Statistica Sinica Preprint No: SS-2025-0318	
Title	Semiparametric Analysis for Paired Comparisons with Covariates
Manuscript ID	SS-2025-0318
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0318
Complete List of Authors	Haoyue Song, Lianqiang Qu, Ting Yan and Yuguo Chen
Corresponding Authors	Ting Yan
E-mails	tingyanty@mail.ccnu.edu.cn
Notice: Accepted author version.	

Statistica Sinica

SEMIPARAMETRIC ANALYSIS FOR PAIRED COMPARISONS WITH COVARIATES

Haoyue Song[1], Lianqiang Qu[1], Ting Yan[1], Yuguo Chen[2]

[1] *Central China Normal University*

[2] *University of Illinois Urbana-Champaign*

Abstract: Statistical inference in parametric models (e.g., the Bradley–Terry model and its variants) for paired-comparison data has been explored in the high-dimensional regime, in which the number of items involving in paired comparisons diverges. However, parametric models are highly susceptible to model misspecification. To relax the assumption of known distributions and provide flexibility, we propose a semiparametric framework for modeling the merits of items and covariate effects (e.g., home-field advantage) by introducing latent random variables with unspecified distributions. As the number of parameters increases with the number of items, semiparametric inference is highly nontrivial. To address this issue, we employ a kernel-based least squares approach to estimate all unknown parameters. When each pair of items has a fixed number of comparisons and the number of items tends to infinity, we prove the consistency of all resulting estimators and derive their asymptotic normal distributions. To the best of our knowledge, this is the first study to conduct a semiparametric analysis of paired comparisons with an increasing dimension. We conduct sim-

ulations to evaluate the finite-sample performance of the proposed method and illustrate its practical utility by analyzing an NBA dataset.

Key words and phrases: Asymptotic normality, Consistency, Covariate effects, Paired comparison, Semiparametric model.

1. Introduction

Paired comparison is a popular approach for ranking a set of items, where the preference between two items is judged each time. This technique is especially useful when the evaluation criteria are objective by nature. Paired-comparison data are naturally generated in sports tournaments, where multiple teams compete against each other. A central problem in paired-comparison analysis is quantitatively characterizing the merits of items to derive a ranking. In balanced-comparison designs, where each pair of items has the same number of comparisons, the ranking is based on the win counts. However, in unbalanced-comparison designs, there is no natural ranking. To address this issue, statistical models, such as the Bradley–Terry (Bradley and Terry, 1952) and Thurstone (Thurstone, 1927) models, are employed to estimate the merit parameters of items in unbalanced comparisons, and rank is produced according to their estimated values. The applications of paired-comparison include the ranking of classical

sports teams (e.g. Masarotto and Varin, 2012), scientific journals (Stigler, 1994; Varin et al., 2016), product brands (Radlinski and Joachims, 2007), and crowd-sourced labels (Chen et al., 2016).

Classical paired-comparison models, including the Bradley–Terry and the Thurstone models, assign a merit parameter θ_i to each item and assume that the probability of one item defeating another item depends only on the relative difference in their merit parameters. As argued by David (1988, page 7), the merit of an item i can be represented by a latent random variable ϵ_i, capturing uncertainty across observations. In a paired comparison of items i and j, items i wins if $\theta_i - \theta_j > \epsilon_i - \epsilon_j$; otherwise, items j wins. Parametric models require the distribution of ϵ_i to be known. When ϵ_i follows a normal or doubly exponential distribution, it corresponds to the Thurstone or Bradley–Terry model (David, 1988, page 8), respectively. However, the distribution of ϵ_i is usually unknown. Hence, it may lead to incorrect inference if not correctly specified. Moreover, covariate information may accompany paired-comparison data. For example, a team is more likely to win when playing in its home city in most sports, which is referred to as the home-field advantage (e.g. Agresti, 2012, page 455). The winning rate of a team in the previous regular season could also be used to predict its performance in the current season. Ignoring the covariate effects may

also lead to invalid inference, as demonstrated in Yan (2025).

In this study, we propose a general semiparametric framework for modeling the merits of items and covariate effects, when the distribution of the latent random variable is unknown. Let $\boldsymbol{X}_{ijt} = (X_{ijt,0}, X_{ijt,1}, \ldots, X_{ijt,p})^\top$ denote the covariate information associated with the tth comparison of item i against j, where p is fixed. We require $\boldsymbol{X}_{ijt} = -\boldsymbol{X}_{jit}$ because if something is advantageous to i, then it is disadvantageous to j. We adopt a random-effects design for the covariates $\boldsymbol{X}_{ijt}$, that is, $\boldsymbol{X}_{ijt}$ are random vectors. Let a_{ijt} indicate whether item i defeats item j in the tth comparison: $a_{ijt} = 1$ if item i defeats item j and 0 otherwise. The semiparametric framework assumes that the winning probability of i against j conditional on $\boldsymbol{X}_{ijt}$ is

$$\mathbb{P}(a_{ijt} = 1 | \boldsymbol{X}_{ijt}, \boldsymbol{\gamma}, \theta_i, \theta_j) = F(\theta_i - \theta_j + \boldsymbol{X}_{ijt}^\top \boldsymbol{\gamma}), \tag{1.1}$$

where $\boldsymbol{\gamma}$ is a $(p+1)$-dimensional regression coefficient of the covariates, θ_i is the merit parameter of item i. Here, $F(x)$ denotes the cumulative distribution function of latent random variables that satifying $F(x) + F(-x) = 1$. Under the restriction $\boldsymbol{X}_{ijt} = -\boldsymbol{X}_{jit}$, the probability distribution in (1.1) is well defined. Hereafter, we call it the semiparametric paired-comparison model. When $F(x)$ is the cumulative distribution function of a logistic

or standard normal variable, it becomes the Bradley–Terry or Thurstone model, respectively.

The covariate $\boldsymbol{X}_{ijt}$ can be defined based on the situations of the teams or the items' attributes. If $\boldsymbol{W}_{it,j}$ and $\boldsymbol{W}_{jt,i}$ denote q-dimensional attributes of items i and j in the tth comparison between items i and j, respectively, they can be used to construct the vector $\boldsymbol{X}_{ijt} = \boldsymbol{g}(\boldsymbol{W}_{it,j}, \boldsymbol{W}_{jt,i})$, where $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = -\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x})$. For instance, if we let $\boldsymbol{g}(\boldsymbol{W}_{it,j}, \boldsymbol{W}_{jt,i})$ equal to $\boldsymbol{W}_{it,j} - \boldsymbol{W}_{jt,i}$, then it measures the dissimilarity between the two items. As an example with a one-dimensional covariate, consider the tth game where team i plays at home against team j (the away team), then we let $W_{it,j} = 1$ and $W_{jt,i} = 0$, such that $X_{ijt} = 1$ and $X_{jit} = -1$.

We obtain the conditions for model identification employing a special regressor method (Lewbel, 1998, 2000). Motivated by Qu et al. (2026), we develop a kernel-based least squares approach for estimating the unknown parameters in semiparametric paired-comparison models. We employ a projection matrix to project the covariates onto the subspace spanned by the column vectors of the design matrix made up of merit parameters, and obtain an explicit estimator for the covariate parameter. We then estimate the merit parameters by using a least squares method. The projection procedure eliminates the potential bias of the estimator for the covariate

parameter. We establish consistency and asymptotic normality of the resulting estimators, when the number of items approaches infinity and each pair has a fixed number of comparisons. Numerical studies and a real-world data analysis demonstrate our theoretical findings.

1.1 Literature review

The Bradley–Terry model and its generalized versions have been examined in the high-dimensional regime, where the number of items is large. Simons and Yao (1999) established the consistency and asymptotic normality of the maximum likelihood estimator (MLE) in the Bradley–Terry model under a dense comparison assumption, in which each pair has a fixed number of comparisons. Yan et al. (2012) and Han et al. (2020) generalized the result of Simons and Yao (1999) to a sparse comparison case where paired comparisons exist only for some pairs. Chen et al. (2019) found that the spectral method or the regularized MLE is minimax optimal in terms of the sample complexity—that is, the number of paired comparisons needed to ensure exact top-K identification. Chen et al. (2022) proved that the MLE achieved optimal partial and exact recovery for the top-K ranking problem, while the spectral method is, in general, sub-optimal. Han et al. (2023) proved consistency of the MLE in a class of generalized Bradley–

Terry models. Yan et al. (2025) established Wilks-type results for likelihood ratio tests under some increasing and fixed dimensional null hypotheses. Hunter (2003) developed a minorization-maximization method to obtain the MLE under generalized Bradley–Terry models.

Recently, Fan et al. (2024) extended the Bradley–Terry model to incorporate the covariate information and derived the asymptotic properties of the MLE under an Erdös–Rényi comparison graph, where the covariate X_i of item i enters the model additively. However, this model characterizes only the individual level covariate information and does not address those covariates associated with each paired comparisons (e.g., home-field advantage). Singh et al. (2025) obtained least squares estimators for cardinal paired-comparison data with covariates and established their large sample theory. Fan et al. (2025) studied the ranking problem in a modified version of the Plackett-Luce model for the top choice of multiway comparisons. Dong et al. (2024) explored MLE in a covariate-assisted Plackett–Luce model. However, these works are based on parametric models, while we focus on semiparametric inference for paired-comparison data with covariates, in which the error distribution is unspecified.

Paired-comparison data can be represented as a weighted directed graph, where nodes denote items, and a weighted directed edge from the head node

i to the tail node j denotes the number of wins by i over j out of all their comparisons. Different from the concept of merit parameters in paired comparisons, degree heterogeneity parameters are used to measure the intrinsic nodal merits in forming network connections. Several parametric models have been proposed to characterize degree heterogeneity and covariate effects in networks (e.g., Yan et al., 2019; Dzemski, 2019), where the estimation and inference methods depend on a specified distributional assumption for the edges. To relax the known distribution assumption, semiparametric models have also been developed (Zeleneev, 2020; Candelaria, 2020; Qu et al., 2026). Our study is motivated by Qu et al. (2026), which introduced a semiparametric framework for directed network formation. However, they focus on in- and out-degree heterogeneity and homophily effects, which are different from paired-comparison data.

The remainder of this paper is organized as follows. Section 2 presents the semiparametric paired-comparison model, establishes identification conditions, and develops a kernel-based least squares estimator. Section 3 establishes the consistency and central limit theorem for the proposed estimator. Section 4 contains numerical studies and a real-data application. Section 5 gives some summarizes. The technical details and additional numerical results are in the Supplementary Material.

Notations: Let $[n]_0 := \{0, 1, \ldots, n\}$ and $[n] := \{1, \ldots, n\}$. Let $\boldsymbol{e}_i$ be an n-dimensional standard basis vector with the ith element 1 and 0 otherwise for $i = 1, \ldots, n$. Further, we define $\boldsymbol{e}_0 = (0, \ldots, 0)^\top$ as the n-dimensional zero vector. Define $\mathbf{1}_n$ as the n-dimensional vector with all elements equal to 1. For any $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, we define the ℓ_2-norm $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i\in[n]} x_i^2}$ and the ℓ_∞-norm $\|\boldsymbol{x}\|_\infty = \max_{i\in[n]} |x_i|$. For any $k \in [p+1]$, we define $\boldsymbol{x}_{(-k)} = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)^\top \in \mathbb{R}^{n-1}$. Let $\boldsymbol{I}_{n\times n}$ be the $n \times n$ identity matrix with 1's on the diagonal and 0's elsewhere, and $\mathbb{I}(\cdot)$ denote the indicator function. For any matrix $\boldsymbol{A}$, we define $\lambda_{max}(\boldsymbol{A})$ and $\lambda_{min}(\boldsymbol{A})$ as its largest and smallest eigenvalues, respectively. For a matrix $\boldsymbol{A} = (A_{ij})_{n\times m} \in \mathbb{R}^{n\times m}$, we define $\|\boldsymbol{A}\|_{\max} = \max_{i\in[n],j\in[m]} |A_{ij}|$, $\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}$. For the positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$ if $a_n/b_n \to 0$, as $n \to \infty$; and $a_n = O(b_n)$ if there exists a constant C, such that $a_n \le Cb_n$ for all n. We use the superscript "$*$" to denote the true parameter, under which the data are generated.

2. Model, Identifiability and Estimation

2.1 Semiparametric paired-comparison model

Assume that there are $n+1$ items, labeled as $0, 1, 2, \ldots, n$, involving in the paired comparisons. Let T_{ij} denote the number of comparisons between

items i and j. For convenience, let $T_{ii} = 0$ for $i \in [n]_0$. For easy exposition, we set $T_{ij} = T$ for all $0 \le i \ne j \le n$, where T is a fixed positive integer. Our results can be easily extended to the case $K_1 \le T_{ij} \le K_2$ for two fixed positive integers K_1 and K_2. Define $T_i = \sum_{j=0, j\ne i}^{n} T_{ij}$ as the total number of comparisons of item $i \in [n]_0$. Let a_i be the number of wins for item i, defined as the sum of all its comparison outcomes: $a_i = \sum_{j=0, j\ne i}^{n} \sum_{t\in[T]} a_{ijt}$.

Recall that $\boldsymbol{X}_{ijt} = (X_{ijt,0}, X_{ijt,1}, \ldots, X_{ijt,p})^\top \in \mathbb{R}^{p+1}$ denotes the covariate information associated with the tth comparison of item i against j, which may affect the outcomes. We assume that a_{ijt} are conditionally independent across $0 \le i < j \le n$ and $1 \le t \le T$, given all the covariates $\{\boldsymbol{X}_{ijt}\}_{i,j,t}$.

According to the semiparametric paired-comparison model specified in (1.1), a_{ijt} can be represented as

$$a_{ijt} = \mathbb{I}(\theta_i - \theta_j + \boldsymbol{X}_{ijt}^\top \boldsymbol{\gamma} > \varepsilon_{ijt}) \quad \text{for} \quad i < j \quad \text{and} \quad t \in [T], \tag{2.1}$$

where θ_i ($i \in [n]_0$) are unknown merit parameters, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_p)^\top$ is the regression coefficient vector for the covariates, and ε_{ijt} denotes the latent noise. Under model (2.1), we observe that the larger the parameter θ_i is, the more likely for item i to win. Therefore, θ_i describes the merit

of item i in paired comparisons. The parameter $\boldsymbol{\gamma}$ captures the covariate effects on the comparison outcomes. In the case of home-field advantage, if the parameter $\gamma > 0$, item i has a larger probability to win. The noise term ε_{ijt} represents unobservable factors that influence the comparison outcomes. The distribution of ε_{ijt} is left unspecified.

2.2 Identifiability and Estimation

In this section, we first address the identifiability problem of model (2.1), which is defined over the joint distribution of covariates and the noises. Clearly, for any $c_1 > 0$ and $c_2 \in \mathbb{R}$, we have

$$\begin{aligned} a_{ijt} =& \mathbb{I}(\theta_i - \theta_j + \boldsymbol{X}_{ijt}^\top \boldsymbol{\gamma} > \varepsilon_{ijt}) \\ =& \mathbb{I}\{(c_1\theta_i + c_2) - (c_1\theta_j + c_2) + c_1 \boldsymbol{X}_{ijt}^\top \boldsymbol{\gamma} > c_1 \varepsilon_{ijt}\}. \end{aligned}$$

Therefore, model (2.1) is unidentifiable without constraints. A common approach to avoid this issue is to set $\sum_{i=0}^{n} \theta_i = 0$ or $\theta_0 = 0$ and $\gamma_k = 1$, where γ_k is the kth component of $\boldsymbol{\gamma}$ and k is chosen such that $X_{ijt,k}$ is a continuous random variable.

The identifiability of model (2.1) is also related to the support of the joint distribution of $(\boldsymbol{X}_{ijt}, \varepsilon_{ijt})$. For illustration, we consider a toy example. Let $\boldsymbol{X}_{ijt} \in \mathbb{R}$ be a random variable with support $(-5, -2) \cup (2, 5)$. We set

$\boldsymbol{\gamma} = \widetilde{\boldsymbol{\gamma}} = 1$, $\theta_i - \theta_j = 1$ for $0 \le i \ne j \le n$, and $\widetilde{\theta}_i - \widetilde{\theta}_j = c_1(\theta_i - \theta_j)$, where $c_1 \in (0, 1)$. Further, let ε_{ij} and $\widetilde{\varepsilon}_{ij}$ be obtained from the uniform distribution on $(-1, 1)$. In this scenario, we have

$$\boldsymbol{X}_{ijt} > \varepsilon_{ij} - 1 \quad \text{if} \quad \boldsymbol{X}_{ijt} \in (2, 5) \quad \text{and} \quad \boldsymbol{X}_{ijt} < \varepsilon_{ij} - 1 \quad \text{otherwise},$$

$$\boldsymbol{X}_{ijt} > \widetilde{\varepsilon}_{ij} - c_1 \quad \text{if} \quad \boldsymbol{X}_{ijt} \in (2, 5) \quad \text{and} \quad \boldsymbol{X}_{ijt} < \widetilde{\varepsilon}_{ij} - c_1 \quad \text{otherwise}.$$

This implies that $a_{ijt} = \widetilde{a}_{ijt}$, where $\widetilde{a}_{ijt} = \mathbb{I}\{\widetilde{\theta}_i - \widetilde{\theta}_j + \boldsymbol{X}_{ijt}^\top \widetilde{\boldsymbol{\gamma}} > \widetilde{\varepsilon}_{ij}\}$. Therefore, model (2.1) cannot be identified in the parameter set $\{0 < \theta_i - \theta_j < 1,\ 0 \le i \ne j \le n\}$. However, if we change the support of $\boldsymbol{X}_{ijt}$ to $(-5, 5)$, then the support of $\theta_i - \theta_j - \varepsilon_{ij}$ is a subset of $(-5, 5)$. Consequently, $\mathbb{P}(a_{ijt} \ne \widetilde{a}_{ijt}) > 0$ when $\theta_i - \theta_j \ne \widetilde{\theta}_i - \widetilde{\theta}_j$. In this case, model (2.1) is identifiable. This phenomenon is also observed in semiparametric network formation models (Qu et al., 2026).

We consider the following conditions for the identifiability of model (2.1).

Condition 1. There exists one continuous covariate $X_{ijt,k}$ such that its regression coefficient γ_k is positive. Additionally, the conditional distribution of $X_{ijt,k}$ given $X_{ijt,(-k)}$ is absolutely continuous with respect to the Lebesgue measure with nondegenerate conditional density $f(x|X_{ijt,(-k)})$, where $X_{ijt,(-k)} = (X_{ijt,0}, \ldots, X_{ijt,k-1}, X_{ijt,k+1}, \ldots, X_{ijt,p}) \in \mathbb{R}^p$.

The covariate $X_{ijt,k}$ defined in Condition 1 is referred to as a special regressor (Lewbel, 1998; Candelaria, 2020; Qu et al., 2026). For simplicity, we denote the special regressor as $X_{ijt,0}$ and its conditional density as $f(x|\boldsymbol{Z}_{ijt})$, where $\boldsymbol{Z}_{ijt} = (X_{ijt,1}, \ldots, X_{ijt,p})^\top \in \mathbb{R}^p$. Denote the regression coefficient of the covariate $\boldsymbol{Z}_{ijt}$ by $\boldsymbol{\eta} = (\gamma_1, \ldots, \gamma_p)^\top$, which is identical to γ excluding its first element. For convenience, denote

$$\overline{\boldsymbol{Z}} = \frac{1}{T}\sum_{t\in[T]} \boldsymbol{Z}_t, \;\; \boldsymbol{Z}_t = \left(\boldsymbol{Z}_{01t}, \ldots, \boldsymbol{Z}_{0nt}, \boldsymbol{Z}_{12t}, \ldots, \boldsymbol{Z}_{1nt}, \ldots, \boldsymbol{Z}_{(n-1)nt}\right)^\top \in \mathbb{R}^{N\times p}, \tag{2.2}$$

where $N = n(n+1)/2$.

Condition 2. The conditional density $f(x|\boldsymbol{Z}_{ijt})$ of $X_{ijt,0}$ given $\boldsymbol{Z}_{ijt}$ has support $(-B_U, B_U)$, where $B_U > 0$. Furthermore, the support of $-(\theta_i^* - \theta_j^* + \boldsymbol{Z}_{ijt}^\top\boldsymbol{\eta}^* - \varepsilon_{ijt})/\gamma_0^*$ is a subset of $(-B_U, B_U)$.

Condition 2 restricts the support for $X_{ijt,0}$, which is mild and has been widely adopted (e.g. Manski, 1985; Lewbel, 1998, 2000; Candelaria, 2020). When there are more than one "special regressor", based on the support requirement in Condition 2, we can use the covariate with the largest observed support among all possible candidates as the special regressor. This is a simple principle that can be easily carried out. Conditions 1 and 2 do not impose any restrictions on the distribution of $\boldsymbol{Z}_{ijt}$. Therefore, this

identification strategy accommodates discrete covariates in $\boldsymbol{Z}_{ijt}$.

Condition 3. ε_{ijt} is independent of $\boldsymbol{X}_{ijt}$, and satisfies $\varepsilon_{ijt} = -\varepsilon_{jit}$ and $\mathbb{E}(\varepsilon_{ijt}) = 0$ $(0 \le i < j \le n,\, 1 \le t \le T)$.

Recall that ε_{ijt} has a symmetric distribution assumption. Condition 3 is mild and can be relaxed to the scenario where ε_{ijt} is conditionally independent of $X_{ijt,0}$ given $\boldsymbol{Z}_{ijt}$.

Next, we introduce some notations. Recall that there are $n+1$ items involved in paired comparisons and we set $\theta_0 = 0$ for model identifiability. Let

$$\boldsymbol{U} = (\boldsymbol{U}_{01}, \ldots, \boldsymbol{U}_{0n}, \boldsymbol{U}_{12}, \ldots, \boldsymbol{U}_{1n}, \ldots, \boldsymbol{U}_{(n-1)n})^\top \in \mathbb{R}^{N \times n} \tag{2.3}$$

be the design matrix for the parameter vector $\boldsymbol{\theta}$, where $\boldsymbol{U}_{ij} = \boldsymbol{e}_i - \boldsymbol{e}_j \in \mathbb{R}^n$ and $\boldsymbol{e}_i \in \mathbb{R}^n$ is the standard basis vector with the ith element 1 and others 0. For convenience, define $\boldsymbol{e}_0$ as an n-dimensional column vector with all elements 0. Denote

$$\boldsymbol{V} = (v_{ij})_{n\times n} := \boldsymbol{U}^\top \boldsymbol{U} = (n+1)\boldsymbol{I}_{n\times n} - \mathbf{1}_n \mathbf{1}_n^\top \in \mathbb{R}^{n\times n}, \tag{2.4}$$

where $v_{ii} = n, i = 1, \ldots, n$ and $v_{ij} = v_{ji} = -1, 1 \le i \ne j \le n$. Here, $\boldsymbol{I}_{n\times n}$ denotes an $n \times n$ identity matrix and $\mathbf{1}_n$ an n-dimensional column vector

with all its elements 1. Thus, $\boldsymbol{V}$ can be viewed as the "reduced graph Laplacian" obtained by removing the first row and the first column of the graph Laplacian corresponding to the comparison graph of all $n+1$ items, where we assume that each pair is compared. The inverse matrix $\boldsymbol{V}^{-1}$ has an explicit expression:

$$\boldsymbol{V}^{-1} = \frac{1}{n+1}(\mathbf{1}_n\mathbf{1}_n^\top + \boldsymbol{I}_{n\times n}). \tag{2.5}$$

Define a projection matrix $\boldsymbol{D}$ indicating the orthogonal projection onto the column space of $\boldsymbol{U}$:

$$\boldsymbol{D} = \boldsymbol{I}_{N\times N} - \boldsymbol{U}\boldsymbol{V}^{-1}\boldsymbol{U}^\top \in \mathbb{R}^{N\times N}, \tag{2.6}$$

where the diagonal elements of $\boldsymbol{D}$ are $1-2/(n+1)$, and the off-diagonal elements are either $1/(n+1)$, $-1/(n+1)$, or 0. Note that

$$\boldsymbol{D}\boldsymbol{U} = (\boldsymbol{I}_{N\times N} - \boldsymbol{U}\boldsymbol{V}^{-1}\boldsymbol{U}^\top)\boldsymbol{U} = \boldsymbol{U} - \boldsymbol{U} = \mathbf{0} \in \mathbb{R}^{N\times n}.$$

Condition 4. There is a constant $\lambda > 0$ such that $\lambda_{min}(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}}/N) \geq \lambda$ almost surely.

Since $\boldsymbol{D}$ has an explicit expression, a direct calculation gives

$$\overline{\boldsymbol{Z}}^{\top}\boldsymbol{D}\overline{\boldsymbol{Z}} = \sum_{i<j}\overline{\boldsymbol{Z}}_{ij}\overline{\boldsymbol{Z}}_{ij}^{\top} - \frac{1}{n+1}\sum_{i=0}^{n}\Big(\sum_{j,k=0,j,k\neq i}^{n}\overline{\boldsymbol{Z}}_{ij}\overline{\boldsymbol{Z}}_{ik}^{\top}\Big),$$

whose detailed proofs is in Section S2 of the Supplementary Material. A sufficient condition to guarantee Condition 4 is that Z_{ijt}'s are independently generated from a p-dimensional non-degenerated multivariate symmetric distribution. Let Z denote a general random vector from this distribution. Under this condition, by the large sample theory, $N^{-1}\overline{\boldsymbol{Z}}^{\top}\boldsymbol{D}\overline{\boldsymbol{Z}}$ converges to $\mathrm{Cov}(Z)$ almost surely, which implies Condition 4. An intuitive explanation for Condition 4 is that the amount of information of the covariate matrix $\overline{\boldsymbol{Z}}$ projected into U^{+} is linearly proportional to the total amount of information, where U^{+} denotes the orthogonal complementary space of the linear subspace $\{\boldsymbol{U}\mathbf{x} : \mathbf{x} \in \mathbb{R}^{p}\}$. The independent and identically distributed assumption on covariates implies that this condition holds automatically. In addition, if $\overline{\boldsymbol{Z}}$ lies in $\{\boldsymbol{U}\mathbf{x} : \mathbf{x} \in \mathbb{R}^{p}\}$, then $\boldsymbol{D}\overline{\boldsymbol{Z}} = 0$ and Condition 4 fails. Condition 4 is related to the asymptotic behavior of an estimator of $\boldsymbol{\eta}$. If $\lambda_{min}(\overline{\boldsymbol{Z}}^{\top}\boldsymbol{D}\overline{\boldsymbol{Z}}/N)$ is close to zero, then the asymptotic covariance matrix of $\widehat{\boldsymbol{\eta}}$ is ill-posed, which prevents the estimator from retaining good properties.

We next establish the identifiability of the parameters. Motivated by

Qu et al. (2026), we define

$$Y_{ijt} = \frac{a_{ijt} - \mathbb{I}(X_{ijt,0} > 0)}{f(X_{ijt,0}|\boldsymbol{Z}_{ijt})}. \tag{2.7}$$

Further, denote

$$\overline{\boldsymbol{Y}} = (\overline{Y}_{01}, \ldots, \overline{Y}_{0n}, \overline{Y}_{12}, \ldots, \overline{Y}_{1n}, \ldots, \overline{Y}_{(n-1)n})^\top, \quad \overline{Y}_{ij} = \frac{1}{T}\sum_{t\in[T]} Y_{ijt}.$$

The following lemma and corollary imply the identifiability of the parameters.

Lemma 1. *Under Conditions 1–3, we have*

$$\mathbb{E}(Y_{ijt}|\boldsymbol{Z}_{ijt}) = (\theta_i^* - \theta_j^* + \boldsymbol{Z}_{ijt}^\top \boldsymbol{\eta}^*)/\gamma_0^*.$$

The proof of Lemma 1 is provided in the Supplementary Material. Without loss of generality, we set $\gamma_0^* = 1$ for addressing the scale invariance issue. Alternatively, we could also use $\gamma_0^* = -1$ as the restricted condition. Under the restrication $\gamma_0^* = 1$, we have

$$\mathbb{E}(\overline{\boldsymbol{Y}}|\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{\theta}^* + \overline{\boldsymbol{Z}}\boldsymbol{\eta}^*,$$

where $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)^\top$ is the vector of the true values for $\boldsymbol{\theta}$.

Lemma 1 suggests that if the model is identifiable, we can estimate the unknown parameters using $\overline{\boldsymbol{Y}}$, which implies that following corollary.

Corollary 1. *Under Conditions 1–4, if* $\gamma_0^* = 1$ *and* $\theta_0^* = 0$*, we have*

$$\begin{aligned} \boldsymbol{\theta}^* &= \boldsymbol{V}^{-1}\boldsymbol{U}^\top \mathbb{E}(\overline{\boldsymbol{Y}} - \overline{\boldsymbol{Z}}\boldsymbol{\eta}^*), \\ \boldsymbol{\eta}^* &= \mathbb{E}(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}\mathbb{E}(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Y}}), \end{aligned}$$

where $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_n^*)^\top$ *excludes* θ_0^*.

Corollary 1 establishes the identifiability of the parameters. Additionally, it suggests that we can estimate the unknown parameters using a least squares method. Specifically, let $\hat{f}(x|\boldsymbol{Z}_{ijt})$ be an estimator of $f(x|\boldsymbol{Z}_{ijt})$, which is provided below. Define

$$\widehat{Y}_{ijt} = \frac{a_{ijt} - \mathbb{I}(X_{ijt,0} > 0)}{\hat{f}(X_{ijt,0}|\boldsymbol{Z}_{ijt})}. \tag{2.8}$$

Further, denote

$$\breve{\boldsymbol{Y}} = (\breve{Y}_{01}, \ldots, \breve{Y}_{0n}, \breve{Y}_{12}, \ldots, \breve{Y}_{1n}, \ldots, \breve{Y}_{(n-1)n})^\top, \quad \breve{Y}_{ij} = \frac{1}{T}\sum_{t\in[T]} \widehat{Y}_{ijt}. \tag{2.9}$$

By Corollary 1, we can estimate $\boldsymbol{\eta}^*$ and $\boldsymbol{\theta}^*$, respectively, using

$$\begin{aligned}\widehat{\boldsymbol{\eta}} =&(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\breve{\boldsymbol{Y}},\\ \text{and}\quad \widehat{\boldsymbol{\theta}} =&\boldsymbol{V}^{-1}\boldsymbol{U}^\top[\boldsymbol{I}_N - \overline{\boldsymbol{Z}}(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}\overline{\boldsymbol{Z}}^\top \boldsymbol{D}]\breve{\boldsymbol{Y}} = \boldsymbol{V}^{-1}\boldsymbol{U}^\top(\breve{\boldsymbol{Y}} - \overline{\boldsymbol{Z}}\widehat{\boldsymbol{\eta}}).\end{aligned} \tag{2.10}$$

Indeed, $\widehat{\boldsymbol{\eta}}$ and $\widehat{\boldsymbol{\theta}}$ are the following least squares estimators:

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}), \tag{2.11}$$

where

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \|\breve{\boldsymbol{Y}} - (\boldsymbol{U}\boldsymbol{\theta} + \overline{\boldsymbol{Z}}\boldsymbol{\eta})\|_2^2.$$

The computational complexity of the estimators $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\eta}}$ mainly involves algebraic operations on $\boldsymbol{Z}$, $\breve{\boldsymbol{Y}}$ and $\boldsymbol{D}$. Note that $\breve{\boldsymbol{Y}}$ is given in (2.9), which depends on the computation of nonparametric density estimator $\hat{f}$. Its computational complexity is $O(n^2) \cdot O(n^2) = O(n^4)$ since it requires evaluating $O(n^2)$ terms $\hat{f}_{ijt}(x_{ijt}|\boldsymbol{Z})$ and the computational complexity of each evaluation of $\hat{f}_{ijt}(x_{ijt}|\boldsymbol{Z})$ is $O(n^2)$. Note that $\overline{\boldsymbol{Z}}$ is an $N \times p$ dense matrix with fixed p, and $\boldsymbol{D}$ is a sparse matrix, where each row and each column contain at least $O(n)$ nonzero elements. Therefore, the computational complexity of $\overline{\boldsymbol{Z}}^\top \boldsymbol{D}$ is $O(n^3)$ and the computational complexity of $\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}}$ is $O(n^3)$. The computational complexity of $\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\breve{\boldsymbol{Y}}$ for a given $\breve{\boldsymbol{Y}}$

is $O(n^3)$. The total computational complexity of $\widehat{\boldsymbol{\eta}}$ is $O(n^4)$. Similarly, the total computational complexity of $\widehat{\boldsymbol{\theta}}$ is $O(n^4)$.

We next employ the Nadaraya-Watson type estimator (Nadaraya, 1964; Watson, 1964) for the conditional density $f(x|\boldsymbol{z})$:

$$\hat{f}(x|\boldsymbol{z}) = \frac{\sum_{0\le i\neq j\le n}\sum_{t\in[T]}\mathcal{K}_{xz_1,h}(X_{ijt,0}-x, \boldsymbol{Z}_{ijt,1}-\boldsymbol{z}_1)\mathbb{I}(\boldsymbol{Z}_{ijt,2}=\boldsymbol{z}_2)}{\sum_{0\le i\neq j\le n}\sum_{t\in[T]}\mathcal{K}_{z_1,h}(\boldsymbol{Z}_{ijt,1}-\boldsymbol{z}_1)\mathbb{I}(\boldsymbol{Z}_{ijt,2}=\boldsymbol{z}_2)}, \quad (2.12)$$

where $\boldsymbol{Z}_{ijt,1}$ and $\boldsymbol{Z}_{ijt,2}$ denote the continuous covariates and the discrete covariates in $\boldsymbol{Z}_{ijt}$, respectively. Here,

$$\mathcal{K}_{z_1,h}(\boldsymbol{Z}_{ijt,1}-\boldsymbol{z}_1) = \frac{1}{h^{p_1}}\mathcal{K}_{z_1}\left(\frac{\boldsymbol{Z}_{ijt,1}-\boldsymbol{z}_1}{h}\right),$$

$$\mathcal{K}_{xz_1,h}(X_{ijt,0}-x, \boldsymbol{Z}_{ijt,1}-\boldsymbol{z}_1) = \frac{1}{h^{p_1+1}}\mathcal{K}_{xz_1}\left(\frac{X_{ijt,0}-x}{h}, \frac{\boldsymbol{Z}_{ijt,1}-\boldsymbol{z}_1}{h}\right),$$

where h denotes the bandwidth parameter, p_1 denotes the dimension of the continuous covariates $\boldsymbol{Z}_{ijt,1}$, and $\mathcal{K}_{z_1}(\cdot)$ and $\mathcal{K}_{xz_1}(\cdot)$ are two kernel functions.

3. Theoretical Results

In this section, we establish consistency and asymptotic normality for the estimators. We consider the following conditions.

Condition 5. For all $1 \le i \neq j \le n$ and $t \in [T]$, $\|\boldsymbol{Z}_{ijt}\|_\infty \le \kappa$ almost surely, where κ is allowed to diverge with n.

Condition 6. The density functions satisfy $\min\{f(x_{ijt,0}|\boldsymbol{Z}_{ijt}), f(\boldsymbol{z}_{ijt})\} > \varpi > 0$ on the support of $X_{ijt,0}$ almost surely. In addition, the rth order partial derivative of the density function $f(\boldsymbol{z}_1)$ of $\boldsymbol{Z}_{ijt,1}$ exists and is continuous and bounded. The rth order partial derivative of the joint density function $f(x, \boldsymbol{z}_1)$ of $(X_{ijt,0}, \boldsymbol{Z}_{ijt,1})$ is also continuous and bounded. Here, ϖ is allowed to decrease to zero as $n \to \infty$, and the subscript n in ϖ is suppressed.

Condition 7. The kernel function $\mathcal{K}_{\boldsymbol{z}}(\boldsymbol{z})$ is symmetric and piecewise Lipschitz continuous with order r. That is,

$$\int \cdots \int \mathcal{K}_{\boldsymbol{z}}(z_1, \ldots, z_{p_1}) dz_1 \cdots dz_{p_1} = 1,$$
$$\int \cdots \int z_1^{r_1} \cdots z_{p_1}^{r_{p_1}} \mathcal{K}_{\boldsymbol{z}}(z_1, \ldots, z_{p_1}) dz_1 \cdots dz_{p_1} = 0, \quad (0 < r_1 + \cdots + r_{p_1} < r),$$
$$\int \cdots \int z_1^{r_1} \cdots z_{p_1}^{r_{p_1}} \mathcal{K}_{\boldsymbol{z}}(z_1, \ldots, z_{p_1}) dz_1 \cdots dz_{p_1} \neq 0, \quad (0 < r_1 + \cdots + r_{p_1} = r),$$

where p_1 denotes the dimension of $\boldsymbol{Z}_{ijt,1}$.

Condition 5 assumes that the covariates are bounded above. This assumption significantly simplifies the proofs of the following theorems. However, it can be relaxed to sub-Gaussian covariates. Conditions 6 and 7 are mild and widely used in nonparametric kernel smoothing methods.

Combining Lemma 1, the definition of Y_{ijt} in (2.7) and Condition 6, we

have

$$|\mathbb{E}(Y_{ijt})| = |\theta_i^* - \theta_j^* + \mathbb{E}(\boldsymbol{Z}_{ijt}^\top)\boldsymbol{\eta}^*| = \left|\mathbb{E}\left(\frac{a_{ijt} - \mathbb{I}(X_{ijt,0} > 0)}{f(X_{ijt,0}|\boldsymbol{Z}_{ijt})}\right)\right| \leq O\left(\frac{1}{\varpi}\right).$$

This implies that $|\theta_i^* - \theta_j^*|$ is bounded above by $O(1/\varpi)$.

The following theorem establishes the consistency of the estimators for the merit parameters and the regression coefficients.

Theorem 1. *Suppose that Conditions 1–7 hold. If*

$$\frac{1}{\varpi^2}\left(1 + \frac{\kappa^2}{\lambda}\right)\left(\sqrt{\frac{\log n}{n}} + \frac{\sqrt{\log n}}{nh^{p+1}} + h^r\right) = o(1), \tag{3.1}$$

then we have

$$\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|_\infty = o_p(1), \quad \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty = o_p(1).$$

Condition (3.1) involves the selection of bandwidth h to balance the bias and variance of $\widehat{Y}_{ijt}$ when using the kernel smoothing method. Thus, the selection of bandwidth h influences the performance of the estimator both theoretically and practically. When κ/ϖ is of a constant order, it is necessary for $h \to 0$ and $nh^{p+1}/\sqrt{\log n} \to \infty$ as $n \to \infty$ to ensure the consistency of the estimators.

We now state the asymptotic distribution of the kernel-based least

squares estimator. For convenience, define

$$\overline{\boldsymbol{\tau}} = (\overline{\tau}_{01}, \ldots, \overline{\tau}_{0n}, \overline{\tau}_{12}, \ldots, \overline{\tau}_{1n}, \ldots, \overline{\tau}_{(n-1)n})^\top := \overline{\boldsymbol{Y}} - \mathbb{E}(\overline{\boldsymbol{Y}}|\boldsymbol{X}_0, \boldsymbol{Z}), \quad (3.2)$$

$$\overline{\boldsymbol{\xi}} = (\overline{\xi}_{01}, \ldots, \overline{\xi}_{0n}, \overline{\xi}_{12}, \ldots, \overline{\xi}_{1n}, \ldots, \overline{\xi}_{(n-1)n})^\top := \overline{\boldsymbol{Y}} - \mathbb{E}(\overline{\boldsymbol{Y}}|\boldsymbol{Z}). \quad (3.3)$$

Since all elements of $\overline{\boldsymbol{\tau}}$ are independent, the covariance matrix of $\overline{\boldsymbol{\tau}}$ can be written as

$$\boldsymbol{\Sigma}_\tau = \mathrm{diag}(\sigma^2_{\tau,01}, \ldots, \sigma^2_{\tau,0n}, \sigma^2_{\tau,12}, \ldots, \sigma^2_{\tau,1n}, \ldots, \sigma^2_{\tau,(n-1)n}) := \mathrm{Cov}(\overline{\boldsymbol{\tau}}), \quad (3.4)$$

where $\sigma^2_{\tau,ij} := \mathrm{Var}(\overline{\tau}_{ij})$. Similarly, we define the covariance matrix of $\overline{\boldsymbol{\xi}}$ as

$$\boldsymbol{\Sigma}_\xi = \mathrm{diag}(\sigma^2_{\xi,01}, \ldots, \sigma^2_{\xi,0n}, \sigma^2_{\xi,12}, \ldots, \sigma^2_{\xi,1n}, \ldots, \sigma^2_{\xi,(n-1)n}). \quad (3.5)$$

The asymptotic normality of $\widehat{\boldsymbol{\eta}}$ is stated below.

Theorem 2. *Suppose that Conditions 1–7 hold. If* $\sup_{i,j} \sigma^2_{\tau,ij} < \infty$ *and*

$$\frac{\kappa}{\lambda\varpi}\left(\frac{\sqrt{\log n}}{nh^{p+1}} + nh^r\right) = o(1) \text{ as } n \to \infty, \quad (3.6)$$

then $n(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)$ *is asymptotically normally distributed with mean* $\mathbf{0}$ *and covariance matrix* $n^2\mathbb{E}[(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}]\mathbb{E}(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\boldsymbol{\Sigma}_\tau \boldsymbol{D}\overline{\boldsymbol{Z}})\mathbb{E}[(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}]$, *where* $\boldsymbol{\Sigma}_\tau$ *is defined in* (3.4).

When $\sigma^2_{\tau,ij} = \sigma^2_\tau$ for all $i \neq j$, we have $n(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*)$ is asymptotically normally distributed with mean $\mathbf{0}$ and variance matrix $n^2\sigma^2_\tau \mathbb{E}[(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}]$ by noting that

$$\begin{aligned} & \mathbb{E}[(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}]\mathbb{E}(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\boldsymbol{\Sigma}_\tau \boldsymbol{D}\overline{\boldsymbol{Z}})\mathbb{E}[(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}] \\ = \quad & \sigma^2_\tau \mathbb{E}[(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}]\mathbb{E}(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\boldsymbol{D}\overline{\boldsymbol{Z}})\mathbb{E}[(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}] = \sigma^2_\tau \mathbb{E}[(\overline{\boldsymbol{Z}}^\top \boldsymbol{D}\overline{\boldsymbol{Z}})^{-1}]. \end{aligned}$$

The asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ is stated below.

Theorem 3. *Suppose that Conditions 1–7 hold. If* $\sup_{i,j} \sigma^2_{\xi,ij} := \mathrm{Var}(\overline{\xi}_{ij}) < \infty$ *and*

$$\frac{1}{\varpi^2}\left(1 + \frac{\kappa^2}{\lambda}\right)\left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log n}{nh^{2p+2}}} + \sqrt{n}h^r\right) = o(1) \text{ as } n \to \infty, \qquad (3.7)$$

then, for a constant vector $\boldsymbol{c} = (c_1, \ldots, c_n)^\top$ *satisfying* $\sum_{i=1}^n |c_i| < \infty$, *we have* $\sqrt{n}\boldsymbol{c}^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ *is asymptotically normally distributed with mean* 0 *and variance* $n\boldsymbol{c}^\top \boldsymbol{V}^{-1}\boldsymbol{U}^\top \boldsymbol{\Sigma}_\xi \boldsymbol{U}\boldsymbol{V}^{-1}\boldsymbol{c}$, *where* $\boldsymbol{\Sigma}_\xi$ *is defined in* (3.5).

When $\sigma^2_{\xi,ij} = \sigma^2_\xi$ for all $i \neq j$, we have that $\sqrt{n}\boldsymbol{c}^\top(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is asymptotically normally distributed with mean 0 and variance $n\boldsymbol{c}^\top \boldsymbol{V}^{-1}\boldsymbol{c}$ by noting that

$$\boldsymbol{c}^\top \boldsymbol{V}^{-1}\boldsymbol{U}^\top \boldsymbol{\Sigma}_\xi \boldsymbol{U}\boldsymbol{V}^{-1}\boldsymbol{c} = \boldsymbol{c}^\top \boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{c} = \boldsymbol{c}^\top \boldsymbol{V}^{-1}\boldsymbol{c}.$$

4. Numerical studies

We first consider a data-driven procedure to determine the bandwidth h. Specifically, let δ be an arbitrarily given constant. It is easy to verify that

$$\delta = \mathbb{E}\{[\mathbb{I}(X_{ijt,0} + \delta > 0) - \mathbb{I}(X_{ijt,0} > 0)]/f(X_{ijt,0}|\boldsymbol{Z}_{ijt})\}.$$

Let $\delta_m\,(1 \le m \le M)$ are pre-specified grid points on $(0, 1]$, and M is a pre-specified integer. Define

$$\hat{\delta}_m(h) = \frac{1}{NT} \sum_{0 \le i \ne j \le n} \sum_{t \in [T]} \frac{\mathbb{I}(X_{ijt,0} + \delta_m > 0) - \mathbb{I}(X_{ijt,0} > 0)}{\hat{f}(X_{ijt,0}|\boldsymbol{Z}_{ijt})}.$$

Motivated by Lewbel (1998), we can estimate h by

$$\hat{h} = \arg\min_h \sum_{m=1}^{M} [\delta_m - \hat{\delta}_m(h)]^2.$$

In the simulation studies, we set $\delta_m \in \{0.1, 0.2, \ldots, 0.9\}$.

For the kernel function, we employ the quartic kernel function

$$\mathcal{K}_z(\boldsymbol{z}) = \prod_l \frac{15}{16}(1 - z_l^2)^2 \mathbb{I}(|z_l| \le 1), \tag{4.1}$$

due to its computational efficiency and favorable properties; see Härdle

(1990). The quartic kernel function in (4.1) is symmetric and piecewise Lipschitz continuous with order two, satisfying Condition 7.

According to the definition of $\hat{h}$, its computational complexity depends on the number of candidate bandwidths k_h $(h_1, \ldots, h_{k_h})$, the number of pre-specified grid points δ_m $(1 \leq m \leq M)$ on $(0, 1]$, and the computation of the nonparametric density estimator $\hat{f}(X_{ijt,0}|\boldsymbol{Z}_{ijt})$. Because the dimension of $\boldsymbol{Z}_{ijt}$ is fixed, the computational complexity of $\hat{f}(X_{ijt,0}|\boldsymbol{Z}_{ijt})$ is $O(n^2)$. This leads to the $O(n^4)$-computational complexity of $\hat{\delta}(h)$. As a result, the total computational complexity for computing $\hat{h}$ is $O(k_h M n^4)$.

4.1 Simulation studies

In this section, we evaluate the finite sample performance of the proposed method. We generate the covariates $\boldsymbol{Z}_{ijt}$ from a two-dimensional normal distribution with mean zero and covariance $\boldsymbol{\Sigma} = (\sigma_{ij})_{2\times 2}$. Here, we take $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = \sigma_{21} = 1/4$. To construct a special regressor $\boldsymbol{X}_0$ that depends on covariates $\boldsymbol{Z}$ but does not affect their values, we set $X_{ijt,0} = \boldsymbol{Z}_{ijt}^\top \boldsymbol{b} + \omega_{ijt}$, where $\boldsymbol{b} = (0.5, -0.5)^\top$, and ω_{ijt} is independently generated from the standard normal distribution. The parameter θ_i^* is set to $\theta_i^* = 0.2i \log n/n$. We set $\boldsymbol{\eta}^* = (-0.5, 0.5)^\top$. For the noise term ε_{ijt}, we consider the following three cases:

(i) ε_{ijt} is generated from the standard normal distribution $N(0,1)$.

(ii) ε_{ijt} is generated from the logistic distribution with shift parameter 0 and scale parameter $\sqrt{3}/\pi$, i.e., $\varepsilon_{ijt} \sim \text{Logistic}(0, \sqrt{3}/\pi)$, where $\text{Var}(\varepsilon_{ijt}) = 1$. Here, $\text{Logistic}(\mu, s)$ denotes the logistic distribution with the density function

$$f(x;\mu,s) = \frac{\exp\left(-\frac{x-\mu}{s}\right)}{s\left(1+\exp\left(-\frac{x-\mu}{s}\right)\right)^2}, \tag{4.2}$$

where μ is the shift parameter (controlling the center of the distribution), and $s > 0$ is the scale parameter (controlling the dispersion of the distribution).

(iii) ε_{ijt} is generated from $N(-0.3, 0.91)$ with probability 0.75 and $N(0.9, 0.19)$ with probability 0.25, denoted as Mix-Norm.

The first case corresponds to the probit regression model, while the second case involves a logistic regression model. The last case concerns a mixture of normal distributions, designed to produce a distribution that is skewed and bimodal, while maintaining mean zero and variance one.

Each simulation is repeated 1000 times. The average biases, standard deviations (SD), and coverage frequencies (CP) of the 95% confidence intervals for the estimate $\widehat{\theta}_i$ and $\widehat{\eta}_i$ $(i = 1, 2)$ are recorded. Here,

we present the results for $\widehat{\theta}_i$ with $i = \{1, 12, 25, 37, 50\}$ for $n = 50$ and $i = \{1, 25, 50, 75, 100\}$ for $n = 100$. Table 1 reports the results for Case (i), while the results for Cases (ii) and (iii) are reported in Tables 3 and 4, respectively, in the Supplementary Material.

Table 1 demonstrates the good performance of the proposed method. Specifically, the bias of the proposed estimators is very small, and the simulated coverage frequencies are very close to the 95% target level. As expected, this table shows that the bias decreases as the number of items n increases. The standard deviation also decreases as n increases. When the number of paired comparisons T for each pair increases, the bias and the standard deviation decrease. Similar phenomena are also observed in Tables 3 and 4 in the Supplementary Material.

To assess the asymptotic normality of the parameter estimators, we provide QQ plots (quantile-quantile plots) of the kernel-based least squares estimators. The QQ plots of η_1 and θ_1 for Cases (i)–(iii) are reported in Figures 1–6 in the Supplementary Material. Most of the data points of the estimators from our proposed method fall near the reference lines, indicating that when the sample size is sufficiently large, the sampling distributions of both $\widehat{\eta}_1$ and $\widehat{\theta}_1$ approximately follow normal distributions. However, there is a slight deviation of θ_1 in the tails, which may indicate a mildly heave-

Table 1: Simulation results when $\varepsilon_{ijt} \sim N(0,1)$.

		$T=1$			$T=3$		
n		Bias	SD	CP	Bias	SD	CP
50	θ_1	0.0054	0.348	0.939	0.0042	0.210	0.955
	θ_{12}	−0.0019	0.353	0.948	0.0090	0.208	0.952
	θ_{25}	−0.0113	0.350	0.948	0.0047	0.204	0.945
	θ_{37}	−0.0066	0.361	0.951	−0.0106	0.207	0.954
	θ_{50}	−0.0178	0.367	0.964	−0.0113	0.206	0.955
	η_1	−0.0053	0.047	0.947	−0.0017	0.027	0.954
	η_2	0.0026	0.047	0.949	0.0033	0.027	0.949
100	θ_1	−0.0025	0.283	0.950	−0.0019	0.167	0.956
	θ_{25}	0.0094	0.280	0.959	0.0067	0.156	0.955
	θ_{50}	0.0025	0.283	0.958	−0.0001	0.156	0.950
	θ_{75}	−0.0022	0.282	0.956	−0.0054	0.150	0.958
	θ_{100}	−0.0081	0.276	0.967	−0.0065	0.158	0.947
	η_1	−0.0025	0.025	0.954	−0.0037	0.014	0.934
	η_2	0.0037	0.025	0.947	0.0028	0.015	0.949

tailed behavior. In comparison, η_1 fits better, indicating that its asymptotic normality is more robust.

We conducted simulation studies for comparing the parametric covariate-adjusted Bradley–Terry model in Yan (2025). The simulation results indicate the robustness of the semiparametric model. We have also carried out additional simulations to evaluate the sensitivity of the bandwidth selection and the choice of a kernel function, where the results show that the proposed estimators are not sensitive to the bandwidth selection (if its order is correctly specified) and the choice of a kernel function. Due to the limited length of pages, these simulations are relegated into the supplementary material.

4.2 Real data analysis

In this section, we analyze a real dataset using the proposed method. We consider an American National Basketball Association (NBA) dataset for the season 2018-2019. The data are freely available from the Basketball Excel website (`https://www.basketball-reference.com`), which provides detailed information about each season, such as the result of each game and the winning percentage for each team. The NBA has 30 teams. The league is divided into Eastern and Western conferences, with 15 teams in each conference. From October 2018 to April 2019, each team plays 82 games, 41 at home and 41 away. The total number of matches is 1230. We focus on the results of the games, which have a binary outcome: $a_{ijt} = 1$ when team i defeats team j in the t-th match; otherwise, $a_{ijt} = 0$.

As argued in Cattelan et al. (2013) and Tutz and Schauberger (2015), home-field advantage influences sports matches, because the home team has obvious advantages due to familiarity with the competition environment and support from the home crowd. Conversely, the away team faces the challenge of traveling and adapting to an unfamiliar environment. Let $Z_{ijt,1}$ be the covariate indicating home-field advantage information. If the home team i plays with the visiting team j in the tth match, then $Z_{ijt,1} = 1$ and $Z_{jit,1} = -1$. Additionally, Esteves et al. (2021) highlighted the impor-

tance of taking at least one day of rest to increase the chance of winning a match, demonstrating the detrimental effect of back-to-back matches on game outcomes. In the NBA, "back-to-back" specifically denotes two successive days of away games. Typically, teams experiencing highly congested schedules may be particularly exposed to air travel effects, circadian rhythm disruption, sleep deprivation, decline in physical capacity, and injury risk (Reilly et al., 2001). Consequently, we regard back-to-back matches as a covariate in our analysis to accurately determine if the away team is at a competitive disadvantage due to consecutive matches. Let $Z_{ijt,2}$ be the covariate indicating whether the away team plays back-to-back games. If team i is the away team in a back-to-back game and team j is the home team, then $Z_{ijt,2} = 1$ and $Z_{jit,2} = -1$; otherwise, $Z_{ijt,2} = Z_{jit,2} = 0$.

We now define a special regressor under Condition 1. We employ the win percentage as the continuous covariate, and ensure its timeliness through dynamic adjustment. In the first month of the season, due to the lack of actual data for that season, we calculate the predicted winning percentage using the win percentagwe calculate the predicted winning percentage using the predicted wins released on October 16, 2018 by ESPN's well-known analyst Kevin Pelton, whose methodology incorporates changes in team rosters, player health, and outlook for the season. Starting from

the second month, we utilize a rolling monthly net win percentages metric, calculated as the number of wins of the team divided by the total number of games the team played in the previous month. This hybrid method balances between pre-season analytics and in-season performance. Specifically, let $X_{ijt,0} = W_{it,j} - W_{jt,i}$, where $W_{it,j}$ denotes the winning percentage of the team i at the tth match with j. We first partition the support of $X_{ijt,0}$ into $K = 5$ equal subintervals, denoted by $\mathcal{I}_k = [x_k, x_{k+1})$ $(1 \le k \le K)$, respectively. We then calculate the wining rate when the value of $X_{ijt,0}$ falls within $\mathcal{I}_k$:

$$\mathrm{WR}_k = \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j\neq i} a_{ijt,0}\mathbb{I}(X_{ijt,0} \in \mathcal{I}_k)}{\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j\neq i} \mathbb{I}(X_{ijt,0} \in \mathcal{I}_k)}. \tag{4.3}$$

The values of WR_k are $0.864, 0.764, 0.613, 0.397$, and 0.130 for $k = 1, \ldots, 5$, respectively, exhibiting a decreasing trend as k increases. Therefore, we set $\gamma_0 = -1$ for identifiablity.

We obtain the estimators according to (2.10). For regression coefficients, $\widehat{\boldsymbol{\eta}} = (0.065, -0.098)^\top$, which indicates that home-field advantage positively affects the competition result with an intensity of 0.065, whereas the back-to-back match negatively influences the competition result with an intensity of 0.098. The p-value and a 95% confidence interval (CI) for

the estimated regression coefficient of the home-field advantage are 0.027 and $(0.008, 0.122)$, respectively. The p-value for the estimated regression coefficient of the back-to-back is 0.234. This suggests that the home-field advantage has a significant positive effect on competition outcomes, while the back-to-back has no significant effect on competition outcomes. We normalize the merit parameter of New York Knicks to 0 for identifiability. New York Knicks has the lowest merit parameter estimator in the Eastern Conference, whereas Milwaukee Bucks has the highest with an estimator of 1.981. The team has the highest estimated merit parameter in the Western Conference is Golden State Warriors, whereas the lowest is Phoenix Suns, with estimators of 1.874 and 0.043, respectively. Table 2 shows the estimators of merit parameters, as well as a_i (the number of wins for team i) and rank estimators $\widehat{R}_i$. Due to the restriction of the page width, the 95% confidence intervals of merit parameters are put in Table 13 in the Supplementary Material.

5. Summary

We have proposed a semiparametric paired comparison model that incorporates covariates. By introducing a special regressor, we developed a kernel-based least squares method to estimate all unknown parameters in the

Table 2: The estimates of θ_i in 2018-19 NBA regular season.

Eastern Conference				Western Conference			
Team	a_i	$\widehat{\theta}_i$	$\widehat{R}_i$	Team	a_i	$\widehat{\theta}_i$	$\widehat{R}_i$
Milwaukee Bucks	60	1.981	1	Golden State Warriors	57	1.874	3
Toronto Raptors	58	1.978	2	Denver Nuggets	54	1.669	5
Philadelphia 76ers	51	1.646	6	Houston Rockets	53	1.680	4
Boston Celtics	49	1.608	7	Portland Trail Blazers	53	1.365	11
Indiana Pacers	48	1.579	8	Utah Jazz	50	1.394	10
Orlando Magic	42	0.879	21	Oklahoma City Thunder	49	1.491	9
Brooklyn Nets	42	1.044	16	San Antonio Spurs	48	1.323	12
Detroit Pistons	41	1.241	14	LA Clippers	48	1.243	13
Miami Heat	39	1.062	15	Sacramento Kings	39	0.880	20
Charlotte Hornets	39	0.905	19	Los Angeles Lakers	37	0.970	17
Washington Wizards	32	0.795	23	Minnesota Timberwolves	36	0.950	18
Atlanta Hawks	29	0.448	26	New Orleans Pelicans	33	0.824	22
Chicago Bulls	22	0.200	27	Dallas Mavericks	33	0.696	25
Cleveland Cavaliers	19	0.111	28	Memphis Grizzlies	33	0.713	24
New York Knicks	17	0	30	Phoenix Suns	19	0.043	29

model. When there are several continuous covariates that could be potentially used as the special regressor, except for selecting the one that with the largest observed support, we could loop over each special regressor to estimate the model parameters and then take the average. This model averaging approach can be used to estimate the merit parameters of items, but cannot be used to estimate the regression coefficients of the covariates because different special regressors lead to different regression coefficients of covariates. Another way is to use the cross-validation method based on some criterion (e.g., prediction error), but this method is very time-consuming in semiparametric paired comparison setting.

In Condition 1, we set the coefficient of the special regressor to be 1 for convenience. However, we could also set it to be -1, and all theoretical results still hold. To determine the sign of coefficient for the special regressor, we can partition the support of covariate $X_{ijt,0}$ into K equal subintervals, as described in Section 4.2. If WR_k in (4.3) exhibits an increasing trend as k increases, then the effect of the covariate $X_{ijt,0}$ is likely positive, and we set $\gamma_0^* = 1$. Conversely, if WR_k shows a decreasing trend as k increases, then the effect of the covariate $X_{ijt,0}$ is likely negative, and we set $\gamma_0^* = -1$. However, when there are no such increasing or decreasing trends among all covariates, identifying a covariate as the special regressor becomes challenging. We do not explore this case here and intend to investigate it in future research.

For the identification of the merit parameters, the condition $\mathbf{1}^\top \boldsymbol{\theta}^* = 0$ could also be used. However, our technique for theoretical analysis uses the inverse of $\boldsymbol{V} = \boldsymbol{U}^\top \boldsymbol{U}$ in (2.4). If θ_0^* is included, then $\boldsymbol{V}$ is not invertible. Therefore, we use $\theta_0^* = 0$ as the identification condition as in Simons and Yao (1999). We established the consistency and asymptotic normality of the resulting estimators under some conditions. These conditions may not be best possible. It is of interest to see whether these conditions can be relaxed.

Finally, we discuss extension of the methodology and theory to sparse comparison graphs such as the Erdös–Rényi graph or heterogenous Erdös–Rényi graph (Han et al., 2024; Han and Xu, 2025). Our proposed estimation method does not depend on the structures of comparison graphs and can be directly applied to any comparison graph. However, our techniques for theoretical analyses reply on properties of the complete comparison graph, where the inverse of the information matrix $\boldsymbol{V} = \boldsymbol{U}^\top \boldsymbol{U}$ admits a closed form, with $\boldsymbol{U}$ denoting the design matrix of the merit parameters. In this case, $\boldsymbol{V}$ is a diagonally dominate matrix with positive diagonal elements and negative off-diagonal elements. In sparse comparison graphs, many entries of $\boldsymbol{V}$ are zero, which implies that $\boldsymbol{V}^{-1}$ does not have a closed form. One approach is to approximate $\boldsymbol{V}^{-1}$ using a simple matrix (Simons and Yao, 1999). Another approach is to apply the pseudoinverse technique developed by Han and Xu (2025). This requires a highly non-trivial analysis. The simulations results in Section S8 of the Supplementary Material indicate that the asymptotic properties of the estimators still hold in sparse paired comparison settings. We would like to leave this issue for future research.

Acknowledgements

We are very grateful to three referees, the associated editor, and the editor for their valuable comments that have greatly improved the manuscript. Yan is supported by the National Natural Science Foundation of China (No. 12171188, 12322114).

References

Agresti, A. (2012). *Categorical Data Analysis, 3rd Edition.* Wiley, New York.

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

Candelaria, L. E. (2020). A semiparametric network formation model with unobserved linear heterogeneity. *arXiv preprint arXiv:2007.05403*.

Cattelan, M., Varin, C., and Firth, D. (2013). Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 62(1):135–150.

Chen, B., Escalera, S., Guyon, I., Ponce-López, V., Shah, N., and Oliu Simón, M. (2016). Overcoming calibration problems in pattern labeling with pairwise ratings: application to personality traits. In *European Conference on Computer Vision*, pages 419–432.

Chen, P., Gao, C., and Zhang, A. Y. (2022). Optimal full ranking from pairwise comparisons. *The Annals of Statistics*, 50(3):1775–1805.

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

David, H. A. (1988). *The Method of Paired Comparisons. 2nd Edition.* Oxford University Press, Oxford.

Dong, P., Han, R., Jiang, B., and Xu, Y. (2024). Statistical ranking with dynamic covariates. *arXiv preprint arXiv:2406.16507*.

Dzemski, A. (2019). An empirical model of dyadic link formation in a network with unobserved heterogeneity. *Review of Economics and Statistics*, 101(5):763–776.

Esteves, P. T., Mikolajec, K., Schelling, X., and Sampaio, J. (2021). Basketball performance is affected by the schedule congestion: NBA back-to-backs under the microscope. *European Journal of Sport Science*, 21(1):26–35.

Fan, J., Hou, J., and Yu, M. (2024). Uncertainty quantification of MLE for entity ranking with covariates. *Journal of Machine Learning Research*, 25(358):1–83.

Fan, J., Lou, Z., Wang, W., and Yu, M. (2025). Ranking inferences based on the top choice of multiway comparisons. *Journal of the American Statistical Association*, 120(549):237–250.

Han, R., Tang, W., and Xu, Y. (2024). Statistical inference for pairwise comparison models. *arXiv preprint arXiv:2401.08463*.

Han, R. and Xu, Y. (2025). A unified analysis of likelihood-based estimators in the plackett–luce model. *The Annals of Statistics*, 53(5):2077–2102.

Han, R., Xu, Y., and Chen, K. (2023). A general pairwise comparison model for extremely sparse networks. *Journal of the American Statistical Association*, 118(544):2422–2432.

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

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Hunter, D. R. (2003). MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, 32(1):384–406.

Lewbel, A. (1998). Semiparametric latent variable model estimation with endogenous or mismeasured regressors. *Econometrica*, 66(1):105–121.

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of*

econometrics, 97(1):145–177.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of econometrics*, 27(3):313–333.

Masarotto, G. and Varin, C. (2012). The ranking lasso and its application to sport tournaments. *The Annals of Applied Statistics*, 6(4):1949–1970.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, 9(1):157–159.

Qu, L., Chen, L., Yan, T., and Chen, Y. (2026). Inference in semiparametric formation models for directed networks. *Journal of Business and Economic Statistics*, 44(1):188–202.

Radlinski, F. and Joachims, T. (2007). Active exploration for learning rankings from clickthrough data. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 570–579, New York. Association for Computing Machinery.

Reilly, T., Atkinson, G., and Budgett, R. (2001). Effect of low-dose temazepam on physiological variables and performance tests following a westerly flight across five time zones. *International journal of sports medicine*, 22(03):166–174.

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

Singh, R., Iliopoulos, G., and Davidov, O. (2025). Least squares for cardinal paired comparisons data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page To appear.

Stigler, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science*, 9(1):94–108.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Tutz, G. and Schauberger, G. (2015). Extended ordered paired comparison models with application to football data from German Bundesliga.

AStA Advances in Statistical Analysis, 99:209–227.

Varin, C., Cattelan, M., and Firth, D. (2016). Statistical modelling of citation exchange between statistics journals. *Journal of The Royal Statistical Society Series A-statistics in Society*, 179(1):1–63.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372.

Yan, T. (2025). Inference in a generalized Bradley-Terry model for paired comparisons with covariates and a growing number of subjects. *arXiv:2507.22472*.

Yan, T., Jiang, B., Fienberg, S. E., and Leng, C. (2019). Statistical inference in a directed network model with covariates. *Journal of the American Statistical Association*, 114(526):857–868.

Yan, T., Li, Y., Xu, J., Yang, Y., and Zhu, J. (2025). Likelihood ratio tests in random graph models with increasing dimensions. *Journal of the American Statistical Association*, (just-accepted):1–26.

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

Zeleneev, A. (2020). Identification and estimation of network models with nonparametric unobserved heterogeneity. *Working paper*.

Department of Statistics, Central China Normal University, Wuhan 430079, China.

E-mail: hysong05@mails.ccnu.edu.cn

Department of Statistics, Central China Normal University, Wuhan 430079, China.

E-mail: qulianq@ccnu.edu.cn

Department of Statistics, Central China Normal University, Wuhan 430079, China.

E-mail: tingyanty@mail.ccnu.edu.cn

Department of Statistics, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA.

E-mail: yuguo@illinois.edu