Statistica Sinica Preprint No: SS-2024-0153	
Title	Tests on Dynamic Ranking
Manuscript ID	SS-2024-0153
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0153
Complete List of Authors	Nan Lu, Jian Shi, Xin-Yu Tian and Kai Song
Corresponding Authors	Jian Shi
E-mails	jshi@iss.ac.cn
Notice: Accepted author version.	

Statistica Sinica

TESTS ON DYNAMIC RANKING

Nan Lu, Jian Shi, Xin-Yu Tian and Kai Song

Chinese Academy of Sciences, University of Chinese Academy of Sciences and University of Minnesota

Abstract: In this paper, we investigate the dynamic Bradley-Terry model, a highly acclaimed statistical ranking model, and tackle several crucial inference problems related to score functions and rank properties. Specifically, we tackle the test problems of score function variation and pairwise similarity, providing valuable insights for model determination and simplification. We derive asymptotic null distributions for the proposed test statistics and prove the tests' consistency. Furthermore, we introduce a novel confidence band of dynamic rank and establish an innovative and generally applicable test framework for dynamic ranking properties. To overcome the conservativeness issue brought by the supreme form statistics, we introduce a novel approach based on signed score difference statistics for ranking inferences. We present theoretical guarantees for the proposed scheme. Numerical simulations validate the theories and demonstrate the satisfactory performance of our methods. The proposed methods are applied to a real dataset, yielding insightful results.

Key words and phrases: Bradley-Terry model, Combinatorial inference, Dynamic ranking property, Time-varying test, Uncertainty quantification.

Corresponding author: Jian Shi. E-mail: jshi@iss.ac.cn.

1. Introduction

Ranking problems are widely studied in various disciplines, including information retrieval (Cambazoglu et al., 2010), computational biology (Duh and Kirchhoff, 2008), recommender systems (Lv et al., 2011), and sports games (Bong et al., 2020), among others. These problems have garnered considerable attention and have been extensively explored across different fields (Wang et al., 2013; Chen et al., 2019; Liu et al., 2023). The Bradley-Terry (BT) model (Bradley and Terry, 1952) is one of the most widely acclaimed statistical ranking models.

Though there have been tremendous studies on the BT model, such as Simons and Yao (1999); Hunter (2004); Huang et al. (2006); Yan et al. (2012); Han et al. (2020); Gao et al. (2023), most of them focus on the static case. However, in reality, data is often collected with timestamps, and it is more reasonable to assume that audience preferences for movies change over time, as well as the potential ability of a basketball team is time-varying. Some studies on the dynamic BT model have been conducted from the Bayesian perspective (Glickman, 1999, 2001; Maystre et al., 2019), while McHale and Morton (2011); Bong et al. (2020) employ the maximum likelihood approach. When involved in large-scale dynamic calculations, a spectral method called rank centrality demonstrates better performance (Negahban et al., 2017; Tian et al., 2024). Different from all these papers focusing on score estimation, this paper addresses some critical test problems, which

remain unexplored in the existing literature. Additionally, this paper introduces novel theoretical findings that reveal the properties of functional estimations, which play crucial roles in dynamic inferences.

When generalizing the static BT model to its dynamic version, a primary concern is to determine if there is sufficient evidence indicating that the latent abilities of the model truly vary over time. The more precise dynamic BT model comes at the expense of a slower convergence rate, as detailed in Section 3. Therefore, it would be beneficial to identify items with equal latent abilities since concating their comparison results allows for more efficient use of information. We address these two concerns by discussing several hypothesis test problems. Another interesting inferential problem relates to ranking properties, which have received limited attention in the existing literature. Unlike classical parameters, ranks are discrete transformations of individual ability comparisons, making their inference a novel and distinct combinatorial problem. Liu et al. (2023) first introduce an inferential framework for ranking problems in the static case, and it is extended to the static Plackett-Luce model (Fan et al., 2024). Mogstad et al. (2023) construct confidence sets for static ranks, and Bazylik et al. (2024) apply the methods to rank journals and universities. However, in a dynamic situation, estimations of individual abilities are stochastic processes, giving rise to a different and intriguing problem. In this work, we establish an inferential framework for dynamic

scenarios. Furthermore, different from all the existing work based on an absolute form perturbation term, we introduce a novel signed statistic, effectively addressing the conservatism issue. To the best of our knowledge, we are the first to present the confidence band of dynamic rank and provide the general test scheme for dynamic ranking properties. It is important to note that ranks are discontinuous and complicated functions of eigenvectors in spectral methods, as illustrated in Section 4. As highlighted in Chen et al. (2019), generating confidence regions for deterministic functions of eigenvectors in the spectral method remains an open and challenging problem. We partially address this problem by providing a novel solution to perform efficient statistical inference for a specific class of complicated functions involving eigenvalues.

The rest of the paper is organized as follows. In Section 2, we introduce the dynamic BT model. Section 3 focuses on the tests for variation and equality of score functions. The confidence band of rank and general ranking property test schemes are presented in Section 4. Section 5 and 6 include simulation results and practical applications, followed by concluding remarks in Section 7. Additional simulations and proofs are deferred to the supplementary material.

2. Dynamic BT Model

In this section, we introduce the dynamic BT model and review the estimation algorithm to facilitate further discussions.

Suppose there are n items for comparison. We normalize the time interval to $[0, 1]$ without loss of generality. The dynamic BT model assumes a latent score vector $\pi^*(t) = (\pi_1^*(t), \pi_2^*(t), \ldots, \pi_n^*(t))^\top$ for items. In general, $\pi^*(t)$ is unobservable and represents qualities of goods, abilities of sports teams, or other factors that determine the comparison outcomes and are specific to the application domain. When $\pi^*(t)$ is constant, the model reduces to the static case. For each pair $(i, j) \in \mathcal{E}$, assume the observations take place in the time set T_{ij}. Set $M_{ij} = \#|T_{ij}|$. We use $y := \{y_{ij}(t) \,|\, i, j \in \mathcal{E}, i \neq j, t \in T_{ij}\}$ to denote the comparison results, which are independent random variables following the Bernoulli distribution. Specifically, the comparison $y_{ij}(t) \sim Bernoulli(y_{ij}^*(t))$, where $y_{ij}^*(t) = \frac{\pi_j^*(t)}{\pi_i^*(t)+\pi_j^*(t)}$. We have $y_{ij}(t) = 1$ represent that item j wins, and $y_{ij}(t) = 0$ indicates that item i wins. Let $[n]$ represent $\{1, 2, \ldots, n\}$ and $y^* := \{y_{ij}^*(t) \,|\, i, j \in [n], i \neq j, t \in [0, 1]\}$. Since $y^*(t)$ remain unchanged with $\pi^*(t)$ multiplying a nonzero constant, we assume $\sum_{i=1}^n \pi_i^*(t) = 1$ for identifiability. To model the compared pairs, we introduce the comparison graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$. Items i and j are compared if and only if (i, j) belongs to the edge set $\mathcal{E}$. We let $\mathcal{G}$ be an Erdős-Rényi (ER) graph $\mathcal{G}_{n,p}$, which consists of n vertices and each edge appears independently with probability p. Let

$M = \min_{(i,j)\in\mathcal{E}} M_{ij}$.

We establish our test methods on a spectral method called Rank Centrality (RC) (Negahban et al., 2017), which builds a connection between the latent scores $\pi^*(t)$ and the stationary probability of a random walk on the comparison graph. RC has garnered attention due to its intuitive interpretation, simplified solution form, and faster computation speed. Further, the Kernel Rank Centrality (KRC) method incorporates a kernel function into the transition probability matrix to accommodate the dynamic case (Tian et al., 2024). Specifically, let $K(\cdot)$ be the kernel function and h be the bandwidth, and define $K_h(t,s) = \frac{1}{h}K\left(\frac{t-s}{h}\right)$. The estimation algorithm consists of two main steps. First, we estimate the stochastic transition matrix $P(t) = (P_{ij}(t))_{i,j\in[n]}$.

$$P_{ij}(t) = \begin{cases} \frac{1}{2np}\frac{\sum_{t_k\in T_{ij}} y_{ij}(t_k)K_h(t,t_k)}{\sum_{t_k\in T_{ij}} K_h(t,t_k)} & \text{if } (i,j)\in\mathcal{E}, \\ 1-\sum_{s\neq i} P_{is}(t) & \text{if } i=j, \\ 0 & \text{otherwise.} \end{cases}$$

Then we compute the stationary distribution $\hat{\pi}(t)$ of the Markov chain defined by $P(t)$, i.e., $\hat{\pi}(t)^\top = \hat{\pi}(t)^\top P(t)$. The following assumptions are required to ensure the theoretical properties of the estimations.

Assumption (A)

(A1). $\sup_{t\in[0,1]} \frac{\max_i \pi_i^*(t)}{\min_i \pi_i^*(t)} \leq \kappa$, where $\kappa > 0$ is a constant. $y_{ij}^*(t)$ is three times continuously differentiable.

(A2). The kernel function is symmetric, nonnegative, and satisfies $\int_{-\infty}^{\infty} K(v)\,dv = 1$ and $\int_{-\infty}^{\infty} v^2 K(v)\,dv < \infty$.

(A3). There exists a constant $\kappa_1 > 0$, such that $\frac{\max_{(i,j)\in\mathcal{E}} M_{ij}}{\min_{(i,j)\in\mathcal{E}} M_{ij}} \leq \kappa_1$.

(A4). $K(|x|) = O(\frac{1}{|x|^{\varsigma}})$ for some $\varsigma > 0$.

Assumptions (A1)-(A3) are commonly used in BT model estimation and kernel methods (Gao et al., 2023; Liu et al., 2023; Tian et al., 2024). Most commonly used kernel functions satisfy assumption (A4).

3. Tests on Score Functions

With the increasing amount of data collected with timestamps and the growing interest in dynamic situations, selecting the appropriate estimation model becomes a fundamental challenge to ensure accurate and efficient results. In this section, we address this problem by presenting the variation and equality tests.

3.1 Test of the Variation of Score Functions

When dealing with time-dependent observed pairwise comparison results, it is crucial to choose between the static and dynamic models. Opting for the dynamic method in every case may result in reduced estimation accuracy due to the underutilization of available information. Specifically, the estimation convergence rate for the static case $\|\pi - \pi^*\|_\infty / \|\pi^*\|_\infty$ is $O_p(\sqrt{\frac{\log n}{npM}})$ (Chen et al., 2019; Gao et al.,

2023), whereas the convergence rate for the dynamic case is $O_p(\sqrt{\frac{\log n}{npMh}})$ (see Theorem 3). On the other hand, misusing the static method can also be detrimental as it can lead to imprecise estimation of the score function, subsequently affecting further inference. We focus on the following hypothesis test for $i \in [n]$.

$$H_{i0}^a : \pi_i^*(t) \text{ is a constant for } t \text{ in } [0,1]; \; H_{i1}^a : \pi_i^*(t) \text{ is not a constant for } t \text{ in } [0,1].$$

Define the scaling factor $\alpha_i(t) = \sqrt{\frac{h(\sum_{j:(i,j)\in\mathcal{E}} y_{ij}^*(t))^2}{\int K^2(v)dv \, \pi_i^*(t) \sum_{j:(i,j)\in\mathcal{E}} \frac{\pi_j^*(t)}{M_{ij}}}}$. Let $\hat{\alpha}_i(t)$ be the sample version of $\alpha_i(t)$ with $\hat{\pi}(t)$ plugged in. We construct the test statistic

$$T_{ai} = \frac{\sum_{t\in\mathcal{S}}[\hat{\alpha}_i(t)(\hat{\pi}_i(t) - \frac{1}{m}\sum_{s\in\mathcal{S}}\hat{\pi}_i(s))]^2 - m}{\sqrt{2m}},$$

where set $\mathcal{S}$ is constructed by equidistant time points spanning from 0 to 1 and m is the cardinality of $\mathcal{S}$. We reject H_{i0}^a if $T_{ai} > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$-th quantile of standard normal distribution. The test procedure is based on the following theorem. We first present a lemma analyzing the main term. For notation simplicity, we set $\beta_n = \min\{\sqrt{npMh^2}, \frac{1}{\sqrt{npMh^5}}, \frac{\sqrt{n}}{\log n}, \frac{1}{h^{\frac{2\varsigma}{1+2\varsigma}}}\}$.

Lemma 1. *Let Assumptions (A1)-(A4) hold. If $np > c\log n$ for some sufficiently large c, $nMh^5 = o(1)$, $\frac{\log n}{Mh} = o(1)$ and $m = o(\beta_n)$, then we have*

$$\frac{\sum_{t\in\mathcal{S}}[\alpha_i(t)(\hat{\pi}_i(t) - \pi_i^*(t))]^2 - m}{\sqrt{2m}} \xrightarrow{\mathcal{D}} N(0,1)$$

as $n \to \infty$.

Based on Lemma 1, we have the following theorem, validating our test procedure.

Theorem 1. *Under the conditions of Lemma 1, we have*

$$T_{ai} \xrightarrow{\mathcal{D}} N(0,1)$$

as $n \to \infty$ *under* H_{i0}^a*, and we have*

$$P(T_{ai} > z_{1-\alpha}) \to 1$$

as $n, m \to \infty$ *under* H_{i1}^a.

Remark 1. To differentiate the dynamic scores, we construct the test statistic T_{ai}, whose main component is $\sum_{t\in\mathcal{S}}[\hat{\alpha}_i(t)(\hat{\pi}_i(t) - \frac{1}{m}\sum_{s\in\mathcal{S}}\hat{\pi}_i(s))]^2$. Under H_{i0}^a, as shown in Theorem 1, the leading term is $\sum_{t\in\mathcal{S}}[\alpha_i(t)(\hat{\pi}_i(t) - \pi_i^*(t))]^2$. However, we note that it is not feasible to directly deduce the asymptotic distribution of this summation due to the high correlation among the m terms, which arises from kernel smoothing. Utilizing a precise expansion detailed in Section **??** of the supplementary material, we reformulate the leading term as a quadratic form involving $\sum_{j:(i,j)\in\mathcal{E}} M_{ji}$ independent variables. By carefully controlling the growth of m, we establish the asymptotic distribution utilizing results from de Jong (1987).

Remark 2. The first part of Theorem 1 is valid for both finite m and divergent m, where the divergence speed is governed by β_n. For the second part of Theorem 1,

a divergent m is necessary to adequately capture the behavior of $\hat{\pi}_i$ across the entire time interval, ensuring the test's consistency. Consequently, the divergence of β_n is required. Specifically, the condition $m \to \infty$ implies that $npMh^2 \to \infty$. This ensures that $\beta_n \to \infty$ combining the conditions of Theorem 1. The choice of m is dependent on β_n. For example, if $M \asymp np$, $h \asymp (np)^{-1/2}$ and $\varsigma \geq \frac{1}{2}$, then $m = o((np)^{1/4})$ satisfies the condition of the theorem, where $a \asymp b$ denotes that $a = O(b)$ and $b = O(a)$.

Then we consider the multiple hypothesis testing problem. Let K be a finite subset of $[n]$ with $\#K = k$. We aim to simultaneously test the following hypotheses:

$$H_{i0}^{\tilde{a}} : \pi_i^*(t) \text{ is a constant for } t \text{ in } [0,1];\ H_{i1}^{\tilde{a}} : \pi_i^*(t) \text{ is not a constant for } t \text{ in } [0,1], i \in K.$$

We provide the test procedure as follows utilizing the Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001).

1. Calculate the p-values $p_i = P(z > T_{ai})$ for $i \in K$, where z follows the standard normal distribution.

2. Order the observed p-values as $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(k)}$. Define $s = \max\{i : p_{(i)} \leq \frac{i}{k}\frac{\alpha}{\sum_{i=1}^{k} 1/i}\}$ and reject hypotheses $H_{(1)0}^{\tilde{a}}, H_{(2)0}^{\tilde{a}}, \ldots, H_{(s)0}^{\tilde{a}}$. If no such i exists, no hypotheses are rejected.

Given the number of true null hypotheses k_0, the above procedure controls the false discovery rate (FDR) at level $k_0\alpha/k$ (Benjamini and Yekutieli, 2001).

Remark 3. Extending the results of Theorem 1 to diverging n is challenging. Given the assumption $\sum_{i=1}^{n} \pi_i^*(t) = 1$, we consider the straightforward summation of $n-1$ items, $(\sum_{i=1}^{n-1} \sum_{t\in\mathcal{S}} [\hat{\alpha}_i(t)(\hat{\pi}_i(t) - \frac{1}{m}\sum_{s\in\mathcal{S}} \hat{\pi}_i(s))]^2 - m(n-1))/\sqrt{2m(n-1)}$. With both m and n diverging, stringent conditions are required to ensure the test's validity. Specifically, following the proof of Theorem 1, we decompose the test statistic as follows:

$$\frac{\sum_{i=1}^{n-1} \sum_{t\in\mathcal{S}} [\hat{\alpha}_i(t)(\hat{\pi}_i(t) - \pi_i^*(t))]^2 - m(n-1)}{\sqrt{2m(n-1)}} + \frac{\sum_{i=1}^{n-1} \sum_{t\in\mathcal{S}} \alpha_i^2(t)(2\hat{\pi}_i(t) - \pi_i^*(t) - \frac{1}{m}\sum_{s\in\mathcal{S}} \hat{\pi}_i(s))(\pi_i^*(t) - \frac{1}{m}\sum_{s\in\mathcal{S}} \hat{\pi}_i(s))}{\sqrt{2m(n-1)}}.$$

Intuitively, the first term approaches a normal distribution, while the second term converges to

$$\sqrt{\frac{m}{2(n-1)}} \sum_{i=1}^{n-1} \int_0^1 \alpha_i^2(t) \Big(\pi_i^*(t) - \int_0^1 \pi_i^*(s) ds\Big)^2 dt$$

as $m, n \to \infty$. Letting n_d represent the number of dynamic items, the above term is of order $O(\sqrt{mn} n_d p M h)$. If n_d is a constant, m needs to be sufficiently large such that $\sqrt{mn} pMh \to \infty$ to maintain the test's validity. This condition is stringent, as the growth rate of m is constrained by β_n to control the correlation among different t induced by kernel smoothing.

Remark 4. Given the test results of the BY procedure, identifying even a single dynamic $\pi^*(t)$ provides valuable insight into the necessity of dynamic ranking. To

further make statistical inferences, we recommend allowing h to vary across pairs. A natural approach is to assign larger h to pairs involving items with higher p-values, enabling more efficient utilization of the available information. However, determining the optimal h for different pairs is a challenging problem, which we leave for future work.

3.2 Test of the Equality of Two Score Functions

For two items, a primary concern is whether they essentially have the same ability score, which is difficult to determine directly from the complex dynamic pairwise comparison results. On the other hand, compared to the static estimation method, the dynamic method utilizes less information at a fixed time point. Specifically, if we choose the Epanechnikov kernel function, for instance, $2Mh$ nearby observation results are used on average for each point. The bandwidth h is typically smaller than 0.5, resulting in a slower convergence rate for the dynamic case, as mentioned in Section 3.1. If there is additional information indicating that the score functions of several items are identical, the related observation results can be combined for more accurate estimation. Therefore, we consider the score equality test:

$$H_0^b : \forall t, \pi_i^*(t) = \pi_j^*(t); \ H_1^b : \exists t_0, \pi_i^*(t_0) \neq \pi_j^*(t_0).$$

Enlightened by the results in the preceding part, we define the test statistic as

$$T_b = \frac{\sum_{t\in\mathcal{S}}[\frac{1}{\sqrt{2}}\hat{\alpha}_i(t)(\hat{\pi}_i(t)-\hat{\pi}_j(t))]^2 - m}{\sqrt{2m}}.$$

Recall that we define the scaling factor $\alpha_i(t) = \sqrt{\frac{h(\sum_{j:(i,j)\in\mathcal{E}} y_{ij}^*(t))^2}{\int K^2(v)dv\, \pi_i^*(t) \sum_{j:(i,j)\in\mathcal{E}} \frac{\pi_j^*(t)}{M_{ij}}}}$ in Section 3.1 and let $\hat{\alpha}_i(t)$ be the sample version of $\alpha_i(t)$, obtained by substituting $\hat{\pi}(t)$ for $\pi^*(t)$. We then deduce the asymptotic distribution of T_b and present the test procedure.

Theorem 2. *Under the conditions of Lemma 1, we have*

$$T_b \xrightarrow{\mathcal{D}} N(0,1)$$

as $n \to \infty$ *under* H_0^b*, and we have*

$$P(T_b > z_{1-\alpha}) \to 1$$

as $n, m \to \infty$ *under* H_1^b.

Based on Theorem 2, we reject the null hypothesis if $T_b > z_{1-\alpha}$. We calculate the p-value $p_b = P(z > T_b)$, where z follows the standard normal distribution.

4. Inference of Dynamic Ranking Properties

The ranking problem has received significant attention across diverse fields. In current practical applications, it is common to directly display rank outcomes obtained through estimations of latent abilities. However, the absence of uncertainty

quantification in these estimations and predictions raises doubts about their reliability. To address this limitation, it becomes crucial to develop tools for tackling inference questions, such as ascertaining the probability of a stock ranking first at 0.9 or 0.6. Should the latter probability prove accurate, investors would be better served by allocating their resources toward diversified portfolios. Additionally, simply stating that player i is ranked higher than player j without providing a confidence level diminishes practical informativeness. Our objective is to address these uncertainty problems head-on and provide effective solutions.

Within this section, we focus on the ranking of items. We recognize that the relative value of latent abilities serves as the determining factor for rank positions. Leveraging this insight, our approach begins by approximating the distribution of a supremum of score difference error and deducing the confidence band of rank. Subsequently, we present a comprehensive test scheme that addresses the general ranking property in dynamic settings. This scheme is versatile and can be adapted to various test problems. We provide rigorous theoretical guarantees for the inference process. However, it is worth noting that while the inferential error can be controlled theoretically, the use of a supreme type statistic often results in conservatism, leading to wider confidence intervals and potentially lower test power. To mitigate this issue, we delve deeper into the relationship between scores and ranks and propose a novel perturbation statistic, which yields a more practical inference

method.

Differing from the inference in the static case, the elements of $\hat{\pi}(t)$ are stochastic processes. Therefore, a certain degree of uniformity in $\hat{\pi}(t)$ is necessary for deducing the rank property within an interval. Before presenting the test framework, we establish a uniform error bound for $\hat{\pi}$, which plays a fundamental role in the subsequent deduction and is meaningful in itself. We need the following assumption on the smoothness of y^*.

Assumption (A5). There exist constants c_1, c_2, such that $\sup_t \max_{i,j} |\dot{y}^*_{ij}(t)| \leq c_1$ and $\sup_t \max_{i,j} |\ddot{y}^*_{ij}(t)| \leq c_2$, where $\dot{y}^*_{ij}(t)$ and $\ddot{y}^*_{ij}(t)$ represent the first and second order derivatives respectively.

Theorem 3. *Let Assumptions (A1)-(A3) and (A5) hold. If $np > c \log n$ for some sufficiently large c, $nMh^5 = o(1)$, then there exists constant c_3, such that*

$$P(\sup_t \|\hat{\pi}(t) - \pi^*(t)\|_\infty \leq c_3 \sqrt{\frac{\log n}{n^3 pMh}}) \to 1 \tag{4.1}$$

as $n, Mh \to \infty$.

Remark 5. From Theorem 3, we can obtain the convergence rate that $\sup_t \|\hat{\pi}(t) - \pi^*(t)\|_\infty / \sup_t \|\pi^*(t)\|_\infty = O_p(\sqrt{\frac{\log n}{npMh}})$. By employing a truncation kernel such as the Epanechnikov kernel with bandwidth h, the effective number of observations used for estimation at a single time point is $2Mh$. Our result aligns with the optimal rate $O_p(\sqrt{\frac{\log n}{npL}})$ for the static case (Theorem 3 in Chen et al. (2019)), where L is the

number of comparisons for each pair. For the dynamic case, our rate is consistent with $O_p(\sqrt{\frac{\log n}{np(t)L}})$ (Theorem 7 in Karlé and Tyagi (2023)), where L is the number of comparisons at each time point in the comparison set $\{\frac{i}{T} \,|\, i = 0, \ldots, T\}$ and $p(t)$ is the probability of comparing items at time t.

4.1 Confidence Band of Rank

Suppose $r_i^*(t)$ represents the rank position of item i, where $r_i^*(t) < r_j^*(t)$ if $\pi_i^*(t) < \pi_j^*(t)$. The definition of r_i^* is flexible, allowing for the possibility that the rank of item i is not unique. Specifically, $r_i^*(t)$ can be any value within the range $rr_i^*(t) = [\underline{r}_i(t), \bar{r}_i(t)]$, where $\underline{r}_i(t) = 1 + \sum_{j\in[n]} \mathbb{1}(\pi_j^*(t) > \pi_i^*(t))$, $\bar{r}_i(t) = n - \sum_{j\in[n]} \mathbb{1}(\pi_j^*(t) < \pi_i^*(t))$ represent the smallest and largest possible ranks of item i, respectively. If item i is not tied with any other item at t, then $\underline{r}_i(t) = \bar{r}_i(t)$ and the rank is unique. Conversely, if item i is tied with at least one other item, then $\underline{r}_i(t) < \bar{r}_i(t)$.

Since rank is a global attribute, it is necessary to take into account all the other latent scores, rather than focusing solely on π_i^*. To establish a uniform band for the specified time set $T \subset [0, 1]$, we construct a supreme form statistic of score differences for all points in T. Recognizing the relationship between r_i and $\pi_i - \pi_j$, we consider the random variable

$$S = \sup_{t\in T} \max_{j:j\neq i} \gamma_{ij} |\hat{\pi}_i(t) - \hat{\pi}_j(t) - \pi_i^*(t) + \pi_j^*(t)|,$$

where $\{\gamma_{ij}\}_{i,j\in[n],i\neq j}$ are the scaling parameters. We assume that there exist con-

stants c, C, such that $c\sqrt{n^3pMh} \leq \gamma_{ij} \leq C\sqrt{n^3pMh}$ for all $i, j \in [n], i \neq j$. A natural choice of γ_{ij} is the standard deviation of $\hat{\pi}_i(t) - \hat{\pi}_j(t)$, which are of order $\sqrt{n^3pMh}$. S is dependent on i. We omit the subscript for simplicity without ambiguity.

Let $S_{1-\alpha}$ represent the $(1-\alpha)$-th quantile of S, i.e.,

$$S_{1-\alpha} = \inf\{a \in \mathbb{R} : P(S \leq a) \geq 1 - \alpha\}. \tag{4.2}$$

Define $R_l(t) = 1+\sum_{j:j\neq i} \mathbb{1}(\hat{\pi}_i(t)-\hat{\pi}_j(t)+\frac{S_{1-\alpha}}{\gamma_{ij}} < 0)$, and $R_u(t) = n-\sum_{j:j\neq i} \mathbb{1}(\hat{\pi}_i(t)-\hat{\pi}_j(t) - \frac{S_{1-\alpha}}{\gamma_{ij}} > 0)$. We claim that $\{[R_l(t), R_u(t)]\}_{t\in T}$ is a confidence band for $r_i(t)$, $t \in T$, which is a direct result of the following proposition.

Proposition 1. *Let* $E_1 = \{\text{for all } t \in T, rr_i^*(t) \in [R_l(t), R_u(t)]\}$ *and* $E_2 = \{\sup_{t\in T} \sup_{j:j\neq i} |\pi_i^*(t) - \pi_j^*(t) - (\hat{\pi}_i(t) - \hat{\pi}_j(t))| \leq \frac{S_{1-\alpha}}{\gamma_{ij}}\}$, *then we have* $E_2 \subset E_1$.

From Proposition 1 and (4.2), we have

$$P(E_1) \geq P(E_2) \geq 1 - \alpha,$$

indicating that $\{[R_l(t), R_u(t)]\}_{t\in T}$ is a valid confidence band.

Note that $S_{1-\alpha}$ is an unknown quantity. Estimating $S_{1-\alpha}$ essentially involves determining the asymptotic distribution in a high-dimensional space. Utilizing Theorem 3, we will show that the distribution of the supremum statistic can be approximated by its discretized counterpart using the Gaussian multiplier bootstrap. For simplicity, we assume that $M_{ij} = M$ for all $i, j \in [n]$ in the subsequent

derivation. The following deduction can be readily extended to accommodate the general scenario where the number of observations for different pairs may vary. Let $A_{ij} = 1$ if $(i,j) \in \mathcal{E}$ and $A_{ij} = 0$ otherwise. Let n_i denote the degree of item i in the comparison graph. Building upon the expansion of the estimation

$$\hat{\pi}_i(t) - \pi_i^*(t) = \frac{1}{\sum_{j:(i,j)\in\mathcal{E}} y_{ij}^*(t)} \sum_{j:(i,j)\in\mathcal{E}} (\pi_i^*(t) + \pi_j^*(t))\bar{\Delta}_{ij}(t) + o_p(1),$$

where $\bar{\Delta}_{ij}(t) = \frac{\sum_{t_k\in T_{ji}}(y_{ji}(t_k) - y_{ji}^*(t_k))K_h(t,t_k)}{\sum_{t_k\in T_{ji}} K_h(t,t_k)}$ (see Theorem S?? in the supplementary material for detailed deduction), we have

$$\begin{aligned}
S =& \sup_{t\in T} \max_{j:j\neq i} \gamma_{ij} \left| \frac{\sum_{l:(i,l)\in\mathcal{E}} (\pi_i^*(t) + \pi_l^*(t))\bar{\Delta}_{il}(t)}{\sum_{l:(i,l)\in\mathcal{E}} y_{il}^*(t)} - \frac{\sum_{l:(j,l)\in\mathcal{E}} (\pi_j^*(t) + \pi_l^*(t))\bar{\Delta}_{jl}(t)}{\sum_{l:(j,l)\in\mathcal{E}} y_{jl}^*(t)} \right| + o_p(1) \\
=& \sup_{t\in T} \max_{j:j\neq i} \left| \sqrt{\frac{1}{Mn_{ij}}} \sum_{k=1}^{M} \Big(\sum_{l:(i,l)\in\mathcal{E}, l\neq j} x_{kli}(t) + \sum_{l:(i,l)\in\mathcal{E}, l\neq j} x_{klj}(t) + A_{ij} x_{kji}(t) \Big) \right| + o_p(1) \\
=:& \sup_{t\in T} \max_{j:j\neq i} \left| \sqrt{\frac{1}{Mn_{ij}}} \sum_{k=1}^{M} \sum_{l=1}^{n_{ij}} x_{kl}^{(ij)}(t) \right| + o_p(1), \qquad (4.3)
\end{aligned}$$

where $n_{ij} = n_i + n_j - \mathbb{1}((i,j)\in\mathcal{E})$, and

$$x_{kli}(t) = \frac{\sqrt{\gamma_{ij}^2 n_{ij} M} K_h(t,t_k)(\pi_i^*(t) + \pi_l^*(t))(y_{li}(t_k) - y_{li}^*(t_k))}{\sum_{k=1}^{M} K_h(t,t_k) \sum_{l:(i,l)\in\mathcal{E}} y_{il}^*(t)},$$

$$x_{klj}(t) = \frac{\sqrt{\gamma_{ij}^2 n_{ij} M} K_h(t,t_k)(\pi_j^*(t) + \pi_l^*(t))(-y_{lj}(t_k) + y_{lj}^*(t_k))}{\sum_{k=1}^{M} K_h(t,t_k) \sum_{l:(j,l)\in\mathcal{E}} y_{jl}^*(t)},$$

$$x_{kji}(t) = \frac{\sqrt{\gamma_{ij}^2 n_{ij} M} K_h(t,t_k)(\pi_i^*(t) + \pi_j^*(t))(y_{ji}(t_k) - y_{ji}^*(t_k))}{\sum_{k=1}^{M} K_h(t,t_k)}$$

$$\times \left(\frac{1}{\sum_{l:(i,l)\in\mathcal{E}} y^*_{il}(t)} + \frac{1}{\sum_{l:(j,l)\in\mathcal{E}} y^*_{jl}(t)}\right) \text{ if } (i,j) \in \mathcal{E},$$

for $k \in [M]$, $l \neq i, j$ are $n_{ij}M$ independent variables with unknown variances. t_k is dependent on the corresponding items. For instance, t_k in x_{kli} represents the kth comparison time between items l and i. We omit these symbols for brevity.

From equation (4.3), S represents the supreme coordinate of a $2(n-1)|T|$-dimensional vector, with the last $(n-1)|T|$ components equal to the opposite value of the first $(n-1)|T|$ components to handle the absolute value. We discretize T into v time points $\mathcal{T}$, where $\mathcal{T}$ is a set satisfying $\max_{t,s\in\mathcal{T}} |t-s| / \min_{t,s\in\mathcal{T}} |t-s| \leq \kappa_2$ and $\kappa_2 > 0$ is a constant. If T is a continuous interval, equidistant points are a natural choice. The value of v can be allowed to diverge to infinity.

Given that $2(n-1)v$ can be significantly larger than $n_{ij}M$, and the covariance structure of $\{x^{(ij)}_{kl}(t)\}_{t\in\mathcal{T}, j:j\neq i}$ is complex, we employ the Gaussian multiplier bootstrap for high-dimensional vectors (Chernozhukov et al., 2013). The main idea is to approximate the distribution of the maximum of a sum of independent random vectors with their Gaussian equivalent. We multiply the original vectors with independently and identically distributed standard normal variables to obtain conditional Gaussian vectors through bootstrap. We define the following Gaussian analogue of S that

$$V = \max_{t\in\mathcal{T}} \max_{j:j\neq i} \left| \sqrt{\frac{1}{Mn_{ij}}} \sum_{k=1}^{M} \Big(\sum_{l:(i,l)\in\mathcal{E}, l\neq j} \hat{x}_{kli}(t) z_{kli} + \sum_{l:(i,l)\in\mathcal{E}, l\neq i} \hat{x}_{klj}(t) z_{klj} + A_{ij} \hat{x}_{kji}(t) z_{kji} \Big) \right|$$

$$=: \max_{t\in\mathcal{T}} \max_{j:j\neq i} \left| \sqrt{\frac{1}{Mn_{ij}}} \sum_{k=1}^{M} \sum_{l=1}^{n_{ij}} \hat{x}_{kl}^{(ij)}(t) z_{kl} \right|,$$

where $\hat{x}$ represents x with estimators $\hat{\pi}$ plugged in, for example, $\hat{x}_{kli}(t) = \frac{\sqrt{\gamma_{ij}^2 n_{ij} M} K_h(t,t_k)(\hat{\pi}_i(t)+\hat{\pi}_l(t))(y_{li}(t_k)-\hat{y}_{li}(t_k))}{\sum_{k=1}^{M} K_h(t,t_k) \sum_{l:(i,l)\in\mathcal{E}} \hat{y}_{il}(t)}$. And $\{z_{kl}\}_{k\in[M],l\in[n_{ij}]}$ are independent standard normal random variables. Let $V_{1-\alpha}$ be the $(1-\alpha)$-th quantile of V given y, i.e.,

$$V_{1-\alpha} = \inf\{a \in \mathbb{R} : P(V \leq a|y) \geq 1-\alpha\}.$$

The following theorem verifies the correspondence between S and V. Let $ploy(x)$ represent the polynomial of x.

Theorem 4. *Given that Assumptions (A1)-(A3) and (A5) hold, let* $h \asymp \frac{1}{n^{a_1} M^{b_1}}$ *for* $\frac{1}{5} < a_1 < \frac{1}{2}$, $\frac{1}{5} < b_1 < 1$, $v \asymp \frac{n^a M^b}{h^{1/2}}$ *for* $a, b > \frac{1}{2}$, $\frac{poly(\log M)}{n} = o(1)$. *If* $np > c \log n$ *for some sufficiently large* c, *then we have* $\sup_{\alpha\in(0,1)} |P(S > V_{1-\alpha}) - \alpha| \to 0$ *as* $n, Mh \to \infty$.

Theorem 4 demonstrates that V is a valid approximation of S. By substituting $S_{1-\alpha}$ with $V_{1-\alpha}$ in $R_l(t)$ and $R_u(t)$, we obtain the confidence band $\{[\hat{R}_l(t), \hat{R}_u(t)]\}_{t\in T}$.

Notice that $(1-\alpha)$ is essentially the level of score differences. When we employ a discretized transformation on scores for rank inferences, a wide gap between each pair of items tends to push the rank level toward 1. This phenomenon enhances the

precision of rank inferences. The following proposition provides a more detailed explanation.

Proposition 2. *Under the conditions of Theorem 3, if* $\min_{j:j\neq i}\inf_{t\in T}|\pi_i^*(t)-\pi_j^*(t)|\gg\sqrt{\frac{\log(nM)}{n^3pMh}}$, *then we have* $P(\hat{R}_l(t)=i \text{ and } \hat{R}_u(t)=i,\forall t\in T)\to 1$ *as* $n, Mh\to\infty$.

The proposition demonstrates that as long as the difference among items is higher than the order of $\sqrt{\frac{\log(nM)}{n^3pMh}}$, the coverage probability will no longer be $1-\alpha$, but 1. This distinction is significant compared to classical parameter inference, arising from the unique nature of ranks and contributing to the high accuracy of rank inference.

We now briefly discuss the one-sided confidence band, which is particularly significant for the top-K test discussed in Section 4.2. Given the previous results, the deduction of a one-sided confidence band is straightforward, so we omit the detailed proof. We consider the random variable that

$$\begin{aligned} W =& \sup_{t\in T}\max_{j:j\neq i}\gamma_{ij}(\hat{\pi}_i(t)-\hat{\pi}_j(t)-\pi_i^*(t)+\pi_j^*(t)) \\ =& \sup_{t\in T}\max_{j:j\neq i}\sqrt{\frac{1}{Mn_{ij}}}\sum_{k=1}^{M}\sum_{l=1}^{n_{ij}}x_{kl}^{(ij)}(t)+o_p(1), \end{aligned}$$

and define its Gaussian analog that

$$U=\max_{t\in\mathcal{T}}\max_{j:j\neq i}\sqrt{\frac{1}{Mn_{ij}}}\sum_{k=1}^{M}\sum_{l=1}^{n_{ij}}\hat{x}_{kl}^{(ij)}(t)z_{kl}.$$

Let $U_{1-\alpha}$ be the $(1-\alpha)$-th quantile of U given y. Under the conditions of Theorem 4, we can obtain $\sup_{\alpha\in(0,1)}|P(W > U_{1-\alpha}) - \alpha| \to 0$. Let $\tilde{R}_u(t) = n - \sum_{j:j\neq i}\mathbb{1}(\hat{\pi}_i(t) - \hat{\pi}_j(t) - \frac{W_{1-\alpha}}{\gamma_{ij}} > 0)$. Then we have the one-sided confidence band estimation $\{[1, \hat{\tilde{R}}_u(t)]\}_{t\in T}$ by substituting $W_{1-\alpha}$ with $U_{1-\alpha}$.

4.2 Dynamic Ranking Inference (DRI)

We introduce formal definitions of the dynamic ranking property first. We denote the rank of n items by $r = (r_1, \ldots, r_n)^\top$, where r_i is the rank of item i. We use $rr(\pi)$ to represent the corresponding rank of score $\pi = (\pi_1, \ldots, \pi_n)^\top$. Specifically, we let $rr_i(t) = [1 + \sum_{j\in[n]}\mathbb{1}(\pi_j(t) > \pi_i(t)), n - \sum_{j\in[n]}\mathbb{1}(\pi_j(t) < \pi_i(t))]$ to include the presence of ties. The collection of all possible ranks is denoted as R. Recall that we use r^* to denote the truth. Let $R_i(t) \subset R$ be the subset of ranks that exhibit a specific ranking property regarding item i. For instance, $R_i(t)$ could indicate that item i is ranked among the top K items or that it is ranked higher than a fixed item j. In other words, if $r(t) \in R_i(t)$, then $r(t)$ satisfies the property represented by $R_i(t)$. We still use T to represent the time set of interest. The general ranking property test takes the following form:

$$H_0^c : \exists t_0 \in T, r^*(t_0) \notin R_i(t_0);\ H_1^c : \forall t \in T, r^*(t) \in R_i(t), \tag{4.4}$$

which covers various scenarios with different settings of T and $R_i := \{R_i(t)\}_{t\in T}$. As a simple example, letting $T = \{t_0\}$ and $R_i(t_0) = \{r : r_i(t_0) \leq K\}$, the test

concerns whether item i ranks top K at t_0:

$$H_0^c : r_i^*(t_0) > K;\ H_1^c : r_i^*(t_0) \leq K.$$

Notice that the correspondence between $\{\pi_i\}_{i\in[n]}$ and $\{r_i\}_{i\in[n]}$ is not one-to-one, as different latent scores can result in the same rank. To measure the distance between H_0^c and H_1^c, we adapt the concept of legal pair (Liu et al., 2023) to dynamic scenarios. Supposing that $r(t) \in R_i(t)$, if the new rank $r'(t) \notin R_i(t)$ after swapping the scores of items i and j, then we call (i, j) a legal pair. Define $\Delta(\pi, R_i, t) = \min_{j:(i,j)} |\pi_i(t) - \pi_j(t)|$ is a legal pair for $R_i(t)$ and $\tilde{\Delta}(\pi, R_i, T) = \inf_{t\in T} \Delta(\pi, R_i, t)$.

Then we present a general test procedure for (4.4). Define the score set $\tilde{\Pi}_t(\alpha) = \{\pi(t) : \pi_j(t) = \hat{\pi}_j(t) \text{ for } j \neq i, \pi_i(t) \in [\hat{\pi}_i(t) - \frac{V_{1-\alpha}}{\gamma_{ij}}, \hat{\pi}_i(t) + \frac{V_{1-\alpha}}{\gamma_{ij}}]\}$. For a given α, if for all $t \in T$, for all $\pi \in \tilde{\Pi}_t(\alpha)$, for all $r \in rr(\pi)$, if we have $r \in R_i(t)$, then we reject H_0^c. In other words, we reject the null hypothesis if the event $\cap_{t\in T} \cap_{\pi\in\tilde{\Pi}_t(\alpha)} \cap_{r\in rr(\pi)}\{r \in R_i(t)\}$ holds. Intuitively, the rejection set is given by introducing a perturbation term of $\frac{V_{1-\alpha}}{\gamma_{ij}}$ to $\hat{\pi}_i(t)$.

Theorem 5. *Under the conditions of Theorem 4, as $n, Mh \to \infty$, we have*

$$P(\text{reject } H_0^c | H_0^c) \leq \alpha.$$

Further, if $\tilde{\Delta}(\pi^*, R_i, T) > c_4\sqrt{\frac{\log(nM)}{n^3pMh}}$, where c_4 is a constant, then we have

$$P(\text{reject } H_0^c | H_1^c) \to 1.$$

According to the theorem, the type I error can be controlled, and the power of the test approaches 1 with proper $\tilde{\Delta}(\pi, R_i, T)$.

Then we focus on the specific top-K test, for which we can develop a more efficient test procedure utilizing the one-sided confidence band. The formalization of the top-K test is as follows:

$$H_0^c : \exists t_0 \in T, r_i^*(t_0) > K; \ H_1^c : \forall t \in T, r_i^*(t) \leq K.$$

If for any $t \in T$, we have $\hat{\bar{R}}_u(t) \leq K$, then we reject H_0^c. We have the type I error below α since that

$$\begin{aligned} &P(\text{reject } H_0^c | H_0^c) = P(\forall t \in T, \hat{\bar{R}}_u(t) \leq K | \exists t_0 \in T, r_i^*(t_0) > K) \\ &\leq P(\exists t_0 \in T, r_i^*(t_0) \notin [1, \hat{\bar{R}}_u(t_0)]) \leq \alpha + o(1). \end{aligned}$$

In addition, the test power tends to 1 under the conditions of Theorem 5, with $\tilde{\Delta}(\pi, R_i, T)$ replaced by $\inf_{t \in T} |\pi_i^*(t) - \pi_{(K+1)}^*(t)|$, where $\pi_{(K+1)}^*(t)$ represents the $(K+1)$-th largest score at time point t.

4.3 Dynamic Rank Inference Using Signed Differences (DRIS)

Since the relationship between score differences and rank positions is inherently flexible, the combinatorial inference of the ranking properties is conservative, which is a challenge that persists even in static scenarios (Liu et al., 2023; Mogstad et al., 2023; Fan et al., 2024). To address the limitation, we delve deeper into the interplay

between scores and ranks and propose the creation of precise bands based on a novel signed score difference perturbation.

We define the perturbation that

$$\begin{aligned} S^{\dagger} =& \sup_{t\in T} \max_{j:j\neq i} \gamma_{ij}\, s_j(t)(\pi_i^*(t) - \pi_j^*(t) - \hat{\pi}_i(t) + \hat{\pi}_j(t)) \\ =& \sup_{t\in T} \max_{j:j\neq i} (-s_j(t)) \sqrt{\frac{1}{M n_{ij}}} \sum_{k=1}^{M} \sum_{l=1}^{n_{ij}} x_{kl}^{(ij)}(t) + o_p(1), \end{aligned}$$

where $s_j(t) = \text{sign}(\hat{\pi}_j(t) - \hat{\pi}_i(t))$. Based on $S^{\dagger}$, we construct the confidence band $\{[R_l^{\dagger}(t), R_u^{\dagger}(t)]\}_{t\in T}$, where $R_l^{\dagger}(t) = 1 + \sum_{j:s_j(t)>0} \mathbb{1}(\hat{\pi}_i(t) - \hat{\pi}_j(t) + \frac{S_{1-\alpha}^{\dagger}}{\gamma_{ij}} < 0)$ and $R_u^{\dagger}(t) = n - \sum_{j:s_j(t)<0} \mathbb{1}(\hat{\pi}_i(t) - \hat{\pi}_j(t) - \frac{S_{1-\alpha}^{\dagger}}{\gamma_{ij}} > 0)$. Proposition 3 verifies the effectiveness of the confidence band.

Proposition 3. *Let* $E_3 = \{\text{for all } t \in T, rr_i^*(t) \in [R_l^{\dagger}(t), R_u^{\dagger}(t)]\}$ *and* $E_4 = \{\sup_{t\in T} \sup_{j:j\neq i} s_j(t)(\pi_i^*(t) - \pi_j^*(t) - \hat{\pi}_i(t) + \hat{\pi}_j(t)) \leq \frac{S_{1-\alpha}^{\dagger}}{\gamma_{ij}}\}$, *then* $E_4 \subset E_3$.

Proposition 3 ensures the coverage probability of the interval, as indicated by $P(E_3) \geq P(E_4) \geq 1 - \alpha$. Furthermore, we demonstrate that the newly imposed sign $s_j(t)$ induces a narrower confidence band. Specifically, we can obtain that

$$\begin{aligned} R_l^{\dagger}(t) =& 1 + \sum_{j:s_j(t)>0} \mathbb{1}(\hat{\pi}_i(t) - \hat{\pi}_j(t) + \frac{S_{1-\alpha}^{\dagger}}{\gamma_{ij}} < 0) \\ \geq& 1 + \sum_{j:s_j(t)>0} \mathbb{1}(\hat{\pi}_i(t) - \hat{\pi}_j(t) + \frac{S_{1-\alpha}}{\gamma_{ij}} < 0) \\ =& 1 + \sum_{j:j\neq i} \mathbb{1}(\hat{\pi}_i(t) - \hat{\pi}_j(t) + \frac{S_{1-\alpha}}{\gamma_{ij}} < 0) = R_l(t). \end{aligned} \tag{4.5}$$

The second line holds since $S^{\dagger}_{1-\alpha} \leq S_{1-\alpha}$. We have $R^{\dagger}_u(t) \leq R_u(t)$ following a similar manner. We then consider the estimation of $S^{\dagger}$. It's noteworthy that $S^{\dagger}$ exhibits greater complexity compared to S. While S signifies the supremum of $\hat{\pi}_i(t) - \hat{\pi}_j(t) - \pi^*_i(t) + \pi^*_j(t)$, which can be uniformly approximated by the sum of independent variables, $s_j(t)$ in $S^{\dagger}$ is discontinuous and entails an intricate relationship with $\hat{\pi}_i(t) - \hat{\pi}_j(t) - \pi^*_i(t) + \pi^*_j(t)$. This complexity poses challenges in studying the distribution of $S^{\dagger}$.

We use the Gaussian multiplier bootstrap to approximate $S^{\dagger}$. Let

$$V^{\dagger} = \max_{t\in\mathcal{T}} \max_{j:j\neq i} (-s_j(t)) \sqrt{\frac{1}{Mn_{ij}}} \sum_{k=1}^{M} \sum_{l=1}^{n_{ij}} \hat{x}^{(ij)}_{kl}(t) z_{kl}.$$

By substituting $s_j(t)$ with $s^*_j := \text{sign}(\pi^*_j(t) - \pi^*_i(t))$, Theorem 4 holds for $S^{\dagger}$ and $V^{\dagger}$ with a simple modification. However, due to the discontinuity and randomness of $s_j(t)$, it is challenging to provide an analog theorem for $s_j(t)$. We omit the rigorous proof and consider it an approximation based on the given $s_j(t)$. Substituting $S^{\dagger}_{1-\alpha}$ with $V^{\dagger}_{1-\alpha}$, we obtain the modified confidence band $\{[\hat{R}^{\dagger}_l(t), \hat{R}^{\dagger}_u(t)]\}_{t\in T}$.

Accordingly, we have the one-sided confidence band $\{[1, \tilde{R}^{\dagger}_u(t)]\}_{t\in T}$, where $\tilde{R}^{\dagger}_u(t) = n - \sum_{j:s_j(t)<0} \mathbb{1}(\hat{\pi}_i(t) - \hat{\pi}_j(t) - \frac{W^{\dagger}_{1-\alpha}}{\gamma_{ij}} > 0)$ and

$$\begin{aligned} W^{\dagger} =& \sup_{t\in T} \max_{j:s_j(t)<0} \gamma_{ij}(\hat{\pi}_i(t) - \hat{\pi}_j(t) - \pi^*_i(t) + \pi^*_j(t)) \\ =& \sup_{t\in T} \max_{j:s_j(t)<0} \sqrt{\frac{1}{Mn_{ij}}} \sum_{k=1}^{M} \sum_{l=1}^{n_{ij}} x^{(ij)}_{kl}(t) + o_p(1), \end{aligned}$$

If α is less than 0.5, the band is asymptotically narrower than $\{[1, \tilde{R}_u(t)]\}_{t\in T}$ in that

$$\begin{aligned}\tilde{R}^{\dagger}_u(t) =& n - \sum_{j:s_j(t)<0} \mathbb{1}(\hat{\pi}_i(t) - \hat{\pi}_j(t) - \frac{W^{\dagger}_{1-\alpha}}{\gamma_{ij}} > 0)\\ \leq& n - \sum_{j:s_j(t)<0} \mathbb{1}(\hat{\pi}_i(t) - \hat{\pi}_j(t) - \frac{W_{1-\alpha}}{\gamma_{ij}} > 0)\\ =& n - \sum_{j:j\neq i} \mathbb{1}(\hat{\pi}_i(t) - \hat{\pi}_j(t) - \frac{W_{1-\alpha}}{\gamma_{ij}} > 0) = \tilde{R}_u(t).\end{aligned}$$

The second line holds since $W^{\dagger} \leq W$ and $W_{1-\alpha} \geq 0$ asymptotically when $\alpha \leq 0.5$. By substituting $W^{\dagger}$ with its corresponding Gaussian analog $U^{\dagger} = \max_{t\in\mathcal{T}} \max_{j:s_j(t)<0} \sqrt{\frac{1}{Mn_{ij}}} \sum_{k=1}^{M} \sum_{l=1}^{n_{ij}} \hat{x}^{(ij)}_{kl}(t) z_{kl}$, we can obtain the confidence band estimation.

Remark 6. We point out that the DRIS test framework is universally applicable to score-based ranking models like Elo rating (Elo, 1967) beyond the dynamic BT model. Provided the item ability estimations and their variances, the distribution of $S^{\dagger}$ can be approximated through the bootstrap technique, rendering the validity of the framework.

Remark 7. The test procedure can be modified for situations where the ranking property R_i is only relevant to specific items $I \subset [n]$. An example of such a situation arises when comparing global universities and aiming to draw inferences within a specific region. For instance, suppose we wish to determine whether King's

College London ranks among the top 3 to 5 universities in the United Kingdom. In this case, I represents all universities in the UK. It is important to note that we typically do not directly exclude data related to other items, as they can provide additional information and help mitigate the risk of disconnection between items. We can adjust the perturbation term as $\bar{S} = \sup_{t\in T} \max_{j\in I\setminus\{i\}} \gamma_{ij}(\hat{\pi}_i(t) - \pi_i^*(t) - \hat{\pi}_j(t) + \pi_j^*(t))$, and maintaining the above test framework. The type I error can be controlled through a simple modification of Theorem 5. Additionally, by noticing that we enlarge the rejection region, the consistency of the test still holds.

Remark 8. Notice that if all pairwise comparisons appear without timestamps and we set T as a single point, the dynamic ranking inference methods (DRI and DRIS) reduce to a static scenario, which we refer to as SRI and SRIS. It's important to emphasize that SRIS, based on a novel signed perturbation statistic $S^\dagger$, is more powerful than the SRI proposed in Liu et al. (2023), Mogstad et al. (2023), and Fan et al. (2024), while still maintaining the coverage probability and test level. We demonstrate this in (4.5) and further validate it in Section 5.1.

5. Simulation Study

In our simulations, we observe effective control over type I errors and increasing test power trends as sample sizes grow in all proposed test frameworks. We provide detailed simulation results in the supplementary material, while we emphasize a

compelling comparison with existing methods for the static case in this section.

5.1 Comparisons in the Static Case

As pointed out in Remark 8, the test procedure can be naturally adapted to the static case. In this section, we demonstrate that SRIS is significantly more powerful than SRI proposed in Liu et al. (2023).

We set $n = 100$, $p = 1$, $\pi_i^* = \exp(\theta_i^*)$ and consider two sets of parameters. For case 1, the parameter θ_i^* is set to 10 for i ranging from 1 to 30, then decreases by δ to 10-δ for $i = 31$, and finally becomes 7.5 for i ranging from 32 to 100. For case 2, the parameter θ_i^* is set to 12 for i ranging from 1 to 15, remains at 10 for i ranging from 16 to 30, decreases by δ to 10-δ for $i = 31$, and finally becomes 8 for i ranging from 32 to 100. The first setting is similar to those in Liu et al. (2023), and case 2 is a more complex scenario. We employ 500 bootstrap repetitions and repeat 500 times. We set $\delta = 10^{-5}$ and test whether item 31 is ranked among the top 30 items under H_0. For the empirical power, we vary δ and test whether item 30 ranks in the top 30. We set $\gamma_{ij} = \sqrt{n^3 pMh}$, $i, j \in [n], i \neq j$ for the simplicity of computation. In practice, we observe that using either standard deviation estimation or the constants exhibits similar empirical performances.

Table 1 indicates that for both sets of parameters, the type I errors are effectively controlled close to 0.05 for both methods. However, in the second set of

Table 1: Type I error of top-K test in the static cases.

	Case1		Case2	
M	SRI	SRIS	SRI	SRIS
50	0.026	0.018	0.000	0.040
100	0.032	0.030	0.000	0.040
200	0.044	0.040	0.000	0.042
300	0.020	0.020	0.000	0.048
500	0.050	0.034	0.000	0.054

parameters, the SRI method appears to be overly conservative and does not exhibit the expected trend toward the nominal significance level of 0.05.

Based on Figure 1, it is evident that SRIS outperforms SRI in terms of statistical power, particularly when considering the second set of parameters. The difference in power between the two methods is quite pronounced, with SRIS demonstrating a higher increasing speed compared with SRI.

6. Empirical Study

We collected the NBA regular season game results from season 2008-2009 to 2018-2019, which comprises 13,289 pairwise comparison results among 30 teams. We set the bandwidth to 0.1. We calculate the p-values for each team as described in Section 3.1, which are represented by the dashed line in Figure 2. The solid

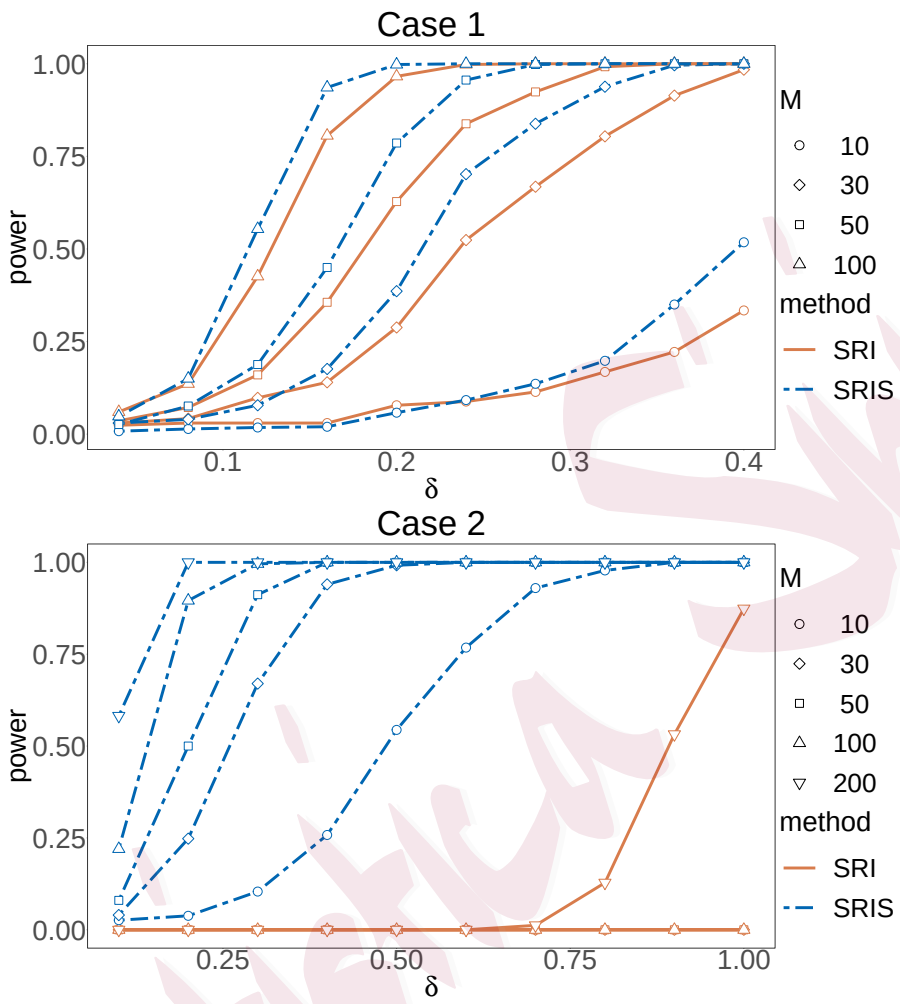

Figure 1: Test power of top-K test in the static cases.

line depicts the threshold determined by the Benjamini-Yekutieli Procedure. We observe that all p-values are considerably smaller than 0.05, except for the team POR, whose p-value is 0.099982. Notably, the p-values for 29 teams fall below the threshold line, suggesting that adopting the dynamic model is reasonable.

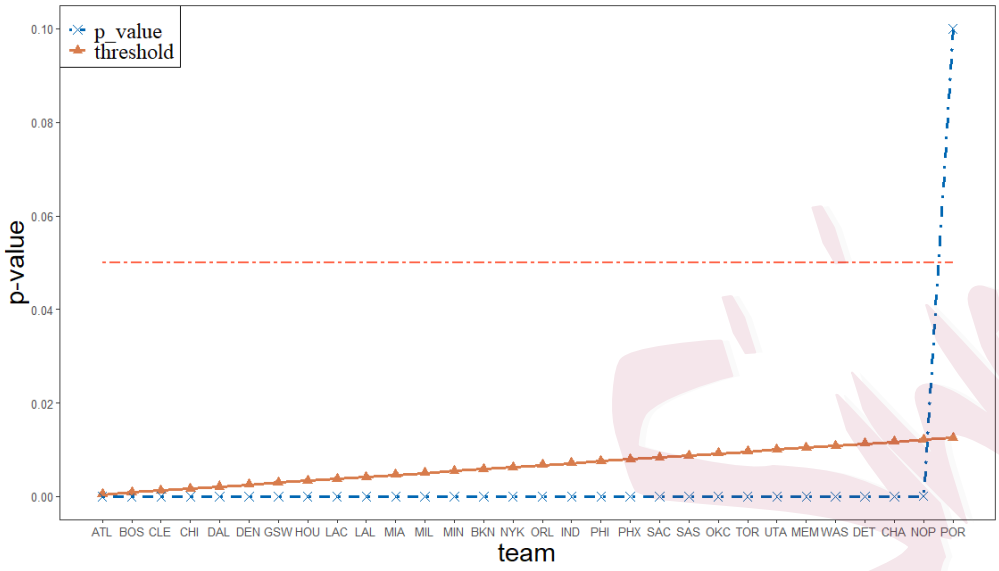

Figure 2: The p-values of the score function invariance test for the NBA data.

Next, we test whether the score functions of two selected teams are equal. We consider three pairs: MIN and SAC, DET and CHA, and MIN and SAS. The corresponding score function estimations are displayed in Figure 3, and the resulting p-values are 0.967391, 0.023472, and 0, respectively. Consequently, we can't reject the null hypothesis that the teams have the same score function for the first pair (MIN and SAC), while we reject the null hypotheses for the last two pairs (DET and CHA, MIN and SAS), which aligns with the direct impression in Figure 3.

For ranking inference, we employ the DRIS method. We first conduct top3, top5, top10, and top15 tests for team SAS at the midpoint of each season. We set

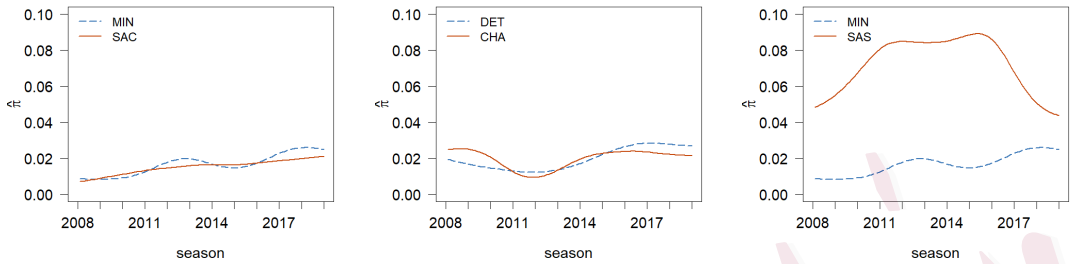

Figure 3: Score function estimators of three selected pairs in the NBA data.

the number of bootstrap repetitions to 1000. The results are presented in Table 2. We then test whether a team consistently ranks top K for a selected season. We set $v = 10$ and perform 1000 bootstraps for quantile estimation. Table 3 presents the resulting p-values $\inf\{\alpha : \cap_{t\in\mathcal{T}} \cap_{\pi\in\tilde{\Pi}_t(\alpha)} \cap_{r\in rr(\pi)}\{r \in R_i(t)\}\}$ for team i, which is a direct extension of the p-values in Liu et al. (2023). The findings align with actual competition outcomes. For example, team GSW had a winning rate of 81.7% in the 2016-2017 season, while the winning rate of BKN was 46.3% in the 2014-2015 season.

7. Conclusion

This paper focuses on a series of inferential problems of the dynamic BT model and ranking properties. We first address a fundamental and critical concern, distinguishing between static and dynamic cases. Subsequently, we develop a test

Table 2: The p-values of top-K test for team SAS at time points.

Time point	Top 3	Top 5	Top 10	Top 15
0809Mid	1.000	1.000	0.527	0.044
0910Mid	0.970	0.895	0.272	0.044
1011Mid	0.521	0.400	0.025	0.001
1112Mid	0.088	0.060	0.002	0.001
1213Mid	0.241	0.026	0.004	0.000
1314Mid	0.254	0.126	0.004	0.000
1415Mid	0.060	0.022	0.002	0.000
1516Mid	0.017	0.003	0.000	0.000
1617Mid	0.077	0.040	0.000	0.000
1718Mid	0.984	0.677	0.156	0.010
1819Mid	1.000	1.000	0.930	0.163

Table 3: The p-values of the top-K test at time intervals.

Team	Season	Test type	P-value
GSW	2016-2017	top 3	0.047
SAS	2015-2016	top 5	0.021
LAL	2009-2010	top 15	0.063
ORL	2009-2010	top 15	0.311
BKN	2014-2015	top 20	1.000

procedure to identify identical individuals, offering a fresh perspective for the clustering problem in the BT model. Furthermore, we propose an innovative dynamic ranking inference scheme. In addition to establishing a novel confidence band of dynamic rank, we present a test procedure for general dynamic ranking properties. The DRIS method employs a novel signed score differences statistic, effectively addressing the overly conservative issue, and the framework is widely applicable to various score estimation methods. The extensive experiments substantiate the satisfactory performance of our methods.

Supplementary Materials

The supplementary material contains additional simulation results and proofs.

References

Bazylik, S., M. Mogstad, J. Romano, A. Shaikh, and D. Wilhelm (2024). Finite-and large-sample inference for ranks using multinomial data with an application to ranking political parties. *arXiv e-prints*, arXiv–2402.

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

Bong, H., W. Li, S. Shrotriya, and A. Rinaldo (2020). Nonparametric estimation in the dynamic Bradley-Terry model. In *International Conference on Artificial Intelligence and Statistics*, pp. 3317–3326.

PMLR.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*(3/4), 324–345.

Cambazoglu, B. B., H. Zaragoza, O. Chapelle, J. Chen, C. Liao, Z. Zheng, and J. Degenhardt (2010). Early exit optimizations for additive machine learned ranking systems. In *Proceedings of the third ACM international conference on Web search and data mining*, pp. 411–420.

Chen, Y., J. Fan, C. Ma, and K. Wang (2019). Spectral method and regularized MLE are both optimal for top-K ranking. *Annals of Statistics 47*(4), 2204–2235.

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

de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields 75*, 261–277.

Duh, K. and K. Kirchhoff (2008). Learning to rank with partially-labeled data. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 251–258.

Elo, A. E. (1967). The proposed uscf rating system, its development, theory, and applications. *Chess life 22*(8), 242–247.

Fan, J., Z. Lou, W. Wang, and M. Yu (2024). Ranking inferences based on the top choice of multiway comparisons. *Journal of the American Statistical Association 0*(0), 1–14.

Gao, C., Y. Shen, and A. Y. Zhang (2023). Uncertainty quantification in the Bradley-Terry-Luce model. *Information and Inference: A Journal of the IMA 12*(2), 1073–1140.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society Series C: Applied Statistics 48*(3), 377–394.

Glickman, M. E. (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics 28*(6), 673–689.

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

Huang, T.-K., R. C. Weng, and C.-J. Lin (2006). Generalized Bradley-Terry models and multi-class probability estimates. *Journal of Machine Learning Research 7*(4), 85–115.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *Annals of Statistics 32*(1), 384–406.

Karlé, E. and H. Tyagi (2023). Dynamic ranking with the btl model: a nearest neighbor based rank centrality method. *Journal of Machine Learning Research 24*(269), 1–57.

Liu, Y., E. X. Fang, and J. Lu (2023). Lagrangian inference for ranking problems. *Operations Research 71*(1), 202–223.

Lv, Y., T. Moon, P. Kolari, Z. Zheng, X. Wang, and Y. Chang (2011). Learning to model relatedness for news recommendation. In *Proceedings of the 20th international conference on World wide web*, pp. 57–66.

REFERENCES

Maystre, L., V. Kristof, and M. Grossglauser (2019). Pairwise comparisons with flexible time-dynamics. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1236–1246.

McHale, I. and A. Morton (2011). A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting 27*(2), 619–630.

Mogstad, M., J. P. Romano, A. M. Shaikh, and D. Wilhelm (2023, 01). Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *The Review of Economic Studies 91*(1), 476–518.

Negahban, S., S. Oh, and D. Shah (2017). Rank centrality: Ranking from pairwise comparisons. *Operations Research 65*, 266–287.

Simons, G. and Y.-C. Yao (1999). Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *Annals of Statistics 27*(3), 1041–1060.

Tian, X., J. Shi, X. Shen, and K. Song (2024). A spectral approach for the dynamic bradley–terry model. *Stat 13*(3), e722.

Wang, Y., L. Wang, Y. Li, D. He, and T.-Y. Liu (2013). A theoretical analysis of NDCG type ranking measures. In *Conference on learning theory*, pp. 25–54. PMLR.

Yan, T., Y. Yang, and J. Xu (2012). Sparse paired comparisons in the Bradley-Terry model. *Statistica Sinica 22*(3), 1305–1318.

Nan Lu

REFERENCES

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China; School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

E-mail: lunan2021@amss.ac.cn

Jian Shi

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China; School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

E-mail: jshi@iss.ac.cn

Xin-Yu Tian

School of Statistics, University of Minnesota, Minneapolis, USA

E-mail: tianx@umn.edu

Kai Song

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

E-mail: kaisong@amss.ac.cn