hypotheses $\mathcal{H}_M = \{H_i : \theta^* = \theta^{(i)}, 0 \leq i \leq M\}$ with deduced tests $\psi^{(i)} \in \{0,1\}^n$ where $\psi_j^{(i)} = 1$ means $\gamma(\theta^{(i)}) \in \mathcal{R}_j$ and $\psi_j^{(i)} = 0$ means $\gamma(\theta^{(i)}) \notin \mathcal{R}_j$ for $0 \leq i \leq M$, $1 \leq j \leq n$. If these hypotheses satisfy $d(\psi^{(i_1)},\psi^{(i_2)}) \geq 1$ for any $i_1 \neq i_2$, we have*

$$\mathfrak{R} = \inf_{\psi} \sup_{\theta^* \in \Xi} \mathbb{P}(d(\psi,\psi^*) \geq 1/2) \geq \mathbb{P}_{e,M}.$$

Thus, if $\mathbb{P}_{e,M} > 1-\epsilon$, we have $\mathfrak{R} > 1-\epsilon$; that is, for any selection procedure ψ, there exists a preference vector $\theta \in \Xi$ such that the probability of making at least one type I or type II error in the family is uncontrollable. This shows that if we can find a finite set $\{\theta^{(i)}, 0 \leq i \leq M\}$ satisfying the conditions of Lemma 5.1, we can derive the lower bound. We construct this set in Step 2.

Step 2. In this step, we construct a finite set of hypotheses for deriving the lower bound. We choose a base preference score vector θ first. Without loss of

generality, we assume that the preference score θ is in descending order, that is,

$$\theta_1 > \theta_2 > \cdots > \theta_n. \tag{5.5}$$

Recall that the divider set $\mathcal{M}(\theta) \subseteq [n] \times [n]$ defined in Definition 5.1 is a collection of pairs with cardinality $|\mathcal{M}(\theta)|$. Denote this set by $\mathcal{M}(\theta) = \{(k_i, k_i'), i = 1, \cdots, |\mathcal{M}(\theta)|\}$. Based on the base preference score θ and the divider set, we construct a set of hypotheses $\mathcal{H}_{\mathcal{M}} = \{H_0, H_1, \cdots, H_{|\mathcal{M}(\theta)|}\}$ such that

$$H_0 : \theta^* = \theta^{(0)},\ H_i : \theta^* = \theta^{(i)}, 1 \le i \le |\mathcal{M}(\theta)|, \tag{5.6}$$

where $\theta^{(0)} = \theta$, and $\theta^{(i)}$ is obtained by swapping scores of the ith pair $(k_i, k_i') \in \mathcal{M}(\theta)$ in the base score vector θ. For each hypothesis $H_i : \theta^* = \theta^{(i)}$, we also have the induced rank $\gamma^{(i)} = \gamma(\theta^{(i)})$ and vector $\psi^{(i)} \in \{0,1\}^n$ with $\psi_j^{(i)} = 1$ meaning $\gamma^{(i)} \in \mathcal{R}_j$ and $\psi_j^{(i)} = 0$ meaning $\gamma^{(i)} \notin \mathcal{R}_j$, $j = 1, \cdots, n$.

By the construction of $\mathcal{M}(\theta)$ in Definition 5.1, for $i_1 \neq i_2$, we have $\{j : \gamma^{(i_1)} \in \mathcal{R}_j\} \neq \{j : \gamma^{(i_2)} \in \mathcal{R}_j\}$, which gives $\psi^{(i_1)} \neq \psi^{(i_2)}$, and we have $d(\psi^{(i_1)}, \psi^{(i_2)}) \ge 1$ for any $i_1 \neq i_2$, which satisfies the condition in Lemma 5.1.

Step 3. In this step, we obtain a lower bound on the minimax risk by Fano-type bounds.

Here we denote $\mathcal{M}(\theta)$ by $\mathcal{M}$ for simplicity. Let the average probability of error and the minimum average probability of error of a test $\phi : \boldsymbol{Y} = \{\boldsymbol{Y}^{(\ell)}\}_{\ell=1}^{L} \longmapsto \{0, 1, \cdots, M\}$ be

$$\bar{\mathbb{P}}_{e,\mathcal{M}}(\phi) = \frac{1}{|\mathcal{M}|+1} \sum_{j=0}^{|\mathcal{M}|} \mathbb{P}_j(\phi \neq j), \quad \text{and}$$

$$\bar{\mathbb{P}}_{\mathrm{e},\mathcal{M}} = \inf_{\phi} \bar{\mathbb{P}}_{\mathrm{e},\mathcal{M}}(\phi).$$

One can easily verify that $\mathbb{P}_{e,\mathcal{M}} \ge \bar{\mathbb{P}}_{e,\mathcal{M}}$, where the minimax probability of error $\mathbb{P}_{e,\mathcal{M}}$ is defined in (5.4). Then we reduce bounding the minimax probability of error to bounding the minimum average probability of error because, if $\bar{\mathbb{P}}_{e,\mathcal{M}} > 1 - \epsilon$, we have $\mathfrak{R} > 1 - \epsilon$.

Following the argument in Chen and Suh (2015), we also apply the generalized Fano inequality in Verdú (1994), which gives a lower bound for $\bar{\mathbb{P}}_{\mathrm{e},\mathcal{M}}$. We summarize the result in the following lemma and provide the proof in the online appendix, Section G.2.

Lemma 5.2. *Under the BTL model, let $\mathcal{M} = \mathcal{M}(\theta)$ be the divider set defined in Definition 5.1. Given the set of hypotheses $\mathcal{H}_{\mathcal{M}}$ constructed in Step 2, we have*

$$\bar{\mathbb{P}}_{\mathrm{e},\mathcal{M}} \ge 1 - \frac{1}{\log(|\mathcal{M}|+1)} \left\{ \frac{pL}{(|\mathcal{M}|+1)^2} \sum_{H, \widetilde{H} \in \mathcal{H}_{\mathcal{M}}} \sum_{i<j} \mathrm{KL}\left(\mathbb{P}_{y_{i,j}^{(1)}|H} \middle\| \mathbb{P}_{y_{i,j}^{(1)}|\widetilde{H}} \right) + \log 2 \right\}. \tag{5.7}$$

Let $\omega = \exp(\theta)$, and define

$$d_{\mathcal{M}}(\omega) = \max_{(k_1,k_2)\in\mathcal{M}} \left(\frac{\omega_{\min(k_1,k_2)}}{\omega_{\min(k_1,k_2)+1}} + \cdots + \frac{\omega_{\max(k_1,k_2)-1}}{\omega_{\max(k_1,k_2)}} - |k_1 - k_2| \right).$$

We further have the following lemma to control the right-hand side of (5.7). We provide the proof in the online appendix, Section G.3.

Lemma 5.3. *Under the BTL model, let $\mathcal{M} = \mathcal{M}(\theta)$ be the divider set defined in Definition 5.1. Given score vector ω $(\omega_1 > \cdots > \omega_n)$ and the set of hypotheses $\mathcal{H}_{\mathcal{M}}$ constructed in Step 2, we have*

$$\sum_{H, \widetilde{H} \in \mathcal{H}_{\mathcal{M}}} \sum_{i<j} \mathrm{KL}\left(\mathbb{P}_{y_{i,j}^{(1)}|H} \middle\| \mathbb{P}_{y_{i,j}^{(1)}|\widetilde{H}} \right) \le 4n\,|\mathcal{M}|^2 (d_{\mathcal{M}}^2(\omega) + \mathcal{O}(d_{\mathcal{M}}^3(\omega))). \tag{5.8}$$

Plugging (5.8) into (5.7), we have $\bar{\mathbb{P}}_{\mathrm{e},\mathcal{M}} > 1 - \epsilon$ if

$$\frac{1}{\log(|\mathcal{M}|+1)} \left\{ 4npL \frac{|\mathcal{M}|^2}{(|\mathcal{M}|+1)^2} (d_{\mathcal{M}}^2(\omega) + \mathcal{O}(d_{\mathcal{M}}^3(\omega))) + \log 2 \right\} < \epsilon. \tag{5.9}$$

Finally, we have the following lemma, and the proof is provided in the online appendix, Section G.4.

Lemma 5.4. *Under the BTL model where the pairwise comparison probability p in Erdös-Rényi Graph satisfies $p \gtrsim \frac{\log n}{n}$, let $\mathcal{M} = \mathcal{M}(\theta)$ be the divider set defined in Definition 5.1, and θ $(\theta_1 > \theta_2 > \cdots > \theta_n)$ be the score vector in Step 2. Assuming that*

$$\Delta(\theta) \lesssim \sqrt{\frac{\epsilon \log(|\mathcal{M}|+1) - \log 2}{npL}},$$

we have

$$d_{\mathcal{M}}(\omega) \lesssim \sqrt{\frac{\epsilon \log(|\mathcal{M}|+1) - \log 2}{npL}}. \tag{5.10}$$

We have that (5.10) implies (5.9). This essentially concludes the proof that when (5.2) holds, we have the minimax risk $\mathfrak{R} \ge \bar{\mathbb{P}}_{\mathrm{e},\mathcal{M}} > 1 - \epsilon$. □

5.2. Applications

We provide some examples of ranking property testing to illustrate the lower bound.

Example 5.1 (Top-K Items Inference). Consider the problem of selecting all items ranked among top-K. We construct a base preference score θ satisfying $\theta_1 = \cdots = \theta_K > \theta_{K+1} = \cdots = \theta_n$. Here the target select set is $\{i : \gamma(\theta) \in \mathcal{R}_i\} = \{1, \cdots, K\}$. Intuitively, for this example, if θ_K and θ_{K+1} are close, this multiple testing

problem becomes difficult, and Theorem 5.1 justifies this intuition.

In particular, because swapping scores θ_i $(1 \le i \le K)$ with θ_j $(K+1 \le j \le n)$ changes the top K items, and we have $|\theta_i - \theta_j| = \theta_K - \theta_{K+1}$, we have $\{(i,j)\}_{1\le i\le K, K+1\le j\le n}$ are all legal pairs, and the distance is $\Delta(\theta) = \theta_K - \theta_{K+1}$. Among all legal pairs, after swapping scores θ_i $(1 \le i \le K)$ with θ_j $(K+1 \le j \le n)$, the target selection set becomes $\{[n]/\{i\}\} \cup \{j\}$. Thus, all legal pairs are included in divider set $\mathcal{M}(\theta) = \{(i,j)\}_{1\le i\le K, K+1\le j\le n}$, which gives us that $\log(|\mathcal{M}(\theta)|) \asymp \log(n)$. Thus, Theorem 5.1 implies that in this example, the minimax risk $\mathfrak{R}$ is uncontrollable if

$$\Delta(\theta) = \theta_K - \theta_{K+1} \lesssim \sqrt{\frac{\epsilon \log(n)}{npL}}.$$

Example 5.2. Consider the problem of inferring all items ranked higher than item k with score θ_k. We construct a base preference score θ satisfying $\theta_1 = \cdots = \theta_{k-1} > \theta_k = \cdots = \theta_n$. In this example, we have $\{i : \gamma(\theta) \in \mathcal{R}_i\} = \{1, \cdots, k-1\}$. Swapping score θ_i $(1 \le i \le k-1)$ with θ_j $(k \le j \le n)$ changes the target selection set $\{i : \gamma(\theta) \in \mathcal{R}_i\} = \{1, \cdots, k-1\}$, and $|\theta_i - \theta_j| = \theta_{k-1} - \theta_k$. Thus, $\{(i,j)\}_{1\le i\le k-1, k\le j\le n}$ are all legal pairs, and the distance is $\Delta(\theta) = \theta_{k-1} - \theta_k$. Among all legal pairs, after swapping score θ_i $(1 \le i \le k-1)$ with θ_k, we have the same target selection set $\{i : \gamma(\theta) \in \mathcal{R}_i\} = \emptyset$, so only one of them can be included in set $\mathcal{M}(\theta)$. We then have $\mathcal{M}(\theta) = \{(i,j)\}_{1\le i\le k-1, k+1\le j\le n} \cup \{(k-1,k)\}$ with $\log(|\mathcal{M}(\theta)|) = \log((k-1)(n-k)+1) \asymp \log(n)$. Theorem 5.1 implies that in this example, the minimax risk $\mathfrak{R}$ is uncontrollable if $\Delta(\theta) = \theta_{k-1} - \theta_k \lesssim \sqrt{\frac{\epsilon \log(n)}{npL}}$.

Remark 5.1 (Upper Bound). In the previous two examples, we have that if $\Delta(\theta) \lesssim \sqrt{\frac{\epsilon \log n}{npL}}$, any procedure fails to control the risk. Meanwhile, Theorem 4.1 shows that if $\Delta(\theta) \gtrsim \sqrt{\frac{\log n}{npL}}$, we can control the FWER and achieve power one asymptotically at the same time, which shows our procedure achieves the minimax optimality.

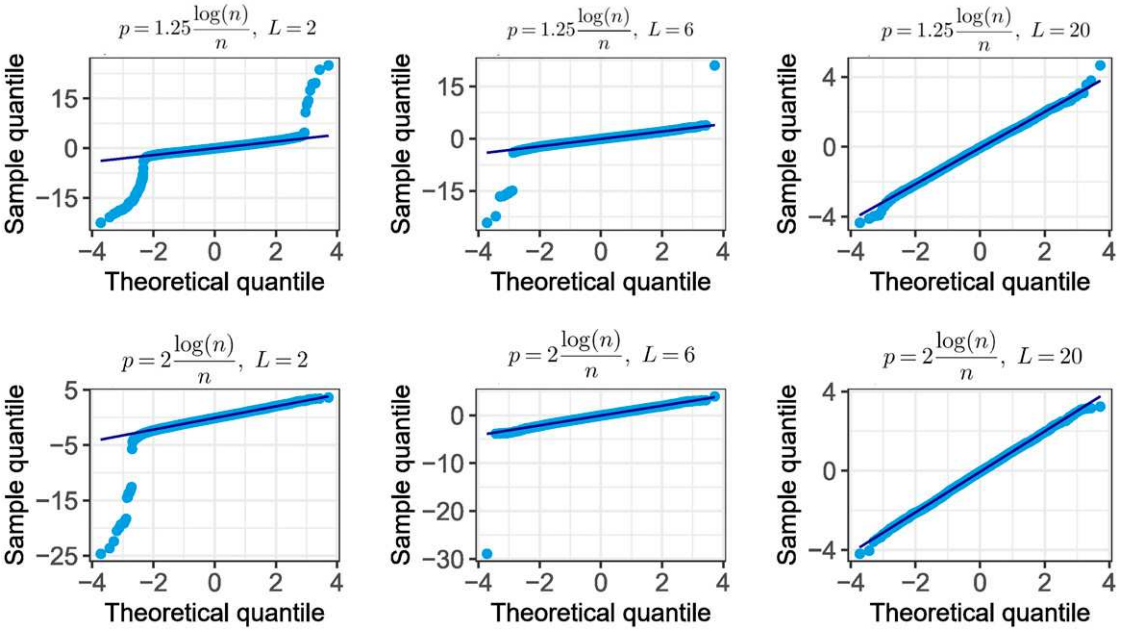
6. Numerical Experiments

In this section, we conduct extensive numerical studies to test the empirical performance of the proposed methods using both synthetic data and two real datasets.

6.1. Synthetic Data

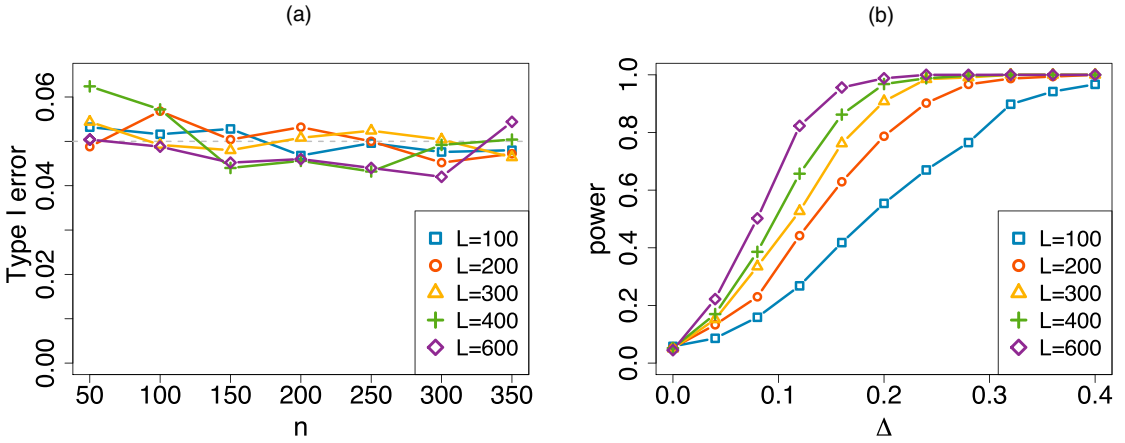
Using synthetic data, we first investigate the asymptotic normality of Lagrangian debiased estimators for latent preference scores. Specifically, we generate the latent scores θ_i^* s independently from a uniform distribution over $[8, 10]$, where we set the number of items $n = 100$. We let $L \in \{2, 6, 20\}$, $p \in \{1.25, 2\} \times \frac{\log(n)}{n}$. Following the procedure developed in Section 3.1, we repeat the generating scheme 2,500 times and present the empirical distribution of the estimator. In particular, Figure 3

Figure 3. (Color online) Q-Q Plots for Lagrangian Debiased Estimators, Comparing Quantiles of Standardized Lagrangian Debiased Estimators with Standard Normal



Note. We fix $n = 100$ and let $L \in \{2, 6, 20\}$, $p \in \{1.25, 2\} \times \frac{\log(n)}{n}$.

Figure 4. (Color online) Performance of Pairwise Test $H_0 : \theta_1^* \le \theta_2^*$ vs. $H_a : \theta_1^* > \theta_2^*$



Notes. (a) Type I error for our proposed pairwise test procedure with different n and L. (b) Power curve as a function of signal strength $\Delta = |\theta_1^* - \theta_2^*|$ with $n = 100$, $p = 0.2$, and $L \in \{100, 200, 300, 400, 600\}$.

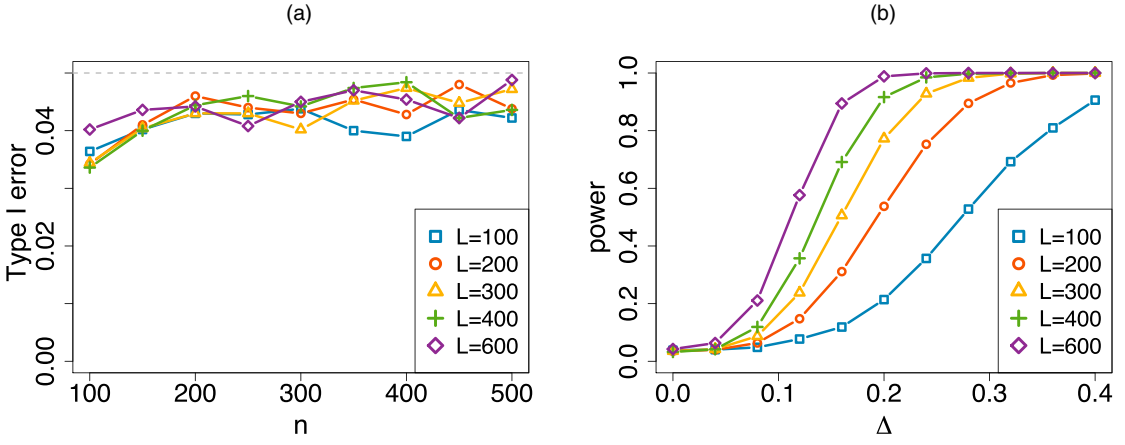
displays the Q-Q plots of θ_1, and we find that the empirical distribution of our estimator is closed to a normal distribution, especially when the number of repeated comparisons L is large. This justifies our result in Theorem 3.2 that our estimator weakly converges to a normal distribution.

Next, we examine the performance of the pairwise test and top-K test procedure in Section 3.2. For pairwise test, we test $H_0 : \theta_1^* \le \theta_2^*$ v.s. $H_a : \theta_1^* > \theta_2^*$. Figure 4(a) displays the type I error with $p = 0.2$ for different total number of items n and number of repeated comparisons L. Here we fix $\theta_1^* = \theta_2^* = 10$ and $\theta_i^* = 7.5$ for $3 \le i \le n$. As seen from this figure, the empirical type I error rate is close to the nominal $\alpha = 0.05$. Figure 4(b) shows the empirical power of this test with different $\Delta = |\theta_1^* - \theta_2^*|$. Here we set $\theta_1^* = 10, \theta_2^* = 10 - \Delta$ and $\theta_i^* = 7.5$ for $3 \le i \le n$, and we let $n = 100$ and $p = 0.2$ and change Δ and L. We observe that the empirical power goes to one quickly.

For top-K test, we test whether item $K + 1$ is ranked among the top K items. We fix $K = 30$ and $p = 0.2$, and we let $\theta_i^* = 10$ for $1 \le i \le K+1$, and $\theta_j^* = 7.5$ for $K + 2 \le j \le n$. Figure 5(a) displays the empirical type I error rate with different n and L, and we find that that the empirical type I error is close to the nominal $\alpha = 0.05$. Figure 5(b) displays the empirical power of this test with different separation Δ between top K items and other items that $\Delta = |\theta_{(K)}^* - \theta_{(K+1)}^*|$. We let $n = 100$, $L \in \{100, 200, 300, 400, 600\}$, and we set $\theta_i^* = 10$ for $1 \le i \le K$, $\theta_{K+1}^* = 10 - \Delta$, and $\theta_j^* = 7.5$ for $K + 2 \le j \le n$. As seen from this plot, the empirical power goes to one as Δ and L increase.

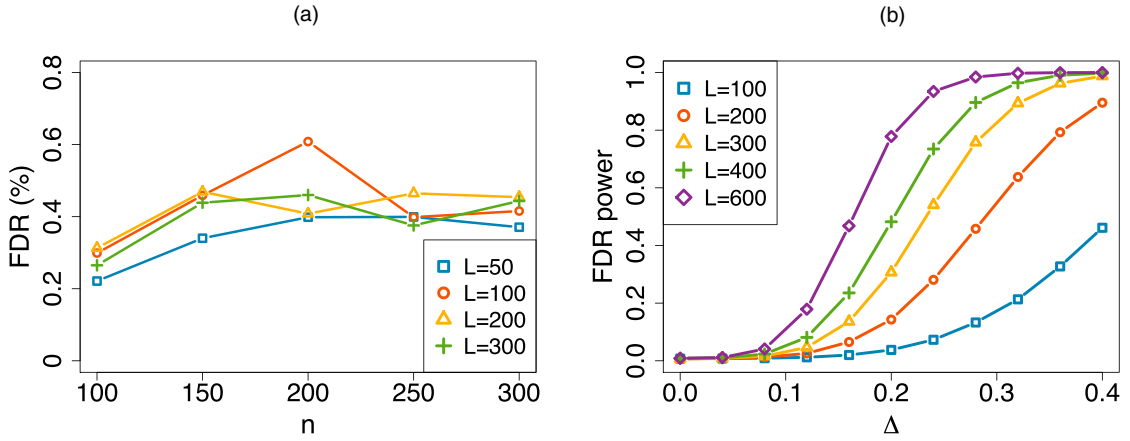
Finally, we evaluate the empirical performance of our FDR procedure in Section 4.2 by considering the

Figure 5. (Color online) Performance of Top 30 Test Using Our Proposed Simplified Testing Procedure in Remark 3.5



Notes. (a) Averaged type I error with different n and L. (b) Averaged power as a function of signal strength $\Delta = |\theta_{(K)}^* - \theta_{(K+1)}^*|$ with $n = 100$, $p = 0.2$, and $L \in \{100, 200, 300, 400, 600\}$.

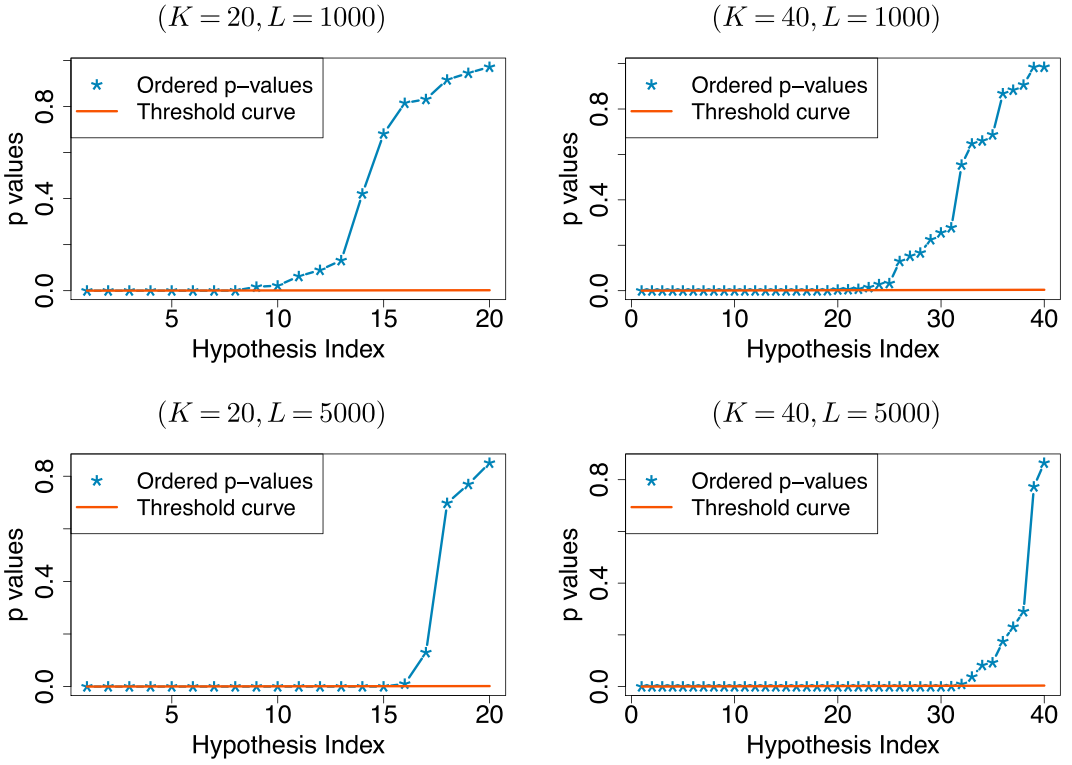
Figure 6. (Color online) Performance of Benjamini-Yekutieli–Based FDR Procedure for Selecting Top 30 Items



Notes. (a) Empirical FDR with different n and L. (b) Empirical FDR power as a function of signal strength $\Delta = |\theta_{(K)} - \theta_{(K+1)}|$ with $n = 100$, $p = 0.2$, and $L \in \{100, 200, 300, 400, 600\}$.

top-K test. We let $K = 30$, $\theta_i^* = 10$ for $1 \le i \le K + 10$, and $\theta_j^* = 6.5$ for $K + 11 \le j \le n$. Figure 6(a) displays the empirical FDR based on our procedure with $p = 0.2$ and different n, L. This figure illustrates that the FDR is well controlled below the nominal $\alpha = 0.05$, consistent with our results in Theorem 4.2. Figure 6(b) displays the empirical power of the FDR procedure, which is defined as true positive rate, with different separation $\Delta = |\theta_{(K)}^* - \theta_{(K+1)}^*|$. We set $\theta_i^* = 10$ for $1 \le i \le K$, $\theta_{K+1}^* = 10 - \Delta$, and $\theta_j^* = 6.5$ for $K + 2 \le j \le n$, and we let $n = 100$, $p = 0.2$, and $L \in \{100, 200, 300, 400, 600\}$. As shown in this plot, the empirical power

Figure 7. (Color online) Application of Our FDR Procedure on Jester Data Set to Select Top K Jokes



Notes. The four panels display the ordered p values and adjusted threshold by Benjamini-Yekutieli procedure with $K \in \{20, 40\}$ and $L \in \{1000, 5000\}$. The horizontal red line represents 0.05.

Table 1. Top 10 Jokes on the Jester Data Set and Their Estimated Scores Obtained from the Spectral Method and Our Lagrangian Debiasing Method

	Spectral method		Debiasing method	
Joke ID	Score	Rank	Score	Rank
89	0.841	1	0.840	1
50	0.799	2	0.801	2
29	0.651	3	0.645	3
36	0.623	4	0.628	4
27	0.621	5	0.620	5
62	0.616	6	0.616	6
32	0.603	7	0.599	7
35	0.596	8	0.596	8
54	0.527	9	0.526	9
69	0.515	10	0.516	10

increases and goes to one as Δ and L increase, showing that the proposed FDR procedure is able to identify the top K items with well-controlled FDR.

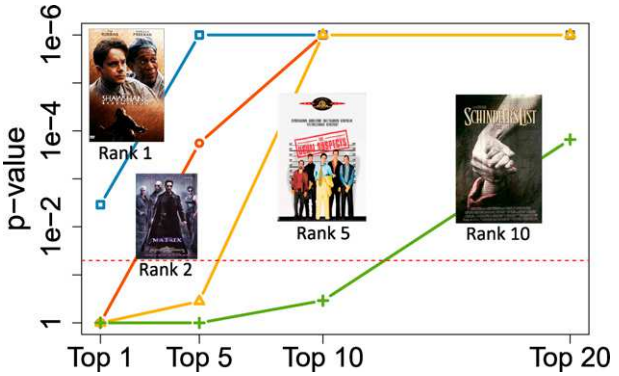
6.2. Real Data

In this section, we apply our method to analyze two real data sets.

6.2.1. Jester Data Set. We first apply our method to the Jester data set from Goldberg et al. (2001). This data set contains ratings of 100 jokes from 73,421 users. The more detailed description and the data set are available through http://eigentaste.berkeley.edu/dataset/. In this data set, 14,116 users rated all 100 jokes, whereas others only rated some of jokes. We only use samples from users who ranked all 100 jokes for our experiments. Because we need pairwise comparisons for our ranking analysis, we generate Erdös-Rényi comparison graph randomly with $p = 0.3$ and obtain each pairwise comparison results based on the relative rating of compared pairs by the same user. To be specific, if joke 1 receives a higher rating than joke 2 from a same user, we have joke 1 beats joke 2 in this comparison. Negahban et al. (2017) and Kim et al. (2017) also use similar approaches to break rating results into pairwise comparisons. Furthermore, we randomly choose L samples from the total 14,116 samples.

Table 1 displays the top 10 jokes' IDs and their estimated scores obtained from the spectral method and our debiasing method with $L = 1{,}000$. Furthermore, we evaluate the performance of our FDR-controlling procedure with $K \in \{20, 40\}$ and $L \in \{1000, 5000\}$. Figure 7 presents the p values from multiple testing and threshold obtained from Benjamini-Yekutieli procedure described in Section 4.2. It shows that our FDR procedure gains more power as L increases.

Figure 8. (Color online) Application of Top-K Test on Four Movies (*The Shawshank Redemption*, *The Matrix*, *The Usual Suspects*, *Schindler's List*) in the MovieLens Data Set, Which Are Ranked 1, 2, 5, and 10 Based on a Lagrangian Debiasing Estimator



Notes. The figure displays the change of their p values in the top 1, 5, 10, and 20 test by our proposed testing procedure in Remark 3.5. The horizontal dotted red line represents 0.05.

6.2.2. MovieLens Data Set. We also apply our method to analyze the MovieLens data set (Harper and Konstan 2015). Similar to our analysis before, we obtain pairwise comparisons results based on the relative ratings of two movies by the same user. In particular, we analyze $n = 218$ movies with the largest number of ratings and randomly sample $L = 1{,}000$ comparisons.

Table 2. Top 10 Movies in the MovieLens Data Set Based on a Lagrangian Debiased Estimator

Rank	Movie title	Average rating	Debiased score	p value in top 10 test	p value in top 20 test
1	*The Shawshank Redemption* (1994)	4.42	1.985	$< 1e-6$	$< 1e-6$
2	*The Matrix* (1999)	4.16	1.766	$< 1e-6$	$< 1e-6$
3	*The Godfather* (1972)	4.25	1.755	$< 1e-6$	$< 1e-6$
4	*Star Wars: Episode V—The Empire Strikes Back* (1980)	4.12	1.684	$< 1e-6$	$< 1e-6$
5	*The Usual Suspects* (1995)	4.28	1.530	$< 1e-6$	$< 1e-6$
6	*Star Wars: Episode IV—A New Hope* (1977)	4.10	1.471	0.0010	$< 1e-6$
7	*The Silence of the Lambs* (1991)	4.15	1.418	0.0070	$< 1e-6$
8	*Seven* (a.k.a. *Se7en*) (1995)	4.08	1.367	0.0401	$< 1e-6$
9	*Lord of the Rings: The Fellowship of the Ring* (2001)	4.10	1.329	0.1133	$< 1e-6$
10	*Schindler's List* (1993)	4.25	1.288	0.3431	0.0002

Notes. The table includes their titles, average rating in https://movielens.org/, estimated scores by Lagrangian debiased procedure, and their corresponding p values in the top 10 and top 20 test by our proposed testing procedure in Remark 3.5.

Table 2 shows the top 10 movies with highest scores based on Lagrangian debiasing procedure and the corresponding p values, where we test whether they are ranked among the top 10 and top 20 movies. Figure 8 displays the change of p values when each movie is tested if it is among top 1, 5, 10, or 20 ranked movies. We display the p values of four movies (*The Shawshank Redemption*, *The Matrix*, *The Usual Suspects*, *Schindler's List*), which are ranked 1, 2, 5, and 10 based on our debiased estimator.

7. Conclusion

To conclude, to the best of our knowledge, we propose the first general framework for conducting inference and quantifying uncertainties for ranking problems. Under the BTL model, we first propose a Lagrangian debiasing method to infer the latent score for each item, where we can then test "local" properties. Next, by leveraging the powerfulness of Gaussian multiplier bootstrap, we can test more general "global" properties. Furthermore, we extend the framework to multiple testing problems where we control both the familywise type I error and the FDR. We prove the optimality of the proposed method by deriving the minimax lower bound. Using both synthetic and real data sets, we demonstrate that our method works well in practice.

There are still numerous promising directions that would be of interest for future investigations. We point out a few possibilities as follows. First, the Gaussian multiplier bootstrap approach is computationally expensive, and it is worth investigating if we can develop a more computationally efficient approach to make the method more scalable. Second, as we have discussed, the current approach for FDR control is conservative, and we plan to develop a more powerful method to tightly control the FDR. In addition, the ranking of different items may change over time, and we plan to develop new models to study dynamic ranking systems.

Acknowledgments

The authors thank the area editor, associate editor, and two reviewers for constructive comments, which led to a significant improvement of the earlier version of this paper and Chao Gao, Yandi Shen, and Anderson Y. Zhang for giving the independent credit. E. X. Fang and J. Lu are joint corresponding authors.

References

Adelman D (2020) An efficient frontier approach to scoring and ranking hospital performance. *Oper. Res.* 68(3):762–792.

Adler M, Gemmell P, Harchol-Balter M, Karp RM, Kenyon C (1994) Selection in the presence of noise: The design of playoff systems. *Proc. 5th Annual ACM-SIAM Sympos. on Discrete Algorithms*, 564–572.

Ailon N (2010) Aggregation of partial rankings, p-ratings and top-m lists. *Algorithmica* 57(2):284–300.

Ailon N, Charikar M, Newman A (2008) Aggregating inconsistent information: ranking and clustering. *J. ACM* 55(5):1–27.

Altman A, Tennenholtz M (2005) Ranking systems: The PageRank axioms. *Proc. 6th ACM Conf. on Electronic Commerce*, 1–8.

Ammar A, Shah D (2011) Ranking: Compare, don't score. *Proc. 49th Annual Allerton Conf. on Comm., Control, and Computing* (IEEE, New York), 776–783.

Ammar A, Shah D (2012) Efficient rank aggregation using partial data. *Performance Evaluation Rev.* 40(1):355–366.

Aouad A, Farias V, Levi R, Segev D (2018) The approximability of assortment optimization under ranking preferences. *Oper. Res.* 66(6):1661–1669.

Azari Soufiani H, Chen WZ, Parkes DC, Xia L (2013) Generalized method-of-moments for rank aggregation. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Adv. Neural Inform. Processing Systems* 2706–2714.

Baltrunas L, Makcinskas T, Ricci F (2010) Group recommendations with rank aggregation and collaborative filtering. *Proc. 4th ACM Conf. on Recommender Systems*, 119–126.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Statist. Soc. B* 57(1):289–300.

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* 29(4) 1165–1188.

Beutel A, Chen J, Doshi T, Qian H, Wei L, Wu Y, Heldt L, et al. (2019) Fairness in recommendation ranking through pairwise comparisons. *Proc. 25th ACM SIGKDD Internat. Conf. on Knowledge Discovery & Data Mining*, 2212–2220.

Bouadjenek MR, Hacid H, Bouzeghoub M (2013) Sopra: A new social personalized ranking function for improving web search. *Proc. 36th Internat. ACM SIGIR Conf. on Res. and Development in Inform. Retrieval*, 861–864.

Boulesteix AL, Slawski M (2009) Stability and aggregation of ranked gene lists. *Brief Bioinform.* 10(5):556–568.

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39(3–4): 324–345.

Brown LD (2003) Ranking journals using social science research network downloads. *Rev. Quant. Finance Accounting* 20(3):291–307.

Busa-Fekete R, Szorenyi B, Cheng W, Weng P, Hüllermeier E (2013) Top-k selection based on adaptive sampling of noisy preferences. *Proc. Internat. Conf. on Machine Learn.*, 1094–1102.

Cai T, Liu W, Luo X (2011) A constrained L1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* 106(494):594–607.

Candès EJ, Recht B (2009) Exact matrix completion via convex optimization. *Foundations Comput. Math.* 9(6):717–772.

Casella G, Berger RL (2021) *Statistical Inference* (Cengage Learning, Boston).

Chen P, Gao C, Zhang AY (2020) Partial recovery for top-*k* ranking: Optimality of MLE and suboptimality of spectral method. Preprint, submitted June 30, https://arxiv.org/abs/2006.16485.

Chen P, Gao C, Zhang AY (2021) Optimal full ranking from pairwise comparisons. Preprint, submitted January 21, https://arxiv.org/abs/2101.08421.

Chen X, Chen Y, Li X (2022) Asymptotically optimal sequential design for rank aggregation. *Math. Oper. Res.* Forthcoming.

Chen X, Jiao K, Lin Q (2016) Bayesian decision process for cost-efficient dynamic ranking via crowdsourcing. *J. Machine Learn. Res.* 17(1):7617–7656.

Chen X, Li Y, Mao J (2018b) A nearly instance optimal algorithm for top-k ranking under the multinomial logit model. *Proc. 29th Annual ACM-SIAM Sympos. on Discrete Algorithms*, 2504–2522.

Chen X, Bennett PN, Collins-Thompson K, Horvitz E (2013) Pairwise ranking aggregation in a crowdsourced setting. *Proc. 6th ACM Internat. Conf. on Web Search and Data Mining*, 193–202.

Chen X, Gopi S, Mao J, Schneider J (2018a) Optimal instance adaptive algorithm for the top-*k* ranking problem. *IEEE Trans. Inform. Theory* 64(9):6139–6160.

Chen Y, Suh C (2015) Spectral MLE: Top-k rank aggregation from pairwise comparisons. *Internat. Conf. on Machine Learn.*, 371–380.

Chen Y, Fan J, Ma C, Wang K (2019) Spectral method and regularized MLE are both optimal for top-*k* ranking. *Ann. Statist.* 47(4): 2204.

Chernoff H (1954) On the distribution of the likelihood ratio. *Ann. Math. Statist.* 25(3):573–578.

Chernozhukov V, Chetverikov D, Kato K (2013) Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* 41(6):2786–2819.

Copeland AH (1951) A reasonable social welfare function. Technical report, University of Michigan, Ann Arbor, MI.

Cremonesi P, Koren Y, Turrin R (2010) Performance of recommender algorithms on top-n recommendation tasks. *Proc. 4th ACM Conf. on Recommender Systems*, 39–46.

Ding J, Han D, Dezert J, Yang Y (2018) A new hierarchical ranking aggregation method. *Inform. Sci.* 453:168–185.

Duchi JC, Mackey LW, Jordan MI (2010) *On the Consistency of Ranking Algorithms* (ICML).

Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the web. *Proc. 10th Internat. Conf. on World Wide Web*, 613–622.

Eisenach C, Bunea F, Ning Y, Dinicu C (2020) High-dimensional inference for cluster-based graphical models. *J. Machine Learn. Res.* 21(53).

Farnoud F, Touri B, Milenkovic O (2012) Novel distance measures for vote aggregation. Preprint, submitted March 28, https://arxiv.org/abs/1203.6371.

Filiberto Y, Bello R, Nowe A (2018) A new method for personnel selection based on ranking aggregation using a reinforcement learning approach. *Comput. Systems* 22(2):537–546.

Ford Jr LR (1957) Solution of a ranking problem from binary comparisons. *Amer. Math. Monthly* 64(8P2):28–33.

Gao C, Shen Y, Zhang AY (2021) Uncertainty quantification in the Bradley-Terry-Luce model. Preprint, submitted October 8, https://arxiv.org/abs/2110.03874.

Ge Y, Dudoit S, Speed TP (2003) Resampling-based multiple testing for microarray data analysis. *TEST.* 12(1):1–77.

Geyik SC, Ambler S, Kenthapadi K (2019) Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. *Proc. 25th ACM SIGKDD Internat. Conf. on Knowledge Discovery & Data Mining*, 2221–2231.

Gleich DF Lim Lh (2011) Rank aggregation via nuclear norm minimization. *Proc. 17th ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining*, 60–68.

Goldberg K, Roeder T, Gupta D, Perkins C (2001) Eigentaste: A constant time collaborative filtering algorithm. *Inform. Retrieval* 4(2):133–151.

Gourieroux C, Holly A, Monfort A (1982) Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica* 50(1):63–80.

Guiver J, Snelson E (2009) Bayesian inference for Plackett-Luce ranking models. *Proc. 26th Annual Internat. Conf. on Machine Learn.*, 377–384.

Guo J, Fan Y, Pang L, Yang L, Ai Q, Zamani H, Wu C, et al. (2020) A deep look into neural ranking models for information retrieval. *Inform. Processing Management* 57(6):102067.

Hajek B, Oh S, Xu J (2014) Minimax-optimal inference from partial rankings. *Adv. Neural Inform. Processing Systems* 1475–1483.

Hall P, Miller H (2009) Using the bootstrap to quantify the authority of an empirical ranking. *Ann. Statist.* 37(6B):3929–3959.

Harper FM, Konstan JA (2015) The movielens datasets: History and context. *ACM Trans. Interactive Intelligent Systems* 5(4):1–19.

He X, He Z, Du X, Chua TS (2018) Adversarial personalized ranking for recommendation. *Proc. 41st Internat. ACM SIGIR Conf. on Res. & Development in Inform. Retrieval*, 355–364.

Heckel R, Shah NB, Ramchandran K, Wainwright MJ (2019) Active ranking from pairwise comparisons and when parametric assumptions do not help. *Ann. Statist.* 47(6):3099–3126.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 6(2):65–70.

Hunter DR (2004) MM algorithms for generalized Bradley-Terry models. *Ann. Statist.* 32(1):384–406.

Jain L, Gilbert A, Varma U (2020) Spectral methods for ranking with scarce data. *Proc. Conf. on Uncertainty in Artificial Intelligence*, 609–618.

Jamieson KG, Nowak RD (2011) Active ranking using pairwise comparisons. Preprint, submitted September 16, https://arxiv.org/abs/1109.3701.

Jang M, Kim S, Suh C (2018) Top-*k* rank aggregation from *m*-wise comparisons. *IEEE J. Sel. Top. Signal Process.* 12(5):989–1004.

Jang M, Kim S, Suh C, Oh S (2016) Top-*k* ranking from pairwise comparisons: When spectral ranking is optimal. Preprint, submitted March 14, https://arxiv.org/abs/1603.04153.

Jang M, Kim S, Suh C, Oh S (2017) Optimal sample complexity of M-wise data for top-K ranking. *Proc. 31st Internat. Conf. on Neural Inform. Processing Systems*, 1685–1695.

Javanmard A, Montanari A (2014a) Confidence intervals and hypothesis testing for high-dimensional regression. *J. Machine Learn. Res.* 15(1):2869–2909.

Javanmard A, Montanari A (2014b) Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inform. Theory* 60(10): 6522–6554.

Jiang X, Lim LH, Yao Y, Ye Y (2011) Statistical ranking and combinatorial Hodge theory. *Math. Programming* 127(1):203–244.

Jin T, Xu P, Gu Q, Farnoud F (2020) Rank aggregation via heterogeneous Thurstone preference models. *Proc. AAAI Conf. on Artificial Intelligence*.

Keener JP (1993) The Perron–Frobenius theorem and the ranking of football teams. *SIAM Rev.* 35(1):80–93.

Kendall MG (1955) Further contributions to the theory of paired comparisons. *Biometrics* 11(1):43–62.

Kendall MG, Smith BB (1940) On the method of paired comparisons. *Biometrika* 31(3–4):324–345.

Kim M, Farnoud F, Milenkovic O (2015) Hydra: Gene prioritization via hybrid distance-score rank aggregation. *Bioinformatics* 31(7): 1034–1043.

Kim Y, Kim W, Shim K (2017) Latent ranking analysis using pairwise comparisons in crowdsourcing platforms. *Inform. Systems* 65:7–21.

Kodde DA, Palm FC (1986) Wald criteria for jointly testing equality and inequality restrictions. *Econometrica* 54(5):1243–1248.

Kolde R, Laur S, Adler P, Vilo J (2012) Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 28(4): 573–580.

Li MM, Liu X, Huang Y, Shi C (2018) Integrating empirical estimation and assortment personalization for e-commerce: A consider-then-choose model. Preprint, submitted September 10, https://dx.doi.org/10.2139/ssrn.3247323.

Liang S, de Alfaro L (2020) Online top-k selection in crowdsourcing environments. *Proc. 6th Internat. Conf. on Comput. and Data Engrg.*, 252–259.

Lu T, Boutilier C (2011) *Learning Mallows Models with Pairwise Preferences* (ICML).

Lu Y, Negahban SN (2015) Individualized rank aggregation using nuclear norm regularization. *Proc. 53rd Annual Allerton Conf. on Comm., Control, and Comput.* (IEEE, New York), 1473–1479.

Luce RD (1959) *Individual Choice Behavior. A Theoretical Analysis* (Wiley, New York).

Mallows CL (1957) Nonnull ranking models. *Biometrika* 44(1–2): 114–130.

Maystre L, Grossglauser M (2015) Fast and accurate inference of Plackett–Luce models. Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, eds. *Adv. Neural Inform. Processing Systems* 28: 172–180.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed. *Frontiers in Econometrics* (Academic Press, New York), 105–142.

Minka T, Cleven R, Zaykov Y (2018) Trueskill 2: An improved Bayesian skill rating system. Technical report MSR-TR-2018-8, Microsoft, Redmond, WA.

Minka T, Graepel T, Herbrich R (2007) Trueskill™: A Bayesian skill rating system. *Adv. Neural Inform. Processing Systems.* 19:569–576.

Mohammadi M, Rezaei J (2020) Ensemble ranking: Aggregation of rankings produced by different multicriteria decision-making methods. *Omega* 96:102254.

Molenberghs G, Verbeke G (2007) Likelihood ratio, score, and Wald tests in a constrained parameter space. *Amer. Statist.* 61(1): 22–27.

Mosteller F (2006) Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Selected Papers of Frederick Mosteller* (Springer, Berlin), 157–162.

Nápoles G, Falcon R, Dikopoulou Z, Papageorgiou E, Bello R, Vanhoof K (2017) Weighted aggregation of partial rankings using ant colony optimization. *Neurocomput.* 250:109–120.

Negahban S, Oh S, Shah D (2017) Rank centrality: Ranking from pairwise comparisons. *Oper. Res.* 65(1):266–287.

Negahban S, Oh S, Thekumparampil KK, Xu J (2018) Learning from comparisons and choices. *J. Machine Learn. Res.* 19(1): 1478–1572.

Negahban S, Wainwright MJ (2012) Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Machine Learn. Res.* 13:1665–1697.

Neudorfer A, Rosset S (2018) Predicting the NCAA basketball tournament using isotonic least squares pairwise comparison model. *J. Quant. Anal. Sports* 14(4):173–183.

Ning Y, Liu H (2017) A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* 45(1):158–195.

Orbán-Mihálykó É, Mihálykó C, Koltay L (2019) A generalization of the Thurstone method for multiple choice and incomplete paired comparisons. *Central Eur. J. Oper. Res.* 27(1):133–159.

Osting B, Brune C, Osher S (2013) Enhanced statistical rankings via targeted data collection. *Internat. Conf. on Machine Learn.*, 489–497.

Pananjady A, Mao C, Muthukumar V, Wainwright MJ, Courtade TA (2017) Worst-case vs average-case design for estimation from fixed pairwise comparisons. Preprint, submitted July 19, https://arxiv.org/abs/1707.06217.

Pelechrinis K, Papalexakis E, Faloutsos C (2016) SportsNetRank: Network-based sports team ranking. *Large Scale Sports Analytics (SIGKDD).*

Pujahari A, Sisodia DS (2020) Aggregation of preference relations to enhance the ranking quality of collaborative filtering based group recommender system. *Expert Systems Appl.* 156:113476.

Rajkumar A, Agarwal S (2014) A statistical convergence perspective of algorithms for rank aggregation from pairwise data. *Proc. Internat. Conf. on Machine Learn.*, 118–126.

Rajkumar A, Agarwal S (2016) When can we rank well from comparisons of o (n\log (n)) nonactively chosen pairs? *Conf. on Learning Theory*, 1376–1401.

Richardson M, Prakash A, Brill E (2006) Beyond PageRank: Machine learning for static ranking. *Proc. 15th Internat. Conf. on World Wide Web*, 707–715.

Rogers AJ (1986) Modified lagrange multiplier tests for problems with one-sided alternatives. *J. Econometrics* 31(3):341–361.

Saaty TL (2003) Decision-making with the AHP: Why is the principal eigenvector necessary. *Eur. J. Oper. Res.* 145(1):85–91.

Shah N, Balakrishnan S, Bradley J, Parekh A, Ramchandran K, Wainwright M (2015) Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Artificial Intelligence and Statistics* (PMLR), 856–865.

Shah N, Balakrishnan S, Guntuboyina A, Wainwright M (2016) Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *Internat. Conf. on Machine Learning*, 11–20 (PMLR).

Shah NB, Wainwright MJ (2017) Simple, robust and optimal ranking from pairwise comparisons. *J. Machine Learn. Res.* 18(1):7246–7283.

Shapiro A (1988) Toward a unified theory of inequality constrained testing in multivariate analysis. *Internat. Statist. Rev.* 56(1):49–62.

Šidák Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* 62(318): 626–633.

Subochev A, Aleskerov F, Pislyakov V (2018) Ranking journals using social choice theory methods: A novel approach in bibliometrics. *J. Informetrics* 12(2):416–429.

Suh C, Tan VY, Zhao R (2017) Adversarial top-*k* ranking. *IEEE Trans. Inform. Theory* 63(4):2201–2225.

Susko E (2013) Likelihood ratio tests with boundary constraints using data-dependent degrees of freedom. *Biometrika* 100(4):1019–1023.

Swain TD, Chandler J, Backman V, Marcelino L (2017) Consensus thermotolerance ranking for 110 symbiodinium phylotypes: An exemplar utilization of a novel iterative partial-rank aggregation tool with broad application potential. *Functional Ecology* 31(1):172–183.

Talluri KT, Van Ryzin GJ (2006) *The Theory and Practice of Revenue Management*, vol. 68 (Springer Science & Business Media, New York).

Thurstone LL (1927) A law of comparative judgment. *Psych. Rev.* 34(4):273.

Tsybakov AB (2008) *Introduction to Nonparametric Estimation* (Springer Science & Business Media, New York).

Van de Geer S, Bühlmann P, Ritov Y, Dezeure R (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* 42(3):1166–1202.

Verdú S (1994) Generalizing the Fano inequality. *IEEE Trans. Inform. Theory* 40(4):1247–1251.

Vigna S (2016) Spectral ranking. *Network Sci.* 4(4):433–445.

Vojnovic M, Yun SY (2017) Parameter estimation for Thurstone choice models. Preprint, submitted April 29, https://arxiv.org/abs/1705.00136.

Volkovs MN, Zemel RS (2012) A flexible generative model for preference aggregation. *Proc. 21st Internat. Conf. on World Wide Web*, 479–488.

Wang J, Shah N, Ravi R (2020) Stretching the effectiveness of MLE from accuracy to bias for pairwise comparisons. *Proc. Internat. Conf. on Artificial Intelligence and Statist.*, 66–76 (PMLR).

Westfall PH, Young SS (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, vol. 279 (John Wiley & Sons, Hoboken, NJ).

Wolak FA (1989) Testing inequality constraints in linear econometric models. *J. Econometrics* 41(2):205–235.

Xia V, Jain K, Krishna A, Brinton CG (2018) A network-driven methodology for sports ranking and prediction. *Proc. 52nd Annual Conf. on Inform. Sci. and Systems* (IEEE, New York), 1–6.

Yu M, Gupta V, Kolar M (2019) Constrained high dimensional statistical inference. Preprint, submitted November 17, https://arxiv.org/abs/1911.07319.

Zhang CH, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. Royal Statist. Soc. Ser. B Statist. Methodology* 76(1):217–242.

Zhang L (2020) Ranking based on triple comparison. PhD thesis, Master's thesis, Texas A&M University, College Station, Texas.

Yue Liu is a third-year PhD student at Harvard University.

Ethan X. Fang is an assistant professor in the Department of Biostatistics and Bioinformatics at Duke University and is affiliated with the Rhodes Information Initiative and Fuqua School of Business.

Junwei Lu is an assistant professor in the Department of Biostatistics at Harvard T.H. Chan School of Public Health.