Ranking Inferences Based on the Top Choice of Multiway Comparisons

Jianqing Fan   Zhipeng Lou   Weichen Wang   Mengxin Yu [1]

Abstract

This paper considers ranking inference of n items based on the observed data on the top choice among M randomly selected items at each trial. This is a useful modification of the Plackett-Luce model for M-way ranking with only the top choice observed and is an extension of the celebrated Bradley-Terry-Luce model that corresponds to $M = 2$. Under a uniform sampling scheme in which any M distinguished items are selected for comparisons with probability p and the selected M items are compared L times with multinomial outcomes, we establish the statistical rates of convergence for underlying n preference scores using both ℓ_2-norm and ℓ_∞-norm, with the minimum sampling complexity. In addition, we establish the asymptotic normality of the maximum likelihood estimator that allows us to construct confidence intervals for the underlying scores. Furthermore, we propose a novel inference framework for ranking items through a sophisticated maximum pairwise difference statistic whose distribution is estimated via a valid Gaussian multiplier bootstrap. The estimated distribution is then used to construct simultaneous confidence intervals for the differences in the preference scores and the ranks of individual items. They also enable us to address various inference questions on the ranks of these items. Extensive simulation studies lend further support to our theoretical results. A real data application illustrates the usefulness of the proposed methods convincingly.

Keyword: Packett-Luce model, Maximum likelihood estimator, Asymptotic distribution, Rank confidence intervals, Gaussian multiplier bootstrap.

1 Introduction

The problem of ranking inference from pairwise comparisons or multiple partial rankings has drew significant attention in recent years, as the ranking problem has always played an important role in many applications such as individual choices in economics [Luce, 2012, McFadden, 1973], psychology [Thurstone, 1927, 2017], online and offline recommendations [Baltrunas et al., 2010, Li et al., 2019], and ranking of items such as journals [Stigler, 1994, Ji et al., 2022], websites [Dwork et al., 2001], colleges and universities [Avery et al., 2013, Caron et al., 2014], sports teams [Massey, 1997, Turner and Firth, 2012], election candidates [Plackett, 1975, Mattei and Walsh, 2013], and even alleles in genetics [Sham and Curtis, 1995]. Previously the ranking problem has mostly focused on parameter estimation and algorithm implementation; see for example Fürnkranz and Hüllermeier [2003], Negahban et al. [2012], Azari Soufiani et al. [2013], Maystre and Grossglauser [2015], Jang et al. [2018]. In addition, a large literature of empirical studies from the above areas of research focused on incorporating individual covariates for personalization [Turner and Firth, 2012, Li et al., 2019]. However, the ranking inference has only received more attention in the statistics community recently.

One of the most celebrated models for ranking problems is the Bradley-Terry-Luce (BTL) model. The model is frequently used to model pairwise comparisons. Specifically, consider a large collection of n items whose true ranking is determined by some unobserved preference scores θ_i^* for $i = 1, \ldots, n$, for example, qualities of products, reputations of education institutes or abilities of sports teams. The BTL model assumes that an individual or a random event ranks item i over j with probability $\mathbb{P}(\text{item i is preferred over j}) = e^{\theta_i^*}/(e^{\theta_i^*} + e^{\theta_j^*})$. The underlying choice axiom states that this probability does not depend on other items. Moreover, for simplicity, the BLT model does not account for data heterogeneity and treats the preference of each individual or event as purely independent Bernoulli. For the theoretical study, one may assume each pair (i, j) is compared with probability p, and once compared, they are compared for L times. Given the model and the collected data of pairwise comparisons, the statistical questions are straightforward: (a) What is the optimal statistical rate of convergence for estimating θ_i^* from the data? (b) What are the proper algorithms to achieve the optimal rate? (c) What is the asymptotic distribution of an estimator $\widehat{\theta}_i$ of θ_i^*? (d) Furthermore, how

can we carry out uncertainty quantification on ranks?

Questions (a) and (b) have been clearly addressed in Negahban et al. [2016] for the ℓ_2-loss of estimating $\pi^* = [\pi_1^*, \ldots, \pi_n^*]^\top$, where $\pi_i^* = e^{\theta_i^*}/\sum_i e^{\theta_i^*}$. They proposed the *rank centrality*, an efficient iterative spectral method. Chen et al. [2019] further studied the estimation of π^* under the ℓ_∞-norm. They delivered the key message that both the spectral method and the regularized maximum likelihood estimator (MLE) can achieve the optimal statistical convergence rate under the ℓ_∞-norm and the sparsest possible sampling regime ($p \gtrsim \log n/n$). Chen et al. [2020] further complements and refines the results of Chen et al. [2019] by concluding that the vanilla MLE without regularization can already achieve the optimal rate of convergence in both ℓ_2- and ℓ_∞- norms for estimating $\theta^* = [\theta_1^*, \ldots, \theta_n^*]^\top$ and the condition for exact recovery of the top-K ranking for MLE is weaker in constant than the spectral method.

Following (a) and (b), researchers also made recent progress on addressing (c) and (d) for the BLT model. Specifically, Han et al. [2020] made contributions to show the asymptotic normality of the MLE estimator with the sampling regime of $p \gtrsim (\log n)^{1/5}/n^{1/10}$, while Gao et al. [2021] fully revealed the asymptotic normality of both the MLE estimator and the spectral estimator under the assumption that $p \gtrsim (\log n)^{1.5}/n$, where 1.5 may be further improved to $1 + \delta$ for arbitrary $\delta > 0$. The authors showed that although the spectral method is optimal in the order of sample complexity, it is less efficient due to its larger asymptotic variance than the MLE. Despite a great theoretical contribution to the asymptotic normality, Gao et al. [2021] did not focus too much on the ranking inference problem (d) and only treated (d) with a crude confidence interval bound. In this work, we will refine the analysis of ranking inference for the MLE and therefore fill an important gap in the literature beyond the work of Gao et al. [2021]. The ranking inference for the BLT model is also studied by Liu et al. [2022], but with a rather strong assumption of $L \gtrsim n^2 \log^2 n$, that is, each compared pair must be compared more than Cn^2 times, which is barely possible in practical applications. In contrast, we only require $L \gtrsim \text{poly}(\log n)$ in this work to carry out our rank hypothesis testing.

Another more general model that extends pairwise comparison is the Plackett-Luce (PL) model, which assumes M-way full ranking. In the PL model, every time an individual provides a personal ranking on all given M items. Denote this full ranking as $i_1 > \cdots > i_M$. It can be understood as

$M-1$ independent events that i_1 is preferred over the set $\{i_1, \ldots, i_M\}$, i_2 is preferred over the set $\{i_2, \ldots, i_M\}$ and so forth. The model assumes that

$$\mathbb{P}(i_1 \succ \cdots \succ i_M) = \prod_{j=1}^{M-1} \left[e^{\theta_{i_j}^*} / \sum_{k=j}^{M} e^{\theta_{i_k}^*} \right].$$

Similar to the BLT model, each M-way comparison $\{i_1, \ldots, i_M\}$ is compared with probability p, and once compared, they are ranked for L times. In practice, L can be different. Nevertheless, for simplicity, in this work, we assume L is a shared quantity for each compared M item to ease the presentation and computation. The set of all M-way comparisons forms a comparison hyper-graph, which we will formally define later. When $M = 2$, the PL model reduces to the BLT model. Again we could ask the same four inference questions above for the PL model.

The PL model has garnered less attention due to the more complicated structure of multiple comparisons, although it fits in with more general and real settings, including for instance, multi-player games and personal preferences with multiple products. Research papers on the inference problems based on the PL model are relatively scarce. Maystre and Grossglauser [2015] introduced the iterative Luce spectral ranking method and showed that it converges to the MLE without providing any statistical rate. Jang et al. [2018] rigorously considered conditions for the exact recovery of top-K ranking and applied the spectral method to achieve this exact recovery under the assumption that $p \gtrsim (M-1)\sqrt{\log n / \binom{n-1}{M-1}}$. However, according to Cooley et al. [2016], the sparsest regime that we can have a connected hyper-graph is when $p \gtrsim \log n / \binom{n}{M-1}$. In this paper, we close this gap by showing that we can achieve optimal estimation error based on the MLE under the sparsest regime with $p \gtrsim \text{poly}(\log n) / \binom{n-1}{M-1}$, even in the harder situation than the traditional PL model when only top choices are observed from the M-way comparisons. Moreover, it is worth noting that all aforementioned works on M-way comparisons only focused on deriving first-order statistical rates of convergence, and the corresponding asymptotic distributions have rarely been investigated. To fill in this blank, we further derive the uncertainty quantification results in this sparest regime and apply them to study the practical ranking inference.

More specifically, instead of working on the PL model, we consider the partial-ranking case in which we only observe the top choice from the choice set $\{i_1, \ldots, i_M\}$. This is motivated from two

perspectives. On the one hand, many applications do not provide the full ranking among all the M selected items, and only the top choice is known. For example, a multi-player game may stop once we get the winner; a shopper may only choose the top item to purchase after presenting a set of products. On the other hand, theoretically speaking, general M with full ranking gives a likelihood function that does not provide much more insight beyond only considering the likelihood for the top choice. The theory will be more concise and intuitive regarding the role of M as we will see in later sections. For $M = 3$, we will also present the asymptotic normality for the PL model when the full ranking of $M = 3$ items is available. For the PL model with general $M > 3$, the MLE theory can be derived similarly, but due to its more tedious notational details, we decide to omit it.

Therefore our main focus of the paper is the multiway comparison model with only the top choice observed. Under this model, we apply the MLE method and analyze its statistical rates in both ℓ_2- and ℓ_∞- norms for estimating θ^* and show that they are *optimal* under the sparsest hyper-graph regime. In specific, when $M = \mathcal{O}(1)$, we achieve the same statistical rate as the PL model presented in Jang et al. [2018] but only requiring $p \gtrsim \mathrm{poly}(\log n)/\binom{n-1}{M-1}$, even if we only observe the top choice. This answers (a) and (b). Furthermore, to respond to the question in (c) for any $M \geqslant 2$, we establish the asymptotic normality of the MLE $\widehat{\theta}_i$, for all $i \in [n]$ in the sparest regime where the sampling probability satisfies $p \gtrsim \mathrm{poly}(\log n)/\binom{n-1}{M-1}$. Finally, for question (d), we address three detailed ranking inference problems: (i) constructing valid confidence intervals for the ranks of a set of items, (ii) testing if an item belongs to the top-K ranked items, (iii) providing a sure screening confidence set that contains all the top-K ranked items with high confidence. All of these are important inference questions in practice. For example, when high school seniors choose their colleges, they often care about the confidence interval for the ranks of a few universities, whether a certain university is within the top 50, and a list of universities that contain top 50 institutes with say a 95% confidence level. In order to complete these tasks, based on the asymptotic normality of the MLE, we propose a novel inference framework for ranking items through a sophisticated maximum pairwise difference statistic whose distribution is estimated via a valid Gaussian multiplier bootstrap [Chernozhukov et al., 2017, 2019]. The estimated distribution is then used to construct simultaneous confidence intervals for the differences in the preference scores and the ranks of individual items. They also enable us to address

the above inference questions on the ranks of these items.

Our main contributions of the work are summarized as follows. Firstly, we study the performance of the MLE on the more complicated general multiway comparison model with only top choice observed and show that MLE can achieve the optimal sample complexity under the sparsest possible regime. Secondly, we quantify the uncertainty of the MLE explicitly. Last but not least, we provide a general framework to conduct effective inference of ranks based on the Gaussian multiplier bootstrap and give answers to three crucial practical inference questions. Specifically, our proposed confidence intervals constructed for individual ranks are provably narrower than the high confidence Bonferroni adjustment in Gao et al. [2021].

1.1 Roadmap

In Section 2, we set up the model with some basic assumptions. Section 3 presents the performance of parameter estimation and asymptotic distribution of the MLE, while Section 4 details newly the proposed framework for constructing rank confidence intervals, rank testing statistics, and top-K sure screening confidence set. Section 5 contains comprehensive numerical studies to verify theoretical results and a real data example to illustrate the usefulness of our ranking inference methods. Finally we conclude the paper with some discussions in Section 6. All the proofs are deferred to the appendix.

1.2 Notation

Throughout this work, we use $[n]$ to denote the index set $\{1, 2, \cdots, n\}$. For any given vector $\mathbf{x} \in \mathbb{R}^n$ and $q \geqslant 0$, we use $\|\mathbf{x}\|_q = (\sum_{i=1}^n |x_i|^q)^{1/q}$ to represent the vector ℓ_q norm. For any given matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, we use $\| \cdot \|$ to denote the spectral norm of $\mathbf{X}$ and write $\mathbf{X} \succcurlyeq 0$ or $\mathbf{X} \preccurlyeq 0$ if $\mathbf{X}$ or $-\mathbf{X}$ is positive semidefinite. For event A, $\mathbb{I}_A$ denotes an indicator random variable which equals 1 if A is true and 0 otherwise. In addition, we let $\nabla L(\cdot), \nabla^2 L(\cdot)$ be the gradient and Hessian of a loss function $L(\cdot)$. For two positive sequences $\{a_n\}_{n\geqslant 1}$, $\{b_n\}_{n\geqslant 1}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \leqslant C \cdot b_n$ and we write $a_n = o(b_n)$ if $a_n/b_n \to 0$. Similarly we have $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$ if $a_n/b_n \geqslant c$ with some constant $c > 0$. We use $a_n = \Theta(b_n)$ (or $a_n \asymp b_n$) if

$a_n = \mathcal{O}(b_n)$ and $a_n = \Omega(b_n)$. Given n items, we use θ_i^* to indicate the underlying preference score of the i-th item. Define $r : [n] \to [n]$ as the rank operator on the n items which maps each item to its population rank based on the preference scores. We write the rank of the i-th item as r_i or $r(i)$. By default, we consider ranking from the largest score to the smallest score.

2 Multiway Comparison Model

We first introduce the formulation of the ranking problem for the multiway comparison model. The model consists of three key components.

- **Preference scores**: For a given group of n items, they are associated n preference scores

$$\theta^* = [\theta_1^*, \cdots, \theta_n^*]^\top,$$

  which are assumed to fall within a range,

$$\theta_i^* \in [\theta_L, \theta_U], \ \forall i \in [n].$$

  with the condition number $\kappa := \theta_U - \theta_L$. This paper considers the case where κ is a fixed constant independent of n. This represents a more challenging scenario in which all items under comparison have preference scores in the same order. Otherwise, we could apply a simple screening to easily differentiate obvious winners or losers and redo the analysis within only items with scores of the same order.

- **Comparison hypergraph**: Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a comparison hypergraph, where the vertex set $\mathcal{V} = \{1, 2, \cdots, n\}$ denotes the n items of interest. M different items $(i_1, \cdots, i_M)$ are compared if $(i_1, \cdots, i_M)$ falls within the edge set $\mathcal{E}$ of a M-way hypergraph. We assume a hyper-edge connecting any set $(i_1, \cdots, i_M)$ of size M with probability p. Let $A_{i_1 \cdots i_M}$ take value 1 if item set $(i_1, \cdots, i_M)$ is compared and 0 otherwise. Then it is a sequence of realizations from independent Bernoulli random variables with parameter p representing whether $(i_1, \cdots, i_M)$ is connected in the hypergraph. When $M = 2$, the hypergraph becomes the well-known Erdos-Renyi graph.

- **Multinomial Comparisons:** For each $(i_1, \cdots, i_M) \in \mathcal{E}$, we observe L independent comparisons among items in $\{i_1, \cdots, i_M\}$ and let $\{y_{i_1}^{(\ell)}, \cdots, y_{i_M}^{(\ell)}\}$ be the ℓ-th outcome of the comparison. If the most preferred item is i_k, then $y_{i_k}^{(\ell)} = 1$ and the others are zero. Thus, for each $\ell \in [L]$, $\{y_{i_k}^{(\ell)}, k \in [M]\}$ follows the multinomial distribution independently with probability

$$\left\{ p_{i_k} = \frac{e^{\theta_{i_k}^*}}{\sum_{j=1}^M e^{\theta_{i_j}^*}}, k \in [M] \right\}.$$

  Further define $y_{i_k} = \frac{1}{L} \sum_{l=1}^L y_{i_k}^{(\ell)}$. We also denote $y_{i_k}^{(\ell)}$ as $y_{i_k, (i_1, \cdots, i_{k-1}, i_{k+1}, \cdots, i_M)}^{(\ell)}$ when we need to emphasize that i_k is preferred over the remaining items $\{i_1, \cdots, i_{k-1}, i_{k+1}, \cdots, i_M\}$. But we prefer the shorter notation when it is clear from the context on the comparison set.

Throughout the paper, we consider $M = \mathcal{O}(1)$. This assumption is trivially satisfied in many practical applications. For example, in a multi-player game, the number of teams or contestants who compete with each other in every game is typically a fixed number or has a fixed upper bound. A student who faces the selection of education programs may only get offers from a few institutes. In a recommendation system such as an online shopping platform, due to the limited space of a webpage, only a fixed number of items can be exhibited on the first page, and a shopper may seldom turn to the second page to make the purchasing decision. Moreover, in all these examples, we typically only have access to the most preferred item instead of knowing the full ranking of all M items. Even when we observe the full ranking, it may not be trusted as much as the top preference due to the challenges to give full ranking of multiple items.

This paper aims to provide the statistical estimation and uncertainty quantification of the underlying scores of all items. More importantly, we study statistical inference for ranks, which is very much underdeveloped for the multiway comparison model.

3   Estimation and Uncertainty Quantification

In this section, we utilize the MLE to derive an estimator for the underlying scores $\{\theta_i^*\}_{i=1}^n$ of n items and establish the statistical convergence rates and asymptotic normality.

3.1 Statistical Estimation

The negative-log-likelihood function for our multiway comparison model is given by

$$\ell_n(\theta) = -\sum_{i_1 \neq \cdots \neq i_M} A_{i_1 \cdots i_M} \left[\sum_{k=1}^{M} y_{i_k} \log \left(\frac{e^{\theta_{i_k}}}{\sum_{j=1}^{M} e^{\theta_{i_j}}} \right) \right]. \tag{1}$$

Here the expression (1) is mainly for the purpose of theoretical analysis. For computation of the MLE, the first summation is over all trials of multiway comparisons and the second sum has only one non-vanishing term.

From the above likelihood function, θ^* is only identifiable up to additive shift. We assume $\mathbf{1}^\top \theta^* = 0$ for model identifiability. Thus, the parameter space for θ^* is the following $\Theta(\kappa)$ for some positive fixed constant $\kappa < \infty$, where

$$\Theta(\kappa) = \left\{ \theta \in \mathbb{R}^n : \max_{1 \leq m \leq n} \theta_m - \min_{1 \leq m \leq n} \theta_m \leq \kappa \text{ and } \mathbf{1}^\top \theta = 0 \right\}. \tag{2}$$

Thus, the MLE is given by

$$\widehat{\theta} = \underset{\mathbf{1}^\top \theta = 0}{\operatorname{argmin}} \ \ell_n(\theta). \tag{3}$$

The next theorem gives the rate of convergence for $\widehat{\theta}$.

Theorem 1. *If* $p \gtrsim poly(\log n)/\binom{n-1}{M-1}$, *then the MLE defined in* (3) *satisfies*

$$\|\widehat{\theta} - \theta^*\|_2^2 \lesssim \frac{n}{\binom{n-1}{M-1}pL}, \tag{4}$$

$$\|\widehat{\theta} - \theta^*\|_\infty^2 \lesssim \frac{\log n}{\binom{n-1}{M-1}pL}. \tag{5}$$

Theorem 1 presents the ℓ_2- and ℓ_∞- statistical convergence rates for $\widehat{\theta}$ when one chooses the most preferred item among M given items. This coincides with the best rate one can hope for if we ignore the logarithmic term. To understand this from the information perspective, note that the parameter θ_i appears only in the comparisons when item i is involved and the expected number of comparisons involving item i is $\binom{n-1}{M-1}pL$ for all $i \in [n]$. Therefore, the best estimation error of θ_i we can achieve is $\mathcal{O}((\binom{n-1}{M-1}pL)^{-1/2})$ for all $i \in [n]$, which matches the obtained bound for $\|\widehat{\theta} - \theta^*\|_\infty$ if we ignore

the logarithmic term. It is also worth mentioning that when $M = 2$, our model reduces to the well-known BTL model. The ℓ_2- and ℓ_∞- statistical rates also match those in estimating BTL model that are optimal up to logarithm terms [Chen et al., 2019, 2020].

We hope to point out that when $M = \mathcal{O}(1)$, our ℓ_2- and ℓ_∞- statistical rates are identical to those in Jang et al. [2018] for the M-way comparisons in the PL model via the spectral method. This reveals that only picking the top item, instead of full ranking of all M items, is sufficient to recover the underlying scores with same order of accuracy. In addition, note that the hypergraph with edge size of M items is connected with high probability when $p \gtrsim \log n / \binom{n}{M-1}$ by Cooley et al. [2016]; otherwise there will be isolated points and the corresponding items are never ranked. Our assumption on the sampling probability $p \gtrsim \mathrm{poly}(\log n) / \binom{n-1}{M-1}$ matches the lower bound on sampling probability up to logarithmic terms, whereas Jang et al. [2018] requires $p \gtrsim (\log n / \binom{n-1}{M-1})^{1/2}$, a order of magnitude larger than ours.

The next corollary provides the conditions on the recovery of the top-K items when there exists a gap between the scores of the true K-th and $(K + 1)$-th items.

Corollary 1. *Under the conditions of Theorem 1, if we have $\theta^* \in \Theta(\kappa)$ with $\theta^*_{(K)} - \theta^*_{(K+1)} \geqslant \Delta$, we are able to recover the true top-K items when the sample complexity satisfies*

$$\binom{n-1}{M-1} pL \gtrsim \Delta^{-2} \cdot \log n.$$

Here $\theta^*_{(i)}$ denotes the underlying score of the item with true rank i for $i \in [n]$.

This corollary follows directly from Theorem 1. Note that when $M = 2$, the requirement for sample complexity exactly reduces to that in Chen et al. [2019], which is minimax optimal up to logarithm factors.

3.2   Uncertainty Quantification

This subsection aims at providing uncertainty quantification of estimator $\widehat{\theta}$. We follow the idea proposed in Gao et al. [2021], depicting the asymptotic behavior of every element of $\widehat{\theta}$ via the likelihood function.

Before proceeding, we first separate out the likelihood terms involving the m-th entry θ_m. Fixing other components $\theta_{-m} = \{\theta_i : i \neq m\}$, we define

$$\ell_n^{(m)}(\theta_m \mid \theta_{-m}) = -M \sum_{(i_1 \neq \cdots \neq i_{M-1}) \neq m} A_{i_1 \cdots i_{M-1}m} \left[\sum_{k=1}^{M-1} y_{i_k} \log \left(\frac{e^{\theta_{i_k}}}{\sum_{j=1}^{M-1} e^{\theta_{i_j}} + e^{\theta_m}} \right) \right. \tag{6}$$
$$\left. + y_m \log \left(\frac{e^{\theta_m}}{\sum_{j=1}^{M-1} e^{\theta_{i_j}} + e^{\theta_m}} \right) \right].$$

Again this includes all terms in the likelihood function (1) that has the information of θ_m. Let $f^{(m)}(\theta_m \mid \theta_{-m})$ be the gradient of $\ell_n^{(m)}(\theta_m \mid \theta_{-m})$ w.r.t. θ_m, which is given by

$$f^{(m)}(\theta_m \mid \theta_{-m}) = M \sum_{(i_1 \neq \cdots \neq i_{M-1}) \neq m} A_{i_1 \cdots i_{M-1}m} \left\{ \frac{e^{\theta_m}}{\sum_{j=1}^{M-1} e^{\theta_{i_j}} + e^{\theta_m}} - y_m \right\}. \tag{7}$$

In addition, we also define $g^{(m)}(\theta_m \mid \theta_{-m})$ as the second derivative of $\ell_n^{(m)}(\theta_m \mid \theta_{-m})$ w.r.t. θ_m, which is given by

$$g^{(m)}(\theta_m \mid \theta_{-m}) = M \sum_{(i_1 \neq \cdots \neq i_{M-1}) \neq m} A_{i_1 \cdots i_{M-1}m} \left\{ \sum_{j=1}^{M-1} \frac{e^{\theta_m + \theta_{i_j}}}{(\sum_{j=1}^{M-1} e^{\theta_{i_j}} + e^{\theta_m})^2} \right\}. \tag{8}$$

Then, to maximize the likelihood, $\widehat{\theta}_m$ must be the minimzer of $\ell_n^{(m)}(\theta \mid \widehat{\theta}_{-m})$. By Taylor expansion of $\ell_n^{(m)}(\theta \mid \widehat{\theta}_{-m})$, a good proxy to $\widehat{\theta}_m - \theta_m^*$ is its score function $-\frac{f^{(m)}(\theta_m^* \mid \widehat{\theta}_{-m})}{g^{(m)}(\theta_m^* \mid \widehat{\theta}_{-m})}$, which is approximately the same as $-\frac{f^{(m)}(\theta_m^* \mid \theta_{-m}^*)}{g^{(m)}(\theta_m^* \mid \theta_{-m}^*)}$. This leads us to consider the heuristic expression

$$\widehat{\theta}_m - \theta_m^* = -\frac{f^{(m)}(\theta_m^* \mid \theta_{-m}^*)}{g^{(m)}(\theta_m^* \mid \theta_{-m}^*)} + \delta_m, \tag{9}$$

where we expect δ_m to be of smaller order. The asymptotic distribution of $\widehat{\theta}_m - \theta_m^*$ will then follow from that of $-\frac{f^{(m)}(\theta_m^* \mid \theta_{-m}^*)}{g^{(m)}(\theta_m^* \mid \theta_{-m}^*)}$. The following theorem makes the above heuristic discussion rigorous.

Theorem 2. *If $p \gtrsim poly(\log n)/\binom{n-1}{M-1}$, the MLE defined in (3) enjoys*

$$\widehat{\theta}_m - \theta_m^* = -\frac{f^{(m)}(\theta_m^* \mid \theta_{-m}^*)}{g^{(m)}(\theta_m^* \mid \theta_{-m}^*)} + \delta_m,$$

for all $m \in [n]$ with $\|\delta\|_\infty = o((1/\binom{n-1}{M-1}pL)^{1/2})$ where $\delta = (\delta_1, \cdots, \delta_n)$. In addition,

$$\rho_m(\theta)(\widehat{\theta}_m - \theta_m^*) \to N(0,1),$$

11

for all $m \in [n]$ with

$$\rho_m(\theta) = \left[\frac{L}{(M-1)!} \sum_{(i_1 \neq \cdots \neq i_{M-1}) \neq m} A_{i_1 \cdots i_{M-1} m} \left\{ \sum_{j=1}^{M-1} \frac{e^{\theta_m + \theta_{i_j}}}{(\sum_{j=1}^{M-1} e^{\theta_{i_j}} + e^{\theta_m})^2} \right\} \right]^{1/2}, \qquad (10)$$

for both $\theta \in \{\theta^*, \widehat{\theta}\}$.

Theorem 2 presents the asymptotic distribution for every element of $\widehat{\theta}$, by approximating $\widehat{\theta}_m - \theta_m^*$ by the score function $-\frac{f^{(m)}(\theta_m^* \mid \theta_{-m}^*)}{g^{(m)}(\theta_m^* \mid \theta_{-m}^*)}$. Although a series of works have studied the ranking estimation problem under pairwise or general multiway comparison [Maystre and Grossglauser, 2015, Jang et al., 2018], results on uncertainty quantification are still very much underdeveloped. To our best knowledge, existing literature only has results on quantifying the uncertainty of the BTL model [Gao et al., 2021]. We take one step further to unravel the uncertainty of the preference score estimator $\widehat{\theta}$ presented in Theorem 2 for the more general M-way comparison with $M \geqslant 2$.

We comment on the connections of our results with several related literature. Firstly, compared with Liu et al. [2022] who study the asymptotic distribution of their estimator via the Lagrangian debiasing method, we have no requirement on the number of comparisons L for establishing the asymptotic distribution whereas they require $L \gtrsim n^2$. Secondly, compared with Han et al. [2020], who established the asymptotic results in the regime when $p \gtrsim 1/n^{1/10}$, we allow much sparser regime for the comparison hypergraph, namely, we allow $p \gtrsim \text{poly}(\log n)/\binom{n-1}{M-1}$. This matches the sparsest sampling regime up to logarithm terms. In addition, when $M = 2$, our model reduces to the BTL model whose uncertainty quantification is provided by Gao et al. [2021] with the requirement of $p \gtrsim \text{poly}(\log n)/n$.

As we mentioned in Section 3.1, the statistical rate achieved under our top choice multiway comparison model is the same as that in estimating the conventional PL model up to logarithm terms. Hence the difference in the asymptotic behaviors under these two models lies in their different asymptotic variances. In specific, for any fixed M, when the underlying scores $\{\theta_i^*\}_{i=1}^n$ are of the same order, we sacrifice a factor of order M in our asymptotic variance compared with using the PL model. We want to emphasize that our analyzing techniques can be easily extended to the scenario where one ranks the top-k items for any $k \leqslant M$ given the M items. This includes the PL model as a particular case with $k = M$. We present a formal theorem on the asymptotic normality of the MLE for the

PL model with $M = 3$ in Appendix B.8. The corresponding theorem for the PL model with general $M = \mathcal{O}(1)$ can be derived similarly, and we leave the details to the interested readers. As we argued, we often only have access to top choices in reality. More importantly, studying the top choice in multiway comparison already conveys our key messages clearly without excessive technicality.

4   Ranking Inferences

In many real applications, people have access to ranking-related data and problems. For instance, multiple sources such as US News and Times Higher Education publish global university rankings every year; sports team rankings are an essential part of our everyday chat; companies try to hire the best candidates based on the evaluations of their interviewers. Most current practical usage of ranks only involves estimating preference scores and displaying the estimated ranks. However, we often lack the tools to address basic inference questions such as the following.

- Is Team A indeed significantly stronger than Team B? Can one build an efficient confidence interval for the ranks of a few items of interest?

- Is the offer from a university truly a good choice to accept? How can we tell whether an item is among the top-K ranking with high confidence?

- How many candidates should a company hire to ensure all the best candidates are selected? How do we get a confidence set of items to ensure the screening of the top-K items?

In this section, we hope to address all these critical statistical inference questions for ranks. To this end, we introduce a novel inference framework for the population ranks $\{r_m\}_{m \in \mathcal{M}}$ simultaneously, where $\mathcal{M}$ is any subset of $[n]$ of interest and $r_m, m \in [n]$ is the true rank in descending order of the m-th item according to its underlying preference score θ_m^*.

4.1 Two-Sided Confidence Intervals

In this subsection, we propose a general framework for constructing two-sided confidence intervals for ranks based on the MLE estimator $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)^\top$ given in (3). To construct the simultaneous confidence intervals for the ranks, a direct approach is to derive the asymptotic distribution of the corresponding empirical ranks $\{\widehat{r}_m\}_{m \in \mathcal{M}}$ and figure out the critical value. However, as $\widehat{r}_m$ is an integer which depends on all $\widehat{\theta}_1, \ldots, \widehat{\theta}_n$ for any $m \in \mathcal{M}$, this is nontrivial.

By exploiting the mutual relationship between the scores and the ranks, we observe that constructing confidence intervals for the ranks can be reduced to constructing simultaneous confidence intervals for the pairwise differences between the population scores, whose empirical counterpart's distribution is easier to depict. Therefore, to circumvent the difficulty of deriving the distribution of ranks directly, we work on the statistics of estimated scores $\{\widehat{\theta}_m\}_{m \in [n]}$ instead. Next, we will illustrate the key intuition of our approach via the following example.

Example 1 (Simultaneous rank confidence intervals). Let $\mathcal{M} = \{m\}$ for some $1 \leqslant m \leqslant n$ be the item of interest. We consider constructing the $(1 - \alpha) \times 100\%$ confidence interval for the population rank r_m, where $\alpha \in (0, 1)$ is a prescribed significance level. Let $\{[\mathcal{C}_L(k, m), \mathcal{C}_U(k, m)]\}_{k \neq m}$ denote the simultaneous confidence intervals of $\{\theta_k^* - \theta_m^*\}_{k \neq m}$ such that with probability at least $1 - \alpha$, we have all $\theta_k^* - \theta_m^* \in [\mathcal{C}_L(k, m), \mathcal{C}_U(k, m)]$. Observe that $\mathcal{C}_U(k, m) < 0$ (resp. $\mathcal{C}_L(k, m) > 0$) implies $\theta_k^* < \theta_m^*$ (resp. $\theta_k^* > \theta_m^*$). Counting the number of items ranked lower than item m by using confidence upper bounds and the number of items ranked above item m by using the confidence lower bounds, we can get a confidence interval for r_m. In other words,

$$\mathbb{P}\left(1 + \sum_{k \neq m} \mathbb{I}\{\mathcal{C}_L(k, m) > 0\} \leqslant r_m \leqslant n - \sum_{k \neq m} \mathbb{I}\{\mathcal{C}_U(k, m) < 0\}\right) \geqslant 1 - \alpha. \tag{11}$$

This yields a $(1 - \alpha) \times 100\%$ confidence interval for r_m.

We now formally introduce the procedure to construct the confidence intervals for multiple ranks $\{r_m\}_{m \in \mathcal{M}}$ simultaneously. Motivated by Example 1, the key step is to construct the simultaneous

confidence intervals for the pairwise score differences $\{\theta_k^* - \theta_m^*\}_{m \in \mathcal{M}, k \neq m}$. Towards this end, define

$$\mathcal{T}_\eta = \max_{m \in \mathcal{M}} \max_{k \neq m} \left| \frac{\sqrt{L}\{\widehat{\theta}_k - \widehat{\theta}_m - (\theta_k^* - \theta_m^*)\}}{\eta_{mk}} \right|. \tag{12}$$

Here $\{\eta_{mk}\}_{1 \leqslant k \neq m \leqslant n}$ is a sequence of positive normalizing constants introduced to account for different scales of $\{\widehat{\theta}_k - \widehat{\theta}_m - (\theta_k^* - \theta_m^*)\}_{1 \leqslant k \neq m \leqslant n}$. A natural choice of $\{\eta_{mk}\}_{\{1 \leqslant k \neq m \leqslant n\}}$ is the uniform consistent estimators of the standard deviations $\{\widehat{\sigma}_{mk}\}$ of $\{\widehat{\theta}_k - \widehat{\theta}_m\}_{\{1 \leqslant k \neq m \leqslant n\}}$, where

$$\widehat{\sigma}_{mk}^2 = \frac{M(M-1)!}{g^{(m)}(\widehat{\theta}_m | \widehat{\theta}_{-m})} + \frac{M(M-1)!}{g^{(k)}(\widehat{\theta}_k | \widehat{\theta}_{-k})}. \tag{13}$$

For any $\alpha \in (0, 1)$, let $\zeta_{1-\alpha}$ denote a consistent estimate of the $(1 - \alpha)$ quantile of the asymptotic distribution of $\mathcal{T}_\eta$ such that

$$\mathbb{P}(\mathcal{T}_\eta \leqslant \zeta_{1-\alpha}) \to 1 - \alpha. \tag{14}$$

Then, motivated by Example 1, the $(1 - \alpha) \times 100\%$ simultaneous confidence intervals for $\{r_m\}_{m \in \mathcal{M}}$ are given by $\{[\mathcal{R}_m^\diamond(\eta, \zeta_{1-\alpha}), \mathcal{R}_m^\sharp(\eta, \zeta_{1-\alpha})]\}_{m \in \mathcal{M}}$, where

$$\mathcal{R}_m^\diamond(\eta, \zeta_{1-\alpha}) = 1 + \sum_{k \neq m} \mathbb{I}\left\{ \widehat{\theta}_k - \widehat{\theta}_m > \frac{\eta_{mk}}{\sqrt{L}} \times \zeta_{1-\alpha} \right\},$$

$$\mathcal{R}_m^\sharp(\eta, \zeta_{1-\alpha}) = n - \sum_{k \neq m} \mathbb{I}\left\{ \widehat{\theta}_k - \widehat{\theta}_m < -\frac{\eta_{mk}}{\sqrt{L}} \times \zeta_{1-\alpha} \right\}. \tag{15}$$

In view of (11) and (14), the constructed simultaneous confidence intervals satisfy that

$$\mathbb{P}\left(\cap_{m \in \mathcal{M}} \{\mathcal{R}_m^\diamond(\eta, \zeta_{1-\alpha}) \leqslant r_m \leqslant \mathcal{R}_m^\sharp(\eta, \zeta_{1-\alpha})\} \right) \geqslant 1 - \alpha - o(1).$$

As noted in (14), the key step for constructing the confidence interval of ranks of interest is to pick the critical value $\zeta_{1-\alpha}$. In the next subsection, we propose to estimate $\zeta_{1-\alpha}$ via the Gaussian multiplier bootstrap procedure.

4.2 Gaussian Multiplier Bootstrap

We now present our proposed estimate of the critical value $\zeta_{1-\alpha}$ for $\mathcal{T}_\eta$, which satisfies (14). We begin by introducing some notations and definitions. For each $1 \leqslant \ell \leqslant L$, define

$$\xi_{m\ell} = \frac{M}{g^{(m)}(\theta_m^* | \theta_{-m}^*)} \sum_{(i_1 \neq \cdots \neq i_{M-1}) \neq m} A_{i_1 \cdots i_{M-1} m} \left\{ \frac{e^{\theta_m^*}}{\sum_{j=1}^{M-1} e^{\theta_{i_j}^*} + e^{\theta_m^*}} - y_m^{(\ell)} \right\}. \tag{16}$$

In addition, let $\xi_m := \frac{f^{(m)}(\theta_m^*|\theta_{-m}^*)}{g^{(m)}(\theta_m^*|\theta_{-m}^*)} = \frac{1}{L}\sum_{\ell=1}^L \xi_{m\ell}$ for each $m \in [n]$. Theorem 2 ensures that $\widehat{\theta}_m - \widehat{\theta}_k - (\theta_m^* - \theta_k^*) \approx \frac{1}{L}\sum_{\ell=1}^L (\xi_{k\ell} - \xi_{m\ell})$ uniformly for all $1 \leqslant k \neq m \leqslant n$. Consequently, we obtain

$$\mathcal{T}_\eta \approx \max_{m\in\mathcal{M}} \max_{k\neq m} \frac{1}{\sqrt{L}} \left| \sum_{\ell=1}^L \left(\frac{\xi_{k\ell} - \xi_{m\ell}}{\eta_{mk}} \right) \right|. \tag{17}$$

Notice that $\{\xi_{k\ell} - \xi_{m\ell}\}_{1\leqslant \ell \leqslant L}$ is a sequence of i.i.d. zero-mean random variables with conditional variance $\sigma_{mk}^2 = \mathrm{Var}(\xi_{k\ell} - \xi_{m\ell}|A)$ where $\mathrm{Var}(\cdot \,|\, A)$ denotes the conditional variance given a fixed comparison hypergraph, i.e.

$$A = \{A_{i_1\cdots i_M}\} \text{ where recall } A_{i_1\cdots i_M} = 1 \text{ if } (i_1, \cdots, i_M) \in \mathcal{E}. \tag{18}$$

In what follows, we also write $\mathcal{T} = \mathcal{T}_\eta$ for simplicity.

Since the dimension of the random vectors $\{(\xi_{k\ell} - \xi_{m\ell})_{m\in\mathcal{M},k\neq m}\}_{1\leqslant \ell \leqslant L}$ is $(n-1)|\mathcal{M}|$ and is usually much larger than L (we only require $\mathrm{poly}(\log n)/L \to 0$ in Theorem 3 below), the classical multivariate central limit theorem cannot be utilized here to derive the asymptotic distribution of $\mathcal{T}$. Instead, we shall invoke the high dimensional Gaussian approximation result [Chernozhukov et al., 2017, 2019] which quantifies the distance between the distribution functions of $\mathcal{T}$ and its Gaussian analogue. Nevertheless, the asymptotic distribution of $\mathcal{T}$ still depends on the unknown population scores $\{\theta_m^*\}_{m\in[n]}$ and the covariance structure of the random vector $(\xi_{m\ell} - \xi_{k\ell})_{m\in\mathcal{M},k\neq m}$. To approximate the asymptotic distribution of $\mathcal{T}$, we apply a practically feasible Gaussian multiplier bootstrap procedure [Chernozhukov et al., 2017, 2019]. First define the empirical version of $\xi_{m\ell}$ as follows,

$$\widehat{\xi}_{m\ell} = \frac{M}{g^{(m)}(\widehat{\theta}_m|\widehat{\theta}_{-m})} \sum_{(i_1\neq\cdots\neq i_{M-1})\neq m} A_{i_1\cdots i_{M-1}m} \left\{ \frac{e^{\widehat{\theta}_m}}{\sum_{j=1}^{M-1} e^{\widehat{\theta}_{i_j}} + e^{\widehat{\theta}_m}} - y_m^{(l)} \right\}.$$

Let $\omega_1, \ldots, \omega_L \in \mathbb{R}$ be i.i.d. standard normal random variables. Then, in view of (17), the bootstrap counterpart of $\mathcal{T}$ is defined as

$$\mathcal{G} = \max_{m\in\mathcal{M}} \max_{k\neq m} \frac{1}{\sqrt{L}} \left| \sum_{\ell=1}^L \left(\frac{\widehat{\xi}_{k\ell} - \widehat{\xi}_{m\ell}}{\widehat{\sigma}_{mk}} \right) \omega_\ell \right|.$$

For any $\alpha \in (0,1)$, let $\mathcal{G}_{1-\alpha}$ denote the $(1-\alpha)$-th quantile of $\mathcal{G}$, that is,

$$\mathcal{G}_{1-\alpha} = \inf\{z \in \mathbb{R} : \mathbb{P}(\mathcal{G} \leqslant z|Y, A) \geqslant 1-\alpha\}. \tag{19}$$

16

Here $\mathbb{P}(\cdot \,|\, Y, A)$ denotes the conditional probability where all randomness from $Y = \{y_{i_k}^{(\ell)} : i_k \in \{i_1, \cdots, i_M\}, \ell \in [L]\}$ and A defined in (18) is fixed.

Theorem 3. *Assume* $poly(\log n)/L \to 0$. *Then, under the conditions of Theorem 2, we have*

$$\left| \mathbb{P} \left\{ \max_{m \in \mathcal{M}} \max_{k \neq m} \left| \frac{\sqrt{L}\{\widehat{\theta}_k - \widehat{\theta}_m - (\theta_k^* - \theta_m^*)\}}{\widehat{\sigma}_{mk}} \right| > \mathcal{G}_{1-\alpha} \right\} - \alpha \right| \to 0. \qquad (20)$$

Theorem 3 indicates that the estimated critical value $\mathcal{G}_{1-\alpha}$ from the Gaussian multiplier bootstrap indeed controls the significance level of the simultaneous confidence intervals for ranks in $\mathcal{M}$ to the prespecified level α. To make the inference valid, we do require L to grow faster than $\mathrm{poly}(\log n)$. So our current proposal does not work for say $L = 1$, where each comparison set is only ranked by one single individual. Fortunately, this requirement is not too restrictive and can be satisfied by many practical datasets such as the more than 30 datasets on elections, Netflix movie ranking, and sports competitions hosted on the PrefLib website [Mattei and Walsh, 2013].

Remark 1. Recall the definition of $\rho_m(\theta)$ in (10) for each $m \in [n]$. In the context of pairwise comparison where $M = 2$, Gao et al. [2021] proposed a $(1-\alpha) \times 100\%$ confidence interval $[\widetilde{\mathcal{R}}_m^\diamond, \widetilde{\mathcal{R}}_m^\sharp]$ for the population rank r_m, where

$$\widetilde{\mathcal{R}}_m^\diamond = 1 + \sum_{k \neq m} \mathbb{I} \left\{ \widehat{\theta}_k - \widehat{\theta}_m > \frac{z_{1-\alpha/2}}{\rho_m(\widehat{\theta})} + \frac{(1+c_0)\sqrt{2 \log n}}{\rho_k(\widehat{\theta})} \right\},$$

$$\widetilde{\mathcal{R}}_m^\sharp = n - \sum_{k \neq m} \mathbb{I} \left\{ \widehat{\theta}_m - \widehat{\theta}_k > \frac{z_{1-\alpha/2}}{\rho_m(\widehat{\theta})} + \frac{(1+c_0)\sqrt{2 \log n}}{\rho_k(\widehat{\theta})} \right\}.$$

Here $z_{1-\alpha/2}$ is the $(1-\alpha)$-th quantile of the standard normal distribution, $c_0 > 0$ is an arbitrary small and fixed positive constant. Note that the uncertainty of $\widehat{\theta}_k$ for all $k \neq m$ is controlled by its high-probability Bonferroni bound, leading to the $\sqrt{2 \log n}$ multiplier (the worst lower or upper bound). Denote $\widetilde{\eta}_{mk} = \frac{z_{1-\alpha/2}}{\rho_m(\widehat{\theta})(1+c_0)\sqrt{2 \log n}} + \frac{1}{\rho_k(\widehat{\theta})}$ for each $1 \leqslant k \neq m \leqslant n$. Then it is straightforward that

$$\mathbb{P}\left(r_m \in [\widetilde{\mathcal{R}}_m^\diamond, \widetilde{\mathcal{R}}_m^\sharp] \right) \geqslant \mathbb{P}\left(\max_{k \neq m} \left| \frac{\widehat{\theta}_k - \widehat{\theta}_m}{\widetilde{\eta}_{mk}} \right| > (1+c_0)\sqrt{2 \log n} \right) \geqslant 1 - \alpha - o(1). \qquad (21)$$

The length of the confidence interval is given by

$$\widetilde{\mathcal{L}}_{1-\alpha} = \widetilde{\mathcal{R}}_m^{\sharp} - \widetilde{\mathcal{R}}_m^{\diamond} = n - 1 - \sum_{k \neq m} \mathbb{I}\left\{ \left| \frac{\widehat{\theta}_k - \widehat{\theta}_m}{\widetilde{\eta}_{mk}} \right| > (1 + c_0)\sqrt{2 \log n} \right\}. \tag{22}$$

In contrast, following (12)–(15), if we set the normalization parameter η_{mk} to be $\widetilde{\eta}_{mk}$, our confidence interval for r_m is given by $[\mathcal{R}_m^{\diamond}(\widetilde{\eta}, \mathcal{G}_{1-\alpha}^{\diamond}), \mathcal{R}_m^{\sharp}(\widetilde{\eta}, \mathcal{G}_{1-\alpha}^{\diamond})]$ with length

$$\mathcal{L}_{1-\alpha} = \mathcal{R}_m^{\sharp}(\widetilde{\eta}, \mathcal{G}_{1-\alpha}^{\diamond}) - \mathcal{R}_m^{\diamond}(\widetilde{\eta}, \mathcal{G}_{1-\alpha}^{\diamond}) = n - 1 - \sum_{k \neq m} \mathbb{I}\left\{ \left| \frac{\widehat{\theta}_k - \widehat{\theta}_m}{\widetilde{\eta}_{mk}} \right| > \mathcal{G}_{1-\alpha}^{\diamond} \right\}, \tag{23}$$

where $\mathcal{G}_{1-\alpha}^{\diamond}$ is the $(1 - \alpha)$-th quantile of $\mathcal{G}^{\diamond} = \max_{k \neq m} \frac{1}{L} \left| \sum_{\ell=1}^{L} \left(\frac{\widehat{\xi}_{k\ell} - \widehat{\xi}_{m\ell}}{\widetilde{\eta}_{mk}} \right) \omega_\ell \right|$. Following the same proof of Theorem 3, we can similarly obtain

$$\mathbb{P}\left(\max_{k \neq m} \left| \frac{\widehat{\theta}_k - \widehat{\theta}_m}{\widetilde{\eta}_{mk}} \right| > \mathcal{G}_{1-\alpha}^{\diamond} \right) \to 1 - \alpha.$$

However, it is easy to verify that the length of our proposed confidence interval in (23) is shorter than that given by Gao et al. [2021] in (22). In detail,

$$\mathbb{P}\left(\widetilde{\mathcal{L}}_{1-\alpha} \geqslant \mathcal{L}_{1-\alpha} \right) \geqslant \mathbb{P}\left(\mathcal{G}_{1-\alpha}^{\diamond} \leqslant (1 + c_0)\sqrt{2 \log n} \right) \to 1.$$

In conclusion, for any $m \in [n]$, our confidence interval of r_m will be narrower than that of Gao et al. [2021] with probability tending to 1. In addition, from the above remark, the confidence interval proposed by Gao et al. [2021] using high-probability Bonferroni adjustment can also be easily extended to the situation of $M > 2$, once we plug-in the updated formula for $\rho_m(\widehat{\theta})$ and $\rho_k(\widehat{\theta})$ as in (10) in the normalization parameter $\widetilde{\eta}_{mk}$. However, their Bonferroni confidence interval cannot be directly extended to simultaneous inference of a few ranks whereas in our proposal we have the freedom to choose any interested set $\mathcal{M}$ of size more than one. In our simulation and real data analyses below, we will demonstrate empirically that there is no big difference on using $\widetilde{\eta}_{mk}$ or $\widehat{\sigma}_{mk}$ as the normalization parameter. Thus, we will see the length of confidence intervals for both choices are strictly smaller than the Bonferroni confidence interval constructed in Gao et al. [2021].

4.3 One-Sided Confidence intervals

In this subsection, we provide details on constructing (simultaneous) one-sided intervals for population ranks, and utilizing the one-sided intervals to resolve two more important questions, namely top-K placement testing and sure screening of top-K candidates. This further illustrates the wide applicability of our methodology.

For one-sided intervals, the overall procedure is similar to constructing two-sided confidence intervals. Specifically, let

$$\mathcal{G}^{\circ} = \max_{m \in \mathcal{M}} \max_{k \neq m} \frac{1}{\sqrt{L}} \sum_{\ell=1}^{L} \left(\frac{\widehat{\xi}_{m\ell} - \widehat{\xi}_{k\ell}}{\widehat{\sigma}_{mk}} \right) \omega_{\ell} \tag{24}$$

where $\omega_1, \ldots, \omega_L \in \mathbb{R}$ are as before i.i.d. standard normal random variables. Correspondingly, let $\mathcal{G}^{\circ}_{1-\alpha}$ be its $(1-\alpha)$-th quantile. Under the conditions of Theorem 3, it follows that

$$\left| \mathbb{P} \left\{ \max_{m \in \mathcal{M}} \max_{k \neq m} \frac{\sqrt{L}\{\widehat{\theta}_k - \widehat{\theta}_m - (\theta_k^* - \theta_m^*)\}}{\widehat{\sigma}_{mk}} > \mathcal{G}^{\circ}_{1-\alpha} \right\} - \alpha \right| \to 0.$$

Then the $(1-\alpha) \times 100\%$ simultaneous left-sided confidence intervals for $\{r_m\}_{m \in \mathcal{M}}$ are given by

$$[\mathcal{R}_m^{\diamond}, n] = \left[1 + \sum_{k \neq m} \mathbb{I} \left\{ \widehat{\theta}_k - \widehat{\theta}_m > \frac{\widehat{\sigma}_{mk}}{\sqrt{L}} \times \mathcal{G}^{\circ}_{1-\alpha} \right\}, n \right], \quad m \in \mathcal{M}. \tag{25}$$

We next show the usefulness of the one-sided confidence intervals with two examples. The first one is on testing whether an item of interest lies in the top-K placement. The mathematical framework is given in Example 2.

Example 2 (Testing top-K placement)**.** Let $\mathcal{M} = \{m\}$ for some $1 \leqslant m \leqslant n$ and let $K \geqslant 1$ be an prescribed positive integer. We are interested in testing whether m-th item is among the top-K ranked items. So the hypotheses under consideration is

$$H_0 : r_m \leqslant K \quad \text{versus} \quad H_1 : r_m > K. \tag{26}$$

Based on the one-sided confidence interval $[\mathcal{R}_m^{\diamond}, n]$ in (25), for any $\alpha \in (0, 1)$, a level α test for (26) is simply given by

$$\psi_{m,K} = \mathbb{I}\{\mathcal{R}_m^{\diamond} > K\}.$$

Proposition 1. *Under the conditions of Theorem 3, we have* $\mathbb{P}(\psi_{m,K} = 1 | H_0) \leqslant \alpha + o(1)$. *In addition,* $\mathbb{P}(\psi_{m,K} = 1 | H_1) \to 1$ *holds when* $\theta^*_{(K)} - \theta^*_m \gtrsim \sqrt{\frac{\log n}{\binom{n-1}{M-1}pL}}$, *where* $\theta^*_{(K)}$ *denotes the underlying score of the item with true rank* K.

Proposition 1 summarizes the size and power of the above test $\psi_{m,K}$. From Proposition 1, we are able to control the type-I error effectively below α under the null hypothesis. Moreover, when the alternative holds, the power of the test goes rapidly to one as long as the score difference is larger than the threshold of order $\mathcal{O}(\sqrt{\log n / (\binom{n-1}{M-1}pL)})$.

Besides testing for the top-K placement, our second crucial example is on constructing a screened candidate set that contains the top-K items with high probability. This is particularly useful in college candidate admission or company hiring decision. Oftentimes, a university or a company would like to design certain admission or hiring policy with the high-probability guarantee of the sure screening of true top-K candidates. We formulate this rigorously in Example 3.

Example 3 (Candidate admission and sure screening). Let $r : [n] \to [n]$ denote the rank operator on n items which maps each item to its population rank and $\mathcal{K} = \{r^{-1}(1), \ldots, r^{-1}(K)\}$ be the top-K ranked items. Our goal is to select a set of candidates $\widehat{\mathcal{I}}_K$ which contains the top-K candidates with a prescribed probability, that is,

$$\mathbb{P}\left(\mathcal{K} \subset \widehat{\mathcal{I}}_K\right) \geqslant 1 - \alpha, \tag{27}$$

for some $\alpha \in (0, 1)$. Let $\mathcal{M} = [n]$ and $\{[\mathcal{R}^\diamond_m, n]\}_{m \in [n]}$ denote the corresponding $(1 - \alpha) \times 100\%$ simultaneous left-sided confidence intervals in (25). Notice that $\mathcal{R}^\diamond_m > K$ implies that $r_m > K$ for each $m \in [n]$. Hence a natural choice of $\widehat{\mathcal{I}}_K$ which satisfies (27) would be

$$\widehat{\mathcal{I}}_K = \{1 \leqslant m \leqslant n : \mathcal{R}^\diamond_m \leqslant K\}.$$

In practice, the candidates are more likely to be admitted based on their empirical ranks $\{\hat{r}(1), \ldots, \hat{r}(n)\}$. In this scenario, we then seek the minimal number $D \geqslant 1$ such that

$$\mathbb{P}\left(\mathcal{K} \subset \{\hat{r}^{-1}(1), \ldots, \hat{r}^{-1}(D)\}\right) \geqslant 1 - \alpha.$$

$$
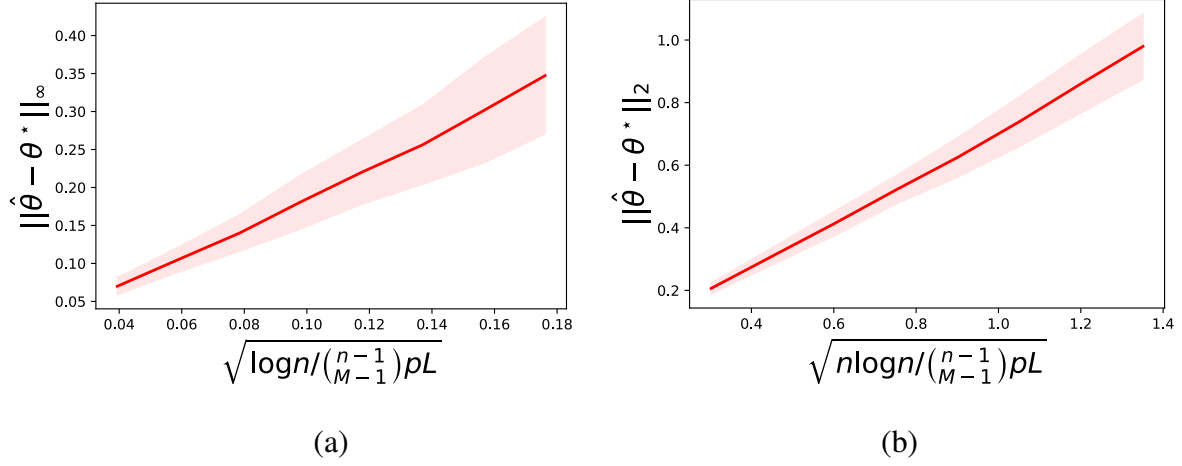\text{(a)} \qquad\qquad\qquad\qquad \text{(b)}
$$

Figure 1: ℓ_∞- and ℓ_2- statistical errors of the MLE $\widehat{\theta}$ against the theoretical rate when p varies. The solid lines represent the averaged statistical errors of 200 repetitions and the light areas denote the standard deviations.

To estimate D, with a slight abuse of notation, set $\widehat{\sigma}_{mk} = 1$ for all $1 \leqslant k \neq m \leqslant n$ in (24) and denote the corresponding $(1 - \alpha) \times 100\%$ simultaneous left-sided confidence intervals as $\{[\widehat{\mathcal{R}}_m^\diamond, n]\}$. Then, similar to $\widehat{\mathcal{I}}_K$, our estimator for D is defined by

$$
\widehat{D} = \max\left\{ 1 \leqslant m \leqslant n : \widehat{\mathcal{R}}_m^\diamond \leqslant K \right\},
$$

where $\widehat{\mathcal{R}}_1^\diamond \leqslant \widehat{\mathcal{R}}_2^\diamond \leqslant \ldots \leqslant \widehat{\mathcal{R}}_n^\diamond$ maintains the same ranking as the empirical ranks.

In Example 3, we constructed sure screening set for top-K candidates. Although many works studied sure screening property of the high-dimensional regression coefficients [Fan and Lv, 2008, Fan and Song, 2010, Zhu et al., 2011, Li et al., 2012, Barut et al., 2016, Wang and Leng, 2016, Fan et al., 2022], the study on the sure screening property of population ranks is much less explored and to our best knowledge, our procedure is the first one to have concrete theoretical guarantee.

5 Numerical Studies

In this section, we first conduct numerical studies via synthetic data to demonstrate our theoretical results in finite samples in Section 5.1 and to illustrate the effectiveness of our proposed confidence
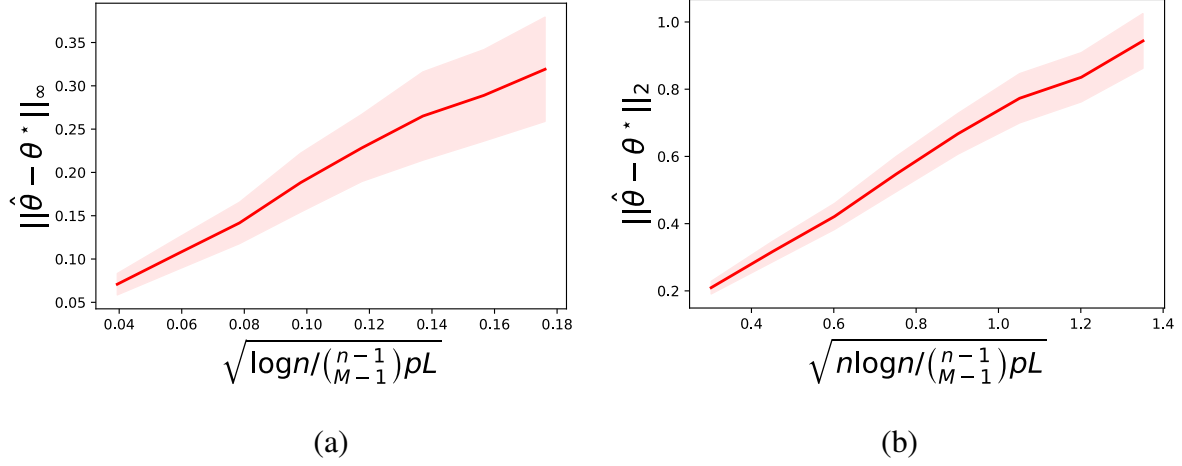
Figure 2: ℓ_∞- and ℓ_2- statistical errors of the MLE $\widehat{\theta}$ against the theoretical rate when L varies. The remaining captions are the same as those in Figure 1.

interval methodology in Section 5.2. We then analyze a real data example in Section 5.3.

5.1 Synthetic Data Analysis

To confrim the correctness of our theorem in finite samples, we will verify the statistical convergence rate, the asympototic normality and the effectiveness of Gaussian multiplier bootstrap in this subsection.

Statistial rates of convergence. We first validate the statistical rates of our MLE estimator $\widehat{\theta}$ in both ℓ_∞- and ℓ_2-norms. In the first simulation, we fix $n = 60, M = 3, L = 20$ and let p vary such that $\sqrt{\log n / \binom{n-1}{M-1} pL}$ takes uniform grids from 0.04 to 0.18. Meanwhile, we generate every entry of the true θ^* independently from Uniform$[2,4]$. We then record the ℓ_∞- and ℓ_2- statistical errors of $\widehat{\theta}$ to θ^* by solving the MLE given in (1). The average errors together with the standard deviations of 200 repetitions for each p are displayed in Figure 1. In the second simulation, we investigate the effects of L on these statistical rates. In this scenario, we fix $n = 60, M = 3, p = 0.05$ and let L vary such that $\sqrt{\log n / \binom{n-1}{M-1} pL}$ takes uniform grids from 0.04 to 0.18. The remaining procedures are the same as above and the results are shown in Figure 2. Clearly, we observe from Figures 1 and 2, the statistical rates are proportional to their theoretical rates, as indicated by the overall linear pattern.

22

These simulation results lend further support to the theoretical results in Theorem 1.

Asymptotic normality. Next, we investigate the uncertainty quantification of the MLE estimator $\widehat{\theta}$. Here we fix $n = 60, M = 3$ and choose L, p from $\{5, 10, 20\}$ and $\{0.008, 0.015, 0.03\}$ respectively, which results in 9 combinations. For each combination, the true θ^* is generated independently from Uniform$[2, 4]$ for 500 times. We record the empirical distributions of the standardized $\widehat{\theta}_1$, that is $\rho_1(\widehat{\theta})(\widehat{\theta}_1 - \theta_1^*)$, of these 500 repetitions, and check its normality via histograms, presented in Figure 3. From Figure 3, we observe that the empirical distribution of $\widehat{\theta}_1$ is well approximated by the standard Gaussian distribution, especially when we have a larger p and L. This is consistent with the theoretical results in Theorem 2.

Gaussian multiplier bootstrap. Finally, we validate the Gaussian approximation results discussed in Section 4. We let $n = 60, M = 3, L = 80, p = 0.05$ and investigate the distribution of $\mathcal{T}$ in (12) with $\mathcal{M} = \{1\}$. By Theorem 3, we have $\mathbb{E}[\mathbb{I}\{\mathcal{T} > \mathcal{G}_{1-\alpha}\}] = \mathbb{P}(\mathcal{T} > \mathcal{G}_{1-\alpha}) \to \alpha$. We then verify this with various $\alpha \in \{0.05, 0.10, \cdots, 0.90\}$. For every α, we bootstrap 300 times to compute the critical value $\mathcal{G}_{1-\alpha}$, and repeat this whole procedure 500 times to calculate $\widehat{\mathbb{P}}(\mathcal{T} > \mathcal{G}_{1-\alpha}) = \frac{1}{500} \sum_{b=1}^{500} \mathbb{I}\{\mathcal{T}^b > \mathcal{G}_{1-\alpha}^b\}$ where $\mathcal{T}^b$ and $\mathcal{G}_{1-\alpha}^b$ are the pairwise score difference statistic and the corresponding bootstrap critical value of the b-th repetition. Note that in each repetition, the true θ^* is generated independently from Uniform$[2, 4]$ as before. Note that $\widehat{\mathbb{P}}(\mathcal{T} \leqslant \mathcal{G}_{1-\alpha})$ is the empirical coverage probability for the pairwise score difference statistic $\mathcal{T}$. Figure 4 gives the so-called PP-plot, which shows $\widehat{\mathbb{P}}(\mathcal{T} > \mathcal{G}_{1-\alpha})$ against the theoretical significance level. From Figure 4, it the clear that the empirical probability match well with the theoretical ones. Especially, when we apply the significant level of $\alpha = 0.05$, the empirical exceptional probability is indeed around 0.05.

5.2 Confidence Interval

Below we provide numerical studies for validating our proposed framework for confidence interval construction in Section 4. Throughout this subsection, we use $n = 60, M = 3, L = 80$ and let p vary in $\{0.05, 0.10, 0.15\}$. In addition, the entries of $\theta \in \mathbb{R}^{60}$ are generated from the uniform grids from 4 to 2, i.e. $\theta_i = 4 - (i - 1)/30$ for $i = 1, \ldots, 60$. For every experiment, we conduct the bootstrap for 500
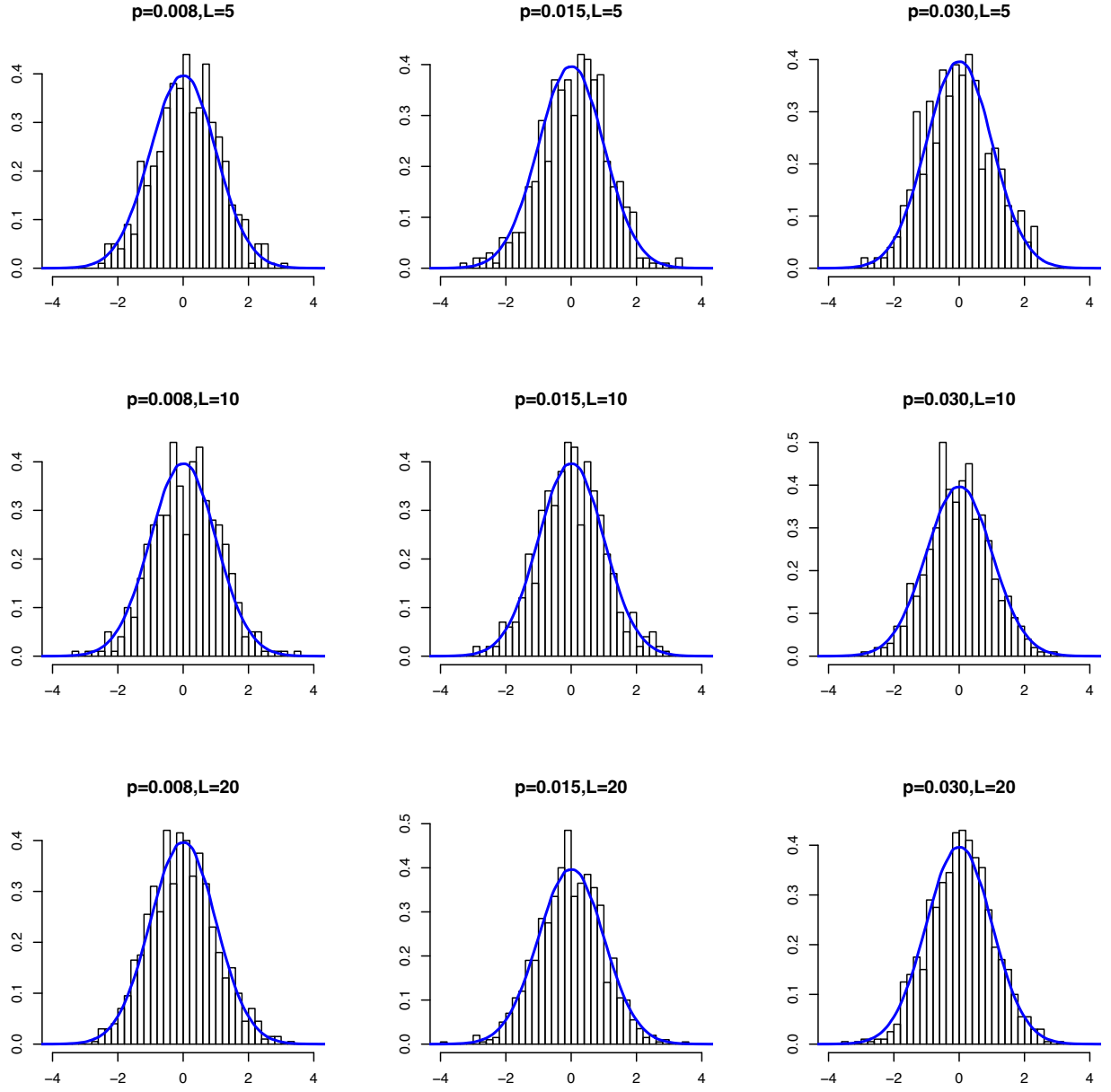
23

Figure 3: Histogram of $\rho_1(\widehat{\theta})(\widehat{\theta}_1 - \theta_1^*)$ against standard Gaussian distribution. The blue curve denotes the standard Gaussian distribution. The experiments were repeated for 500 times for each combination of $L \in \{5, 10, 20\}$ and $p \in \{0.008, 0.015, 0.03\}$.
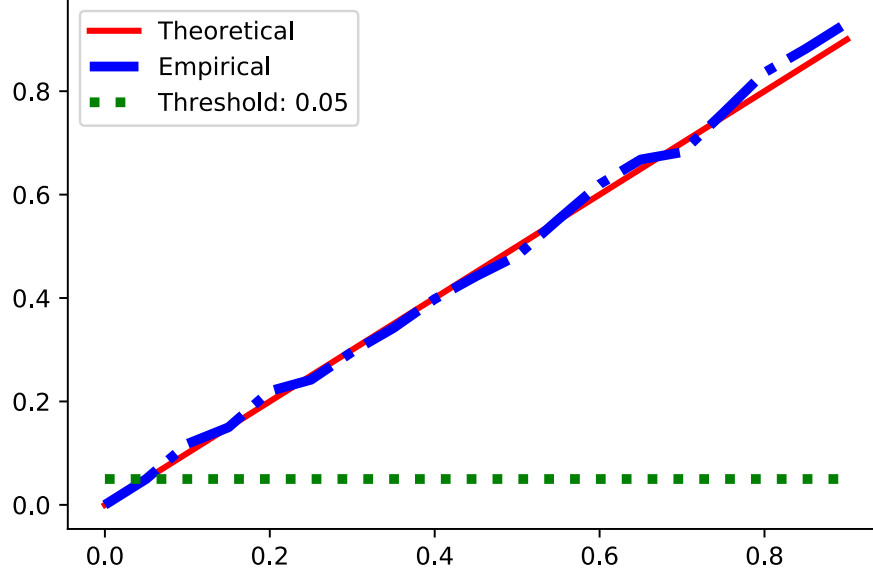
Figure 4: PP-plot of empirical probability $\widehat{\mathbb{P}}(\mathcal{T} > \mathcal{G}_{1-\alpha})$ of $\mathcal{T}$ given in (12) with $\mathcal{M} = \{1\}$ against theoretical significance level α. The red solid and blue dash-dotted lines represent theoretical and empirical probabilities respectively. The green dotted line represents the the case when significance is chosen to be 0.05.

times to compute the critical value according to $\alpha = 0.05$ and further repeat the entire procedure for 500 times.

Two-sided confidence intervals. We construct the confidence interval (CI) for the 10-th item ($\mathcal{M} = \{10\}$) using our method in Section 4 and the Bonferroni correction proposed in Gao et al. [2021] respectively. Concretely, we will compare the following 3 confidence intervals: (i) our bootstrap CI $[\mathcal{R}_{10}^{\diamond}(\widehat{\sigma}, \mathcal{G}_{1-\alpha}), \mathcal{R}_{10}^{\sharp}(\widehat{\sigma}, \mathcal{G}_{1-\alpha})]$ with the normalization $\widehat{\sigma}_{mk}$, (ii) our bootstrap CI $[\mathcal{R}_{10}^{\diamond}(\widetilde{\eta}, \mathcal{G}_{1-\alpha}^{\diamond}), \mathcal{R}_{10}^{\sharp}(\widetilde{\eta}, \mathcal{G}_{1-\alpha}^{\diamond})]$ but with the normalization $\widetilde{\eta}_{mk}$ ($c_0 = 1$), (iii) the Bonferroni CI $[\widetilde{\mathcal{R}}_{10}^{\diamond}, \widetilde{\mathcal{R}}_{10}^{\sharp}]$ given by Gao et al. [2021] extended to $M = 3$ in this simulation (see discussions after Remark 1). For each choice of the above confidence intervals, we report: (a) EC(θ) – the empirical coverage probability $\widehat{\mathbb{P}}(\mathcal{T} \leqslant \mathcal{G}_{1-\alpha})$ for $\mathcal{T}$, which is also the overall empirical coverage probability for all the score differences $\theta_k^* - \theta_{10}^*, \forall k \neq 10$, (b) EC($r$) – the empirical coverage of the confidence interval constructed for rank r_{10}, and furthermore (c) Length – the length of the CI for rank r_{10} which equals $\mathcal{R}_{10}^{\sharp} - \mathcal{R}_{10}^{\diamond}$, where $[\mathcal{R}_{10}^{\diamond}, \mathcal{R}_{10}^{\sharp}]$ is any of the

CI	$[\mathcal{R}_{10}^{\diamond}(\hat{\sigma}, \mathcal{G}_{1-\alpha}), \mathcal{R}_{10}^{\sharp}(\hat{\sigma}, \mathcal{G}_{1-\alpha})]$			$[\mathcal{R}_{10}^{\diamond}(\tilde{\eta}, \mathcal{G}_{1-\alpha}^{\diamond}), \mathcal{R}_{10}^{\sharp}(\tilde{\eta}, \mathcal{G}_{1-\alpha}^{\diamond})]$			$[\widetilde{\mathcal{R}}_{10}^{\diamond}, \widetilde{\mathcal{R}}_{10}^{\sharp}]$		
	$EC(\theta)$	$EC(r)$	Length	$EC(\theta)$	$EC(r)$	Length	$EC(\theta)$	$EC(r)$	Length
$p = 0.05$	0.950	1.000	5.590	0.948	1.000	5.716	1.000	1.000	10.290
$p = 0.10$	0.940	1.000	3.604	0.956	1.000	3.686	1.000	1.000	6.962
$p = 0.15$	0.950	1.000	2.886	0.948	1.000	2.928	1.000	1.000	5.476

Table 1: Empirical coverages and lengths of two-sided confidence intervals (CI). For any choice of CI, "EC(θ)" and "EC(r)" denote the empirical coverages of confidence intervals for score differences and ranks. "Length" denotes the length of the CI for r_{10}. The first two CI's use our bootstrap method in Section 4 and only differ in the normalization parameters used. The third CI is based on the Bonferroni method of Gao et al. [2021]. All numbers are averaged over 500 replications.

above CI in (i)-(iii).

The results of the empirical coverages for score differences and ranks and the length of confidence intervals are summarized in Table 1. Table 1 reveals that the empirical coverage probability for score differences, no matter which normalization it uses, is approximately 0.95, which is consistent with our theory in Section 4. However, the CI for the rank of the 10-th item is more conservative since the rank must be an integer, so we see the empirical coverage of the confidence intervals for the rank stays at one for all cases. In comparison, the confidence interval via Gao et al. [2021]'s method is even more conservative as the empirical coverages of the confidence intervals for θ are already one and the lengths of their confidence intervals are much wider than ours.

One-sided confidence intervals. Next we validate the testing performance of the test $\psi_{m,K} = \mathbb{I}\{\mathcal{R}_{m}^{\diamond} > K\}$ for (26). In this experiment, we choose $K = 10$ in (26) and by default we would like to use the normalization parameter $\hat{\sigma}_{mk}$ in constructing $\mathcal{R}_{m}^{\diamond}$ as in (25). Consider $m \in \{K-2, \cdots, K+5\}$. We computed the proportion of rejection for each given m. If $m \leqslant K$ the null hypothesis is true and the proportion is approximately the sizes of the test, whereas if the alternative is true, the proportion is approximately the power of the test. In addition, when the null holds, we also calculated the size of the

	Null holds: $r_m \leqslant K$			Alternative holds: $r_m > K$				
m	$K-2$	$K-1$	K	$K+1$	$K+2$	$K+3$	$K+4$	$K+5$
$\theta_m^* - \theta_K^*$	$2/30$	$1/30$	0	$-1/30$	$-2/30$	$-3/30$	$-4/30$	$-5/30$
$p = 0.05$	0 (0.036)	0 (0.038)	0 (0.050)	0.008	0.144	0.444	0.822	0.986
$p = 0.10$	0 (0.043)	0 (0.045)	0 (0.054)	0.032	0.372	0.896	0.993	1
$p = 0.15$	0 (0.042)	0 (0.038)	0 (0.046)	0.094	0.624	0.984	1	1

Table 2: Sizes and powers of the test $\psi_{m,K} = \mathbb{I}\{\mathcal{R}_m^\diamond > K\}$ for testing hypothesis in (26). The numbers inside the bracket indicate the sizes of testing score differences as we discussed in the text. The numbers outside the bracket represent the sizes or powers of directing testing the rank, which is our goal here. All displayed numbers are averaged over 500 replications.

test $\mathbb{I}\{\max_{k \neq m} \sqrt{L}\{\widehat{\theta}_k - \widehat{\theta}_m - (\theta_k^* - \theta_m^*)\}/\widehat{\sigma}_{mk} > \mathcal{G}_{1-\alpha}^\circ\}$ for the one-sided hypotheis testing problem $H_0 : \theta_m^* - \theta_k^* \geqslant (k-m)/30$ for all $k \neq m$ for testing score differences.

The results are presented in Table 2. We observe from this table that when the alternative holds and true rank increases, the power of our test rise rapidly to 1. On the other hand, when the null hypothesis holds, we control the test size around $\alpha = 0.05$ if we are testing the score differences and the rank test becomes more conservative and get size equal to zero, which can be a good feature in practice.

Uniform one-sided confidence intervals. In the candidate admission Example 3, to guarantee the sure screening property, we need to build uniform one-sided coverage for all items, i.e. $\mathcal{M} = [n]$. Following Example 3, once we have the one-sided confidence interval for all items, the sure screening set is given by $\widehat{\mathcal{I}}_K = \{1 \leqslant m \leqslant n : \mathcal{R}_m^\diamond \leqslant K\}$. In this simulation, we choose $K = \{5, 10, 15\}$. In Table 3, we report the average length of $\widehat{\mathcal{I}}_K$ and empirical coverages of the confidence intervals for score differences and for true top-K ranks over 500 replications. It turns out that in the simulation the length of the sure screening confidence set is no more than $K + 4$ even when the sampling probability p is as small as 0.05 since the true scores are well-separated. Similar to what we have previously seen, the rank empirical coverage probabilities "EC(r)" are again all one for $K = 5, 10, 15$, indicating that

	EC(θ)	EC(r)	$K = 5$	$K = 10$	$K = 15$
$p = 0.05$	0.964	1.000	8.81	13.82	18.89
$p = 0.10$	0.966	1.000	7.55	12.49	17.54
$p = 0.15$	0.962	1.000	6.98	12.00	17.00

Table 3: Empirical coverages and lengths of sure screening confidence set for Example 3. "EC(θ)" represents the empirical coverage of confidence intervals for score differences. Note EC(θ) does not depend on K. "EC(r)" denotes the empirical coverage of the confidence intervals for true top K items. For all $K \in \{5, 10, 15\}$, we see EC(r) = 1, so we collapse them into just one column. Other numbers are the lengths of the sure screening confidence set $\widehat{\mathcal{I}}_K$. All displayed numbers are averaged over 500 replications.

the sure screening set is already conservative. Finally, since we use $\mathcal{M} = [n]$, the empirical coverage "EC(θ)" for score differences is independent of K. From Table 3, for different p, this coverage via the Gaussian multiplier bootstrap is approximately 0.95 as expected.

5.3 Real Data Analysis

In this subsection, we analyze a real dataset to corroborate the practical effectiveness of our proposed methodology and its associated theoretical guarantees. We choose to use the relatively simple *Jester Dataset* [Goldberg et al., 2001] which contains ratings for 100 jokes from $73,421$ users and is available on the website of https://goldberg.berkeley.edu/jester-data/. Among all the users, $14,116$ rated all 100 jokes. Our analyses are based upon these users who rated all jokes for simplicity.

Results with $M = 3$. As for data generation, we first synthesize an uniform-hypergraph with edge sampling probability $p = 0.05$ and consider the setting $M = 3$, namely, any 3 jokes are chosen for comparisons with probability $p = 0.05$. For any selected tuple, we randomly select $L = 80$ rankings from those $14,116$ users who ranked all jokes, and observe the top rankings for the selected tuple. For CI construction, we follow our methodology discussed in Section 4. We construct both two-sided

and one-sided $1 - \alpha$ confidence intervals for top-15 ranked items with $\alpha = 0.05$. We summarize our inference results in Table 4.

From Table 4, we observe that our bootstrap method using $\widehat{\sigma}_{mk}$ (TC_1) or $\widetilde{\eta}_{mk}$ (TC_2) does not make too much difference. However, it is worth mentioning that our confidence intervals constructed using either of these two normalization parameters are strictly better than the confidence interval constructed via the Bonferroni method (TC_3) of Gao et al. [2021]. Furthermore, we also build the one-sided confidence intervals for each individual item, denotes as OC, and the uniform one-sided confidence intervals for all items together, denoted as UOC, which is wider than OC due to the overall control. OC can be used to conduct the hypothesis testing in (26). For example, if we care to test whether a joke is within the top-3 funniest in this real data, we will reject the hypothesis from the 7-th ranked item. If we test whether a joke is within the top-10 best, we will reject the hypothesis from the 12-th ranked item. UOC can be used to generate the sure screening confidence set in Example 3. For example, a set that contains all the top-5 jokes with high probability should include the first 9 jokes in total.

Results with different M's. According to our asymptotic distribution, the asymptotic variance is of the order $M/[\binom{n-1}{M-1} p_M]$ when we assume $\{\theta_i\}_{i \in [n]}$ are in the same order. Now we let M vary and choose p_M such that $M/[\binom{n-1}{M-1} p_M]$ is a fixed number. Specifically, we fix $p_2 = 0.3$, we will have the following $\{p_M\}_{M \in \{3, \cdots 6\}}$: $p_3 = 9 \times 10^{-3}, p_4 = 3.7 \times 10^{-4}, p_5 = 1.9 \times 10^{-5}, p_6 = 1.2 \times 10^{-6}$. We summarize the two-sided confidence intervals for jokes with ID in $\{10, 30, 50, 70, 90\}$ for each M in Table 5. We observe from Table 5, when we increase M but keep the same $M/\binom{n-1}{M-1} p_M$, we still obtain confidence intervals with comparable length for any given item. We also observe that it allows a much smaller sampling probability p_M when M is large to construct confidence intervals of the same significance level.

As noted before, for each given sampling probability p, the effective number of samples is very different for different M. Therefore, we compare the inference results only for the adjacent M and $M + 1$ with the same p, denoted as $p_{M,M+1}$. Specifically, we pre-select 5 jokes and compute their two-sided confidence intervals based on M-way and $(M + 1)$-way comparisons with a fixed sampling probability $p_{M,M+1}, M \in \{2, 3, 4, 5\}$. The results are presented in Table 6. We observe that for a fixed $p_{M,M+1}$, the two-sided confidence intervals with $M + 1$ are much narrower than those with M.

Moreover, for a given M, if we increase p, the lengths of confidence intervals also become much smaller. Both of these conclusions are due to the increase of sample size in both scenarios.

6   Conclusion and Discussion

This paper studies the ranking inference problem based on multiway comparisons. Unlike the conventional Plackett-Luce model [Plackett, 1975], which models the entire multiway rankings, we considered the more general case of only observing the top choices. Such a model serves as an extension of the famous Bradley-Terry-Luce model and modifies the Plackett-Luce model in a useful and practical direction. Theoretically, under the sparsest uniform sampling regime, we proposed to estimate the underlying preference scores via the MLE and established its *optimal* ℓ_2- and ℓ_∞- statistical rates. This closed the gap of achieving the optimal convergence with a practical algorithm under the sparest comparison hypergraph. Moreover, little has been done to quantify the asymptotic uncertainty of an estimator for the multiway comparisons. To our best knowledge, our work is the first to derive and justify the asymptotic distribution of the MLE for the underlying preference scores in the top-choice multiway comparison model. We should emphasize again that our theoretical contributions are highly nontrivial as the justification for general M is quite mathematically involved. More importantly, we proposed a novel inference framework for building confidence intervals for ranks, which are provably narrower than the confidence intervals with high-probability Bonferroni correction in Gao et al. [2021]. This framework is valuable in solving outstanding inference questions, including testing top-K placement and constructing sure screening confidence sets.

There are a few future directions to improve our work further. Firstly, we studied the ranking problem based on a uniform comparison hypergraph. That is, each comparison is made among M items. It would be interesting to consider the mixed-size choice set where one may observe different numbers of items for each comparison. The Plackett-Luce model can be viewed as choosing the best item from M items and then choosing the best from the remaining $M-1$ items and so on. For general mixed-size comparisons, some analyses need to be modified, and it is interesting to study whether the ranking inference results in this paper can be generalized. Secondly, although the convergence

optimality and asymptotic normality results require no assumption on the number of comparisons L, when we studied the ranking inference in Section 4, we needed L to satisfy $L \gtrsim \text{poly}(\log n)$ in order to establish the theoretical guarantee for the Gaussian multiplier bootstrap. It remains open whether we can further relax the condition to $L \gtrsim 1$ or even allow $L = 1$ to conduct effective ranking inferences. Thirdly, it would be interesting to see if some covariate information can be added to the analysis. In reality, the ranking is sometimes conducted together with item features or expert opinions. It is another exciting topic to study how we may incorporate these pieces of additional information into ranking inferences. Lastly, the time-varying effect of ranks may also be worth further investigation regarding time series of ranks. Over time, we may see underlying scores jump to a different level. Ranking inferences to detect the change point is another promising direction. Overall, we still see many challenges in inference for ranks under various settings, which calls for more research on ranking inference methodologies.

References

C. N. Avery, M. E. Glickman, C. M. Hoxby, and A. Metrick. A revealed preference ranking of US colleges and universities. *The Quarterly Journal of Economics*, 128(1):425–467, 2013.

H. Azari Soufiani, W. Chen, D. C. Parkes, and L. Xia. Generalized method-of-moments for rank aggregation. *Advances in Neural Information Processing Systems*, 26, 2013.

L. Baltrunas, T. Makcinskas, and F. Ricci. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 119–126, 2010.

E. Barut, J. Fan, and A. Verhasselt. Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277, 2016.

F. Caron, Y. W. Teh, and T. B. Murphy. Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics*, 8(2):1145–1181, 2014.

P. Chen, C. Gao, and A. Y. Zhang. Partial recovery for top-k ranking: Optimality of mle and sub-optimality of spectral method. *arXiv preprint arXiv:2006.16485*, 2020.

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

V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.

V. Chernozhukov, D. Chetverikov, K. Kato, and Y. Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *arXiv preprint arXiv:1912.10529*, 2019.

O. Cooley, M. Kang, and C. Koch. Threshold and hitting time for high-order connectedness in random hypergraphs. *the electronic journal of combinatorics*, pages 2–48, 2016.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

J. Fan, Z. Lou, and M. Yu. Are latent factor regression and sparse regression adequate? *arXiv preprint arXiv:2203.01219*, 2022.

J. Fürnkranz and E. Hüllermeier. Pairwise preference learning and ranking. In *European conference on machine learning*, pages 145–156. Springer, 2003.

C. Gao, Y. Shen, and A. Y. Zhang. Uncertainty quantification in the Bradley-Terry-Luce model. *arXiv preprint arXiv:2110.03874*, 2021.

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

R. Han, R. Ye, C. Tan, and K. Chen. Asymptotic theory of sparse Bradley–Terry model. *The Annals of Applied Probability*, 30(5):2491–2515, 2020.

M. Jang, S. Kim, and C. Suh. Top-k rank aggregation from m-wise comparisons. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):989–1004, 2018.

P. Ji, J. Jin, Z. T. Ke, and W. Li. Meta-analysis on citations for statisticians. *Manuscript*, 2022.

G. Li, H. Peng, J. Zhang, and L. Zhu. Robust rank correlation based screening. *The Annals of Statistics*, 40(3): 1846–1877, 2012.

H. Li, D. Simchi-Levi, M. X. Wu, and W. Zhu. Estimating and exploiting the impact of photo layout: A structural approach. *Available at SSRN 3470877*, 2019.

Y. Liu, E. X. Fang, and J. Lu. Lagrangian inference for ranking problems. *Operations Research*, 2022.

R. D. Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

K. Massey. Statistical models applied to the rating of sports teams. *Bluefield College*, 1077, 1997.

N. Mattei and T. Walsh. Preflib: A library for preferences http://www.preflib.org. In *International conference on algorithmic decision theory*, pages 259–270. Springer, 2013.

L. Maystre and M. Grossglauser. Fast and accurate inference of Plackett–Luce models. *Advances in neural information processing systems*, 28, 2015.

D. McFadden. Conditional logit analysis of qualitative choice behavior. 1973.

S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. *Advances in neural information processing systems*, 25, 2012.

S. Negahban, S. Oh, and D. Shah. Rank centrality: Ranking from pairwise comparisons. *Operations research*, 65(1):266–287, 2016.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

P. Sham and D. Curtis. An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Annals of human genetics*, 59(3):323–336, 1995.

S. M. Stigler. Citation patterns in the journals of statistics and probability. *Statistical Science*, pages 94–108, 1994.

L. L. Thurstone. The method of paired comparisons for social values. *Journal of Abnormal Psychology*, 21(4), 1927.

L. L. Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 2017.

H. Turner and D. Firth. Bradley-Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software*, 48:1–21, 2012.

X. Wang and C. Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):589–611, 2016.

L.-P. Zhu, L. Li, R. Li, and L.-X. Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475, 2011.

Rank	Joke ID	Score	TC_1	TC_2	TC_3	OC	UOC
1	89	0.89	[1,2]	[1,2]	[1,2]	[1,100]	[1,100]
2	50	0.85	[1,2]	[1,2]	[1,2]	[1,100]	[1,100]
3	27	0.73	[3,8]	[3,8]	[3,9]	[3,100]	[3,100]
4	36	0.69	[3,8]	[3,8]	[3,10]	[3,100]	[3,100]
5	35	0.69	[3,9]	[3,9]	[3,10]	[3,100]	[3,100]
6	29	0.67	[3,9]	[3,9]	[3,11]	[3,100]	[3,100]
7	32	0.67	[3,9]	[3,9]	[3,11]	[4,100]	[3,100]
8	62	0.66	[3,10]	[3,10]	[3,11]	[4,100]	[3,100]
9	54	0.64	[4,11]	[5,11]	[3,14]	[5,100]	[5,100]
10	53	0.60	[8,12]	[8,12]	[4,16]	[9,100]	[7,100]
11	49	0.57	[9,15]	[9,15]	[6,16]	[10,100]	[9,100]
12	68	0.54	[10,16]	[10,16]	[9,21]	[11,100]	[9,100]
13	72	0.52	[11,16]	[11,16]	[9,21]	[11,100]	[10,100]
14	66	0.52	[11,16]	[11,16]	[9,21]	[11,100]	[10,100]
15	69	0.51	[11,16]	[11,16]	[10,21]	[11,100]	[10,100]

Table 4: Confidence intervals for ranks of jokes in the *Jester Dataset. The columns* Rank*,* Score *and* Joke ID *denote the estimated rank, estimated MLE score and the corresponding joke IDs, respectively.* TC_1, TC_2 *denote our bootstrap two-sided confidence intervals constructed using* $\widehat{\sigma}_{mk}$ *and* $\widetilde{\eta}_{mk}$ ($c_0 = 1$) *as the normalization parameters.* TC_3 *is constructed based on the Bonferroni method of* Gao et al. [2021]*. Furthermore,* OC *represents the one-sided confidence intervals for each individual item and* UOC *denotes the uniform one-sided confidence intervals for all items together, which are used to generate the sure screening confidence set in Example* 3*.*

Joke ID	$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M = 6$
10	[24,51]	[23,50]	[26,48]	[21,45]	[24,43]
30	[63,89]	[78,94]	[80,97]	[79,95]	[84,98]
50	[1,4]	[1,4]	[1,4]	[1,4]	[2,4]
70	[52,81]	[63,79]	[63,80]	[67,81]	[76,89]
90	[47,74]	[40,66]	[43,70]	[41,68]	[38,64]

Table 5: Two-sided confidence intervals constructed for jokes with ID in $\{10, 30, 50, 70, 90\}$. We choose the sampling probability p_M such that $M/[\binom{n-1}{M-1}p_M]$ is a fixed constant for $M \in \{2, \cdots 6\}$.

	$p_{2,3} = 1 \times 10^{-1}$		$p_{3,4} = 3 \times 10^{-3}$		$p_{4,5} = 1.2 \times 10^{-4}$		$p_{5,6} = 6 \times 10^{-5}$	
Joke ID	$M = 2$	$M = 3$	$M = 3$	$M = 4$	$M = 4$	$M = 5$	$M = 5$	$M = 6$
10	[12,69]	[33,39]	[17,63]	[28,45]	[15,50]	[29,36]	[15,47]	[27,36]
30	[54,98]	[83,87]	[66,99]	[86,95]	[78,99]	[88,94]	[78,99]	[88,97]
50	[1,10]	[1,2]	[1,14]	[2,4]	[1,5]	[2,2]	[1,7]	[2,2]
70	[45,92]	[68,75]	[50,90]	[73,81]	[51,87]	[76, 79]	[61,97]	[77,80]
90	[32,88]	[55,63]	[40,84]	[55,68]	[35,77]	[51,63]	[31,72]	[52,64]

Table 6: Two-sided confidence intervals constructed for jokes with ID in $\{10, 30, 50, 70, 90\}$. We compare the confidence intervals for every two adjacent M under the same sampling probability.