INTERACTION SCREENING VIA KENDALL'S RANK CORRELATION FOR IMBALANCED MULTI-CLASS CLASSIFICATION

Tanaka, Shuntaro
The Japan Research Institute, Limited.

Matsui, Hidetoshi
Faculty of Data Science, Shiga University

https://doi.org/10.5109/7234383

出版情報：Bulletin of informatics and cybernetics. 56 (5), pp.1-24, 2024. 統計科学研究会
バージョン：
権利関係：

INTERACTION SCREENING VIA KENDALL'S RANK CORRELATION FOR IMBALANCED MULTI-CLASS CLASSIFICATION

by

Shuntaro Tanaka and **Hidetoshi Matsui**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.56, No. 5

FUKUOKA, JAPAN
2024

Bulletin of Informatics and Cybernetics, Vol. 56, No. 5, 2024

INTERACTION SCREENING VIA KENDALL'S RANK CORRELATION FOR IMBALANCED MULTI-CLASS CLASSIFICATION

By

Shuntaro Tanaka[*†] and **Hidetoshi Matsui**[‡]

Abstract

Screening is a useful method for selecting important variables for high-dimensional data where the number of predictors is much larger than the sample size. Screening methods can eliminate unnecessary variables at a low computational cost by calculating their importance scores, such as the correlation between response and predictor variables. In this study, we consider the problem of selecting variables and interactions in classification problems for data with imbalanced sample sizes between classes. Specifically, we propose a new method called Class-to-Class KIF (CCKIF) to select interactions in imbalanced multi-class classification problems. CCKIF uses the difference in Kendall's rank correlations for each class to calculate the importance scores of the interactions to improve the selection accuracy more than existing methods, even for imbalanced data. We present the theoretical properties of the proposed method. Simulation studies and real data analysis show that the proposed CCKIF appropriately selects important interactions, particularly for data on minor classes.

Key Words and Phrases: Screening, Variable selection, Interaction, Multi-class classification.

1. Introduction

Variable selection is one of the most important issues in regression analysis for modeling the relationships between the response and predictor variables. Variable selection has been applied in various fields including economics, engineering, and biology (Montgomery et al., 2021). In finance, models are constructed to predict company bankruptcy using important variables selected from data on transactions and payment networks (Kou et al., 2021). In computer science, variables closely related to faults are selected and monitored to detect faults that affect product quality (Wu et al., 2020). In environmental science, Valentini et al. (2021) constructed a model for monitoring water quality by setting the water quality index as the response and the number of specific components in the water as predictors.

Sure Independence Screening (SIS) (Fan and Lv, 2008) was proposed as a variable selection method for high-dimensional data where the number of predictors is much

[*] The Japan Research Institute, Ltd. 2–18–1 Higashi–gotanda Shinagawa–ku Tokyo Japan. tanaka.shuntaro@jri.co.jp

[†] Graduate School of Data Science, Shiga University 1–1–1 Banba Hikone Shiga 522–8522 Japan. s7022103@st.shiga-u.ac.jp

[‡] Faculty of Data Science, Shiga University 1–1–1 Banba Hikone Shiga 522–8522 Japan. hmatsui@biwako.shiga-u.ac.jp

larger than the sample size. The SIS assumes a linear regression model for the data and defines the importance of the predictors related to the response using the Pearson correlation between the response and predictors. The SIS has a good property called the sure screening property, which states that the probability that a set of variables selected by the SIS contains a set of important variables converges to one as the sample size increases.

Several extensions of the SIS have been proposed. Fan and Song (2010) extends SIS to generalized linear models, and Fan et al. (2011) proposed Nonparametric Independence Screening (NIS) that selects variables in high-dimensional additive models. Instead of the Pearson correlation used in SIS, Robust Rank Correlation Screening (RRCS) (Li et al., 2012a) uses Kendall's rank correlation coefficient, and Distance Correlation SIS (DC-SIS) (Li et al., 2012b) uses the distance correlation to capture the nonlinear relationship between the response and predictors. The Pearson's Chi-square SIS (PC-SIS) (Huang et al., 2014) and Mean-Variance SIS (MV-SIS) (Cui et al., 2015) enable us to handle categorical and continuous variables.

Many screening methods consider only the main effects of the predictors. However, several methods consider the interactions of predictors to improve prediction performance and model interpretability. There are two methods for selecting interactions: one assumes that only interactions related to variables with main effects are selected, and the other does not. Methods for the former include Interaction FORward selecting procedure under the Marginality principle (iFORM) (Hao and Zhang, 2014) and Stepwise cOnditional likelihood variable selection for Discriminant Analysis (SODA) (Li and Liu, 2019). They selected only interactions related to the selected main effects, which reduced the computational cost. In contrast, methods of the latter type include IP-SIS (Fan et al., 2016) and BCor-SIS (Pan et al., 2019a). To include the effect of interactions on the importance of the predictors, IP-SIS uses the correlation between the squared transformation of the response and the predictors, and BCor-SIS uses the ball correlation (Pan et al., 2019b) between the power transformation of the response and the predictors. This study focuses on methods that do not make any assumptions about the existence of main effects. This is because in some cases, only interactions are meaningful in the real world. For example, in finance, when there are two variables, the interest rate and the economic growth rate, the value of the interest rate has a different meaning depending on whether the economic growth rate is higher or lower, even if it has the same value. In such cases, the main effect may not be significant.

In this study, we consider the problem of interaction selection in classification. Classification problems have been addressed to solve issues in many fields, such as email spam classification and item categorization. Interactions are important for predicting the performance of statistical models in classification problems. Joint Cumulant Interaction Screening (JCIS) (Reese, 2018) and the Kendall Interaction Filter (KIF) (Anzarmou et al., 2023) are proposed for selecting interactions in classification problems. JCIS calculates the importance scores of interactions using a three-way joint cumulant that includes the response and two predictors. The KIF calculates the scores by taking the difference between the Kendall's rank correlation computed for the entire sample and that computed for the data for each class. JCIS supports binary classification only, whereas the KIF supports multiclass classification.

This study considers a multi-class classification problem in which the sample size of a minor class is highly imbalanced. For example, in scenarios where a financial institution

detects illegal transactions or a medical institution diagnoses rare diseases, the sample sizes for classes corresponding to illegal transactions or rare diseases are typically small. The KIF does not perform well with imbalanced data, particularly when important interactions are related to the classification of minor classes. To address this issue, we propose a modified KIF method that improves the calculation of the importance scores of interactions even for imbalanced data. This method is called the Class-to-Class KIF (CCKIF). Using data from each class, CCKIF can adequately reflect the information provided by the data from the minor classes in the importance scores. We also show that the proposed CCKIF satisfies the theoretical properties under several conditions, such as the sure screening and ranking consistency properties. Simulations and real data analysis show that CCKIF selects important interactions even for imbalanced data.

The remainder of this paper is organized as follows: Section 2 provides an overview of existing screening methods. Section 3 introduces the details of the proposed method and presents its theoretical properties. In Section 4, we report the results of the analyses based on the simulated and real data. Finally, Section 5 provides a summary and discussion.

2. KIF methods

Let $X_1, \ldots, X_p$ be p random variables as predictors and $Y \in \{1, \ldots, K\}$ be a variable representing the class label as a response. We assume that each variable has a finite variance. We consider a multiclass classification problem for Y using $X_1, \ldots, X_p$. In particular, we consider the problem of selecting the interactions that contribute to classifying Y from $p(p-1)/2$ interactions constructed using $X_1, \ldots, X_p$. Let $\tilde{X}_1, \ldots, \tilde{X}_p$ and $\tilde{Y}$ be independent copies of $X_1, \ldots, X_p$ and Y, respectively. The global Kendall rank correlation τ of the interaction between the j-th and l-th predictors is given by:

$$\begin{aligned}\tau(X_j, X_l) &= \mathbb{P}\left((X_j - \tilde{X}_j)(X_l - \tilde{X}_l) > 0\right) - \mathbb{P}\left((X_j - \tilde{X}_j)(X_l - \tilde{X}_l) < 0\right) \\ &= 2\mathbb{P}\left((X_j - \tilde{X}_j)(X_l - \tilde{X}_l) > 0\right) - 1,\end{aligned}$$

and the in-class Kendall rank correlation τ_k for class k is given by:

$$\tau_k(X_j, X_l) = 2\mathbb{P}\left((X_j - \tilde{X}_j)(X_l - \tilde{X}_l) > 0 \mid Y = k, \tilde{Y} = k\right) - 1.$$

Then, the KIF score is defined as

$$w^*_{j,l} = \sum_{k=1}^{K} \pi_k \left|\tau_k(X_j, X_l) - \tau(X_j, X_l)\right|,$$

where $\pi_k = \mathbb{P}(Y = k)$. When Y is independent of the interaction (X_j, X_l), $\tau_k(X_j, X_l)$ takes equal value for all k; thus, $\tau_k(X_j, X_l) = \tau(X_j, X_l)$. In this case, $w^*_{j,l}$ equals zero. Conversely, when Y depends on the interaction (X_j, X_l), meaning $\tau_k(X_j, X_l)$ takes different values for each class, $w^*_{j,l}$ increases because $|\tau_k(X_j, X_l) - \tau(X_j, X_l)|$ increases. We consider that interactions may contribute to the classification when they relate to Y. Therefore, we can interpret the KIF score as indicating the importance of the interactions that contribute to the classification of responses. The KIF score has the advantage of being insensitive to the monotonic transformations of predictors because it is based on rank.

Because KIF is a filtering method that aims to select important interactions at a low computational cost, it cannot select variables by considering the main effects. Using the interactions selected by the KIF and the variables selected by other methods that consider the main effects enables us to construct statistical models that consider both the main effects and interactions.

Let us suppose we have n sets of observations $\{(y_i, \boldsymbol{x}_i);\ 1 \leq i \leq n\}$, where $y_i \in \{1, \ldots, K\}$ is a response and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ is a vector of predictors. The empirical version of the global Kendall rank correlation is given by:

$$\hat{\tau}(X_j, X_l) = \frac{4}{n(n-1)} \sum_{i=1}^{n-1} \sum_{s=i+1}^{n} \mathbb{1}\left\{(x_{ij} - x_{sj})(x_{il} - x_{sl}) > 0\right\} - 1,$$

where $\mathbb{1}\{\cdot\}$ is an indicator function and the empirical version of the in-class Kendall rank correlation is given by:

$$\hat{\tau}_k(X_j, X_l) = \frac{4}{n_k(n_k-1)} \sum_{i=1}^{n-1} \sum_{s=i+1}^{n} \mathbb{1}\left\{(x_{ij} - x_{sj})(x_{il} - x_{sl}) > 0, y_i = k, y_s = k\right\} - 1,$$

where n_k denotes the number of observations belonging to class k. The empirical version of the KIF score is then given by:

$$\hat{w}^*_{j,l} = \sum_{k=1}^{K} \hat{\pi}_k \left|\hat{\tau}_k(X_j, X_l) - \hat{\tau}(X_j, X_l)\right|, \tag{1}$$

where $\hat{\pi}_k = n_k/n$. We select the interactions of (X_j, X_l) that satisfy $\hat{w}^*_{j,l} > h$ for threshold $h > 0$.

A drawback of the KIF is that the importance of interactions related to minor classes is sometimes estimated to be smaller than their true values for imbalanced data. A possible reason for this is that the KIF score (1) assigns smaller weights to minor classes both implicitly by using $\hat{\tau}(X_j, X_l)$ calculated using all in the sample and explicitly by assigning weights $\hat{\pi}_k$.

3. Proposed method

3.1. CCKIF score

To resolve the KIF's drawback, we propose a method called CCKIF, which can accurately compute the importance scores of interactions, even for imbalanced data. The CCKIF score is calculated as follows:

$$w_{j,l} = \frac{1}{K^2} \sum_{k=1}^{K} \sum_{m=1}^{K} \pi_{k,m} \left|\tau_k(X_j, X_l) - \tau_m(X_j, X_l)\right|,$$

where $\pi_{k,m}$ is the average of π_k and π_m that satisfies the order of $\pi_{k,m}$, π_k, and π_m with respect to n are the same. Examples of $\pi_{k,m}$ are as follows.

- Arithmetic: $\dfrac{\pi_k + \pi_m}{2}$,

- Geometric: $\sqrt{\pi_k \pi_m}$,
- Harmonic: $\dfrac{2\pi_k \pi_m}{\pi_k + \pi_m}$.

CCKIF score captures the differences in the behavior of interactions across classes more accurately than the KIF score, by using the in-class Kendall rank correlation instead of the global Kendall rank correlation calculated using all in the sample. Furthermore, CCKIF uses $\pi_{k,m}$ to avoid underestimating the score of interactions related to minor classes.

The empirical version of the CCKIF score is given by:

$$\hat{w}_{j,l} = \frac{1}{K^2} \sum_{k=1}^{K} \sum_{m=1}^{K} \hat{\pi}_{k,m} \left| \hat{\tau}_k(X_j, X_l) - \hat{\tau}_m(X_j, X_l) \right|,$$

where $\hat{\pi}_{k,m}$ is the average of $\hat{\pi}_k$ and $\hat{\pi}_m$.

3.2. Theoretical properties of CCKIF

Let $\boldsymbol{S} = \{(j,l) : \text{ the interaction } (X_j, X_l) \text{ is important for } Y\}$ be an active set of important interactions. We consider the interaction (X_j, X_l) important if the empirical CCKIF score satisfies $\hat{w}_{j,l} > cn^{-r}$, where c and r are positive constants. Furthermore, let $\hat{\boldsymbol{S}} = \{(j,l) : \hat{w}_{j,l} > cn^{-r}\}$ be the set of tuples of the indices of predictors whose interactions are estimated to be important. The number of important interactions is $|\boldsymbol{S}| = s$, where s is significantly lower than p.

First, we demonstrate the sure screening property that $\mathbb{P}(\boldsymbol{S} \subseteq \hat{\boldsymbol{S}})$ converges asymptotically to one as the sample size increases. To achieve this, we assume the following conditions:

(C1) There exist two positive constants c_1 and c_2 such that

$$\frac{c_1}{K} \leq \min_{1 \leq k \leq K} \pi_k \leq \max_{1 \leq k \leq K} \pi_k \leq \frac{c_2}{K}.$$

(C2) There exist positive constants $c > 0$ and $0 \leq r < 1/2$ such that $\min_{(j,l) \in \boldsymbol{S}} w_{j,l} > 2cn^{-r}$.

(C3) The number K of classes satisfies $K = O(n^d)$ for $0 \leq d < 1/2 - r$.

(C4) $\log(p) = o(n^{1-2r})$.

(C5) The orders of $\pi_{k,m}$, π_k, and π_m with respect to n are the same.

Under these conditions, Theorem 1 holds. The proof of Theorem 1 is provided in Appendix.

Theorem 1 (Sure screening property) *Under conditions (C1)–(C5), there exists a positive constant b that depends on c, c_1, and c_2 such that*

$$i. \quad \mathbb{P}\Big(\max_{1\le j,l\le p} |\hat{w}_{j,l} - w_{j,l}| \le cn^{-r} \Big) \ge 1 - O\left(p^2 n^{2d} \exp\left(\frac{-c^2 n^{1-2r}}{72b^2} \right) \right),$$

$$ii. \quad \mathbb{P}(\boldsymbol{S} \subseteq \hat{\boldsymbol{S}}) \ge 1 - O\left(sn^{2d} \exp\left(\frac{-c^2 n^{1-2r}}{72b^2} \right) \right).$$

Since $\hat{\boldsymbol{S}}$ uses cn^{-r} as a threshold for selecting important interactions, it may appear that as the sample size increases, cn^{-r} approaches zero, which causes $\hat{\boldsymbol{S}}$ to include all interactions eventually. However, according to Theorem 1 (i), the difference between $\hat{w}_{j,l}$ and $w_{j,l}$ also approaches zero. Therefore, $(j,l) \in \boldsymbol{S}^c$ that satisfy $w_{j,l} \le cn^{-r}$ are unlikely to be included in $\hat{\boldsymbol{S}}$. Theorem 1 (ii) ensures that the set of interactions that CCKIF estimates as important contains important interactions with a probability close to one. If there exists $(j,l) \in \boldsymbol{S}^c$ that satisfies $w_{j,l} > 2cn^{-r}$, then $\hat{\boldsymbol{S}}$ will include some unimportant interactions, and therefore it is unlikely to be $\boldsymbol{S} = \hat{\boldsymbol{S}}$. Although the KIF has the sure screening property, Theorem 1 claims that the probabilities of CCKIF converge to one faster than those of the KIF as the sample size approaches infinity because the order of n is smaller than that of the KIF (Anzarmou et al., 2023).

In addition, we introduce the ranking consistency property, which states that interactions relevant to a response have higher CCKIF scores than irrelevant ones. To achieve this, we consider an additional condition.

(C6) $\liminf_{n\to\infty} \left\{ \min_{(j,l)\in\boldsymbol{S}} w_{j,l} - \max_{(j,l)\in\boldsymbol{S}^c} w_{j,l} \right\} \ge c_3$, where $c_3 > 0$ is a constant.

Then the following result holds. The proof of Theorem 2 is provided in Appendix.

Theorem 2 (Ranking consistency property) *Suppose that (C1)–(C6); then,*

$$\liminf_{n\to\infty} \left\{ \min_{(j,l)\in\boldsymbol{S}} \hat{w}_{j,l} - \max_{(j,l)\in\boldsymbol{S}^c} \hat{w}_{j,l} \right\} > 0 \ \ a.s.$$

From Theorem 2, we can identify $\boldsymbol{S}$ by selecting interactions in the order of CCKIF scores if the number of important interactions is known and the sample size is sufficiently large.

4. Numerical Results

We analyzed two simulated datasets and one real dataset. In these experiments, we compared the variable selection performance of the proposed method with that of the existing methods. Each method calculates the importance scores of the interactions and then selects the interactions based on these scores. A method that selects more interactions relevant to a response is preferable.

4.1. Simulated dataset 1

Let Y be a response that takes one of the values 1, 2, or 3, and $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ be a vector of predictors that follow three multivariate normal distributions, as follows:

$$\boldsymbol{X} \mid Y = 1 \sim N(\mathbf{0}_p, \Sigma_1) \text{ where } \Sigma_{1,(j_1,j_2)} = \begin{cases} 1 & \text{if } j_1 = j_2, \\ 0.8 & \text{if } (j_1, j_2) \in \{(1,2),(2,1)\}, \\ 0.2 & \text{otherwise,} \end{cases}$$

$$\boldsymbol{X} \mid Y = 2 \sim N(\mathbf{0}_p, \Sigma_2) \text{ where } \Sigma_{2,(j_1,j_2)} = \begin{cases} 1 & \text{if } j_1 = j_2, \\ 0.8 & \text{if } (j_1, j_2) \in \{(3,4),(4,3)\}, \\ 0.2 & \text{otherwise,} \end{cases}$$

$$\boldsymbol{X} \mid Y = 3 \sim N(\mathbf{0}_p, \Sigma_3) \text{ where } \Sigma_{3,(j_1,j_2)} = \begin{cases} 1 & \text{if } j_1 = j_2, \\ 0.8 & \text{if } (j_1, j_2) \in \{(5,6),(6,5)\}, \\ 0.2 & \text{otherwise.} \end{cases}$$

This setting indicates that three interactions, X_1X_2, X_3X_4, and X_5X_6, are important for identifying classes 1, 2, and 3. Datasets were generated with several sample sizes n, number of variables p, and sample ratios for Y. The sample ratios were tested using three different imbalance patterns: (0.6, 0.3, 0.1), (0.5, 0.3, 0.2), and (0.4, 0.3, 0.3). We selected $[n/\log(n)]$ interactions in the order of importance score values, as in the KIF, where $[a]$ is the maximum integer that does not exceed a. Note that many screening methods commonly select $[n/\log(n)]$ as the number of variables; therefore, we also use this value.

We compared the proposed CCKIF with the KIF, JCIS, BCor-SIS, IP-SIS, and SODA for each dataset. For $\pi_{k,m}$ in CCKIF, we used the arithmetic average $\pi_{k,m} = (\pi_k + \pi_m)/2$. For methods other than CCKIF and KIF, which do not support multiclass classification, we convert the value of Y to binary values Y_1, Y_2, and Y_3 as follows:

$$Y_1 = \begin{cases} 1 & \text{if } Y = 1 \\ 0 & \text{if } Y \neq 1 \end{cases}, \quad Y_2 = \begin{cases} 1 & \text{if } Y = 2 \\ 0 & \text{if } Y \neq 2 \end{cases}, \quad Y_3 = \begin{cases} 1 & \text{if } Y = 3 \\ 0 & \text{if } Y \neq 3 \end{cases}.$$

We calculated the importance scores of the interactions in each of the three datasets $(Y_1, \boldsymbol{X})$, $(Y_2, \boldsymbol{X})$, and $(Y_3, \boldsymbol{X})$ and then used the average of these three scores as the final importance score. Because BCor-SIS and IP-SIS do not directly compute the importance scores of the interactions, we compute them for each predictor and then select all possible combinations of q predictors, where q is chosen such that the number $q(q-1)/2$ of interactions is close to $[n/\log(n)]$. We repeated the above analysis 50 times and summarized the screening results.

Table 1 lists the results for several values of p with fixed $n = 300$. The values in the columns for each method represent the number of times the three important interactions were selected. BCor-SIS, IP-SIS, and SODA rarely selected important interactions. The proposed CCKIF selected the interaction term X_5X_6 related to the minor class more than existing methods. In particular, when the class labels were highly imbalanced with a sample ratio of (0.6, 0.3, 0.1), methods other than CCKIF rarely selected X_5X_6. When the class labels are not imbalanced, both the KIF and CCKIF select X_5X_6 appropriately. In most cases, CCKIF selected the three interactions the most.

Table 1: Results of the interaction selection for simulated dataset 1 with several p

n	$[n/\log(n)]$	p	number of all interactions	sample ratio	important interactions	CCKIF	KIF	JCIS	BCor-SIS	IP-SIS	SODA
300	52	200	19900	0.6, 0.3, 0.1	X_1X_2	48	49	43	0	0	0
					X_3X_4	48	50	27	0	0	1
					X_5X_6	33	2	2	0	0	0
				0.5, 0.3, 0.2	X_1X_2	50	50	48	0	0	0
					X_3X_4	49	50	31	2	1	2
					X_5X_6	49	46	7	0	0	0
				0.4, 0.3, 0.3	X_1X_2	50	50	46	0	0	1
					X_3X_4	49	49	19	1	0	1
					X_5X_6	50	49	20	0	0	3
300	52	300	44850	0.6, 0.3, 0.1	X_1X_2	50	50	42	0	1	1
					X_3X_4	43	49	25	0	1	2
					X_5X_6	24	2	1	0	1	0
				0.5, 0.3, 0.2	X_1X_2	50	50	47	1	0	1
					X_3X_4	48	50	14	0	0	1
					X_5X_6	48	36	5	0	0	1
				0.4, 0.3, 0.3	X_1X_2	50	50	42	0	1	0
					X_3X_4	50	50	10	0	0	1
					X_5X_6	49	48	15	0	0	1
300	52	400	79800	0.6, 0.3, 0.1	X_1X_2	46	50	41	0	0	1
					X_3X_4	40	49	17	0	0	1
					X_5X_6	23	1	0	0	1	0
				0.5, 0.3, 0.2	X_1X_2	50	50	42	0	0	1
					X_3X_4	48	50	19	0	0	0
					X_5X_6	44	34	5	0	0	0
				0.4, 0.3, 0.3	X_1X_2	48	50	36	0	0	1
					X_3X_4	50	50	16	1	0	0
					X_5X_6	49	49	5	0	0	0

Table 2 presents the results for several values of n at a fixed $p = 300$. CCKIF selected X_5X_6 the most in all cases compared with the other methods. Moreover, CCKIF selected three important interactions most frequently (except in one case). In the case of $n = 200$ and the sample ratio (0.6, 0.3, 0.1), CCKIF selected X_5X_6 the most frequently, whereas KIF selected the other interactions the most. This result suggests that CCKIF may not perform well when the minor class is too small. As the sample size n increases, CCKIF, KIF, and JCIS select the important interactions more frequently. Furthermore, Table 3 shows the median and 90th percentile ranks of the estimated score for important interactions in CCKIF and KIF, based on 50 trials. While KIF outperformed CCKIF in the ranking of X_1X_2 and X_3X_4 in some cases, the rank of CCKIF score for these interactions frequently fell within the top $[n/\log n]$, allowing CCKIF to select these interactions frequently as well. For X_5X_6, CCKIF consistently outperformed KIF across all cases. When the sample ratio was (0.6, 0.3, 0.1), the median rank of the KIF score was worse than $[n/\log n]$, resulting in X_5X_6 being rarely selected. The variability between the median and the 90th percentile ranks indicates that the difference in importance between important and unimportant interactions is often small. The results of Table 3 indicate that the importance score of the interactions can be calculated more accurately with larger sample sizes, supporting Theorem 1 of CCKIF.

4.2. Simulated dataset 2

We analyzed the generated data using multinomial logistic regression models. Let $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ be a vector of predictors that follows $N(\mathbf{0}_p, \Sigma)$ with $\Sigma = (\varphi^{|j-l|})_{1 \le j,l \le p}$, and let Y be a response that takes one of the values 1, 2, or 3 and is assigned with prob-

Table 2: Results of the interaction selection for simulated dataset 1 with several n

n	$[n/\log(n)]$	p	number of all interactions	sample ratio	important interactions	CCKIF	KIF	JCIS	BCor-SIS	IP-SIS	SODA
200	37	300	44850	0.6, 0.3, 0.1	X_1X_2	35	46	24	0	0	2
					X_3X_4	17	44	11	0	0	2
					X_5X_6	8	0	1	1	0	0
				0.5, 0.3, 0.2	X_1X_2	42	47	30	0	0	1
					X_3X_4	36	40	10	0	0	0
					X_5X_6	31	16	0	0	0	0
				0.4, 0.3, 0.3	X_1X_2	42	45	21	0	0	1
					X_3X_4	40	43	3	0	0	0
					X_5X_6	41	41	5	2	0	2
300	52	300	44850	0.6, 0.3, 0.1	X_1X_2	48	50	43	0	0	2
					X_3X_4	42	50	18	1	1	2
					X_5X_6	31	3	1	0	0	0
				0.5, 0.3, 0.2	X_1X_2	50	50	44	0	0	0
					X_3X_4	50	50	20	0	0	0
					X_5X_6	48	39	5	0	1	0
				0.4, 0.3, 0.3	X_1X_2	50	50	39	0	0	1
					X_3X_4	50	50	14	0	0	1
					X_5X_6	50	50	11	0	0	0
400	66	300	44850	0.6, 0.3, 0.1	X_1X_2	50	50	48	0	0	1
					X_3X_4	50	50	32	0	1	1
					X_5X_6	48	8	2	0	0	0
				0.5, 0.3, 0.2	X_1X_2	50	50	50	0	0	1
					X_3X_4	50	50	23	0	0	2
					X_5X_6	50	49	9	1	0	0
				0.4, 0.3, 0.3	X_1X_2	50	50	42	0	0	1
					X_3X_4	50	50	21	0	0	1
					X_5X_6	50	50	20	0	0	0

abilities as follows:

$$\log\left(\frac{\mathbb{P}(Y=2 \mid \boldsymbol{X})}{\mathbb{P}(Y=1 \mid \boldsymbol{X})}\right) = X_1X_2, \quad \log\left(\frac{\mathbb{P}(Y=3 \mid \boldsymbol{X})}{\mathbb{P}(Y=1 \mid \boldsymbol{X})}\right) = X_1X_2 + X_3X_4.$$

In this setting, X_1X_2 and X_3X_4 are important interactions for classifying three classes. Under these conditions, we generated datasets with $n = 300$, $p = 400$, and $\varphi = 0.2$ and 0.5, followed by applying the same six methods used in the simulation analysis. We repeated this process 50 times and summarized the number of correctly selected interactions in each setting.

The results for the simulated dataset 2 are listed in Table 4. For $\varphi = 0.2$, CCKIF and KIF selected two important interactions in all 50 repetitions. However, in the case of $\varphi = 0.5$, CCKIF selected important interactions, especially X_1X_2, more frequently than the other methods. This difference in the results was due to the sample size of each class. Table 5 lists the average sample sizes and ratios for each class Y of the 50 datasets generated using this setting. The datasets with $\varphi = 0.5$ were more imbalanced than those with $\varphi = 0.2$ because of the smaller sample size for $Y = 1$. Therefore, CCKIF, which can select more important variables related to minor classes, selects X_1X_2 more frequently.

4.3. Real data analysis

We applied the proposed screening method to the analysis of a Human Activity dataset (Reyes-Ortiz et al., 2012). This dataset consists of sensor measurements of human activity. Thirty volunteers wore smartphones with accelerometers and gyroscopes and then repeatedly performed six activities: "walking," "walking upstairs," "walking

Table 3: Percentile ranks of estimated score of the important interactions

important interactions	sample ratio	n	p	number of all interactions	$[n/\log(n)]$	rank of CCKIF score median	rank of CCKIF score 90th percentile	rank of KIF score median	rank of KIF score 90th percentile
X_1X_2	0.6, 0.3, 0.1	200	300	44850	37	11	249	1	12
		300			52	1	6	1	2
		400			66	1	3	1	2
	0.5, 0.3, 0.2	200	300	44850	37	3	119	1	22
		300			52	1	2	1	2
		400			66	1	3	1	2
	0.4, 0.3, 0.3	200	300	44850	37	2	55	2	39
		300			52	1	3	1	2
		400			66	1	3	1	2
X_3X_4	0.6, 0.3, 0.1	200	300	44850	37	82	889	4	40
		300			52	8	113	2	3
		400			66	2	10	2	2
	0.5, 0.3, 0.2	200	300	44850	37	12	164	5	80
		300			52	2	7	2	4
		400			66	2	3	2	2
	0.4, 0.3, 0.3	200	300	44850	37	4	69	6	74
		300			52	2	5	2	5
		400			66	2	3	3	3
X_5X_6	0.6, 0.3, 0.1	200	300	44850	37	797	2603	5123	13052
		300			52	29	661	1255	6551
		400			66	4	43	586	1710
	0.5, 0.3, 0.2	200	300	44850	37	16	693	130	1594
		300			52	3	18	12	169
		400			66	3	4	3	13
	0.4, 0.3, 0.3	200	300	44850	37	3	192	3	108
		300			52	2	8	3	4
		400			66	2	3	2	3

Table 4: Results of interaction selection for simulated dataset 2

n	p	number of all interactions	φ	important interactions	CCKIF	KIF	JCIS	BCor-SIS	IP-SIS	SODA
300	400	79800	0.2	X_1X_2	50	50	0	0	0	0
				X_3X_4	50	50	7	1	6	9
			0.5	X_1X_2	42	33	5	1	3	0
				X_3X_4	47	48	23	19	41	35

downstairs," "sitting," "standing," and "laying." The sensors captured the linear acceleration and angular velocity at a constant rate of 50 Hz. We treated the labels corresponding to the six types of activities in the dataset as responses. The sample sizes for each label are 1722 for "walking," 1544 for "walking upstairs," 1406 for "walking downstairs," 1777 for "sitting," 1906 for "standing," and 1944 for "laying." In total, 561 predictors were identified from the sensors. We artificially created imbalanced datasets from this dataset by reducing the size of one of the six classes by one-tenth and then applying KIF and CCKIF to these six datasets to select the interactions. We repeated this analysis six times, changing the role of the minor class, and then compared the results of the KIF and CCKIF. For CCKIF score, we used $\pi_{k,m} = (\pi_k + \pi_m)/2$ as the settings for the simulated datasets.

First, we calculated the importance scores of all interaction terms using KIF and CCKIF and then selected the top 15 interactions based on these scores. If we select $[n/\log n]$ variables, this will exceed the total number of predictors, which is 561, resulting in the selection of all variables. However, based on the results from the analysis of the simulated data in Table 3, it is likely that the most important interactions are included

Table 5: Averages of sample sizes and ratios for each label in simulated dataset 2

	$\varphi = 0.2$			$\varphi = 0.5$		
	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 1$	$Y = 2$	$Y = 3$
sample size	91.30	94.68	114.02	68.74	89.84	141.42
sample ratio	0.304	0.316	0.380	0.229	0.300	0.471

Table 6: Classification accuracy for real dataset. Setting numbers 1 to 6 indicate the analysis of datasets with minor classes "walking," "walking upstairs," "walking downstairs," "sitting," "standing," and "laying," respectively.

	Setting number					
	1	2	3	4	5	6
KIF	0.765	0.735	0.775	0.758	0.706	0.585
CCKIF	0.765	0.735	0.783	0.741	0.711	0.712

among the top-ranked variables in terms of the importance scores. Therefore, we select 15 interactions, which corresponds to approximately 5% of the variables, or 30 variables, out of the 561. Next, we constructed decision-tree models to classify the data into six labels using the variables included in the selected interactions. The decision-tree model has internal conditional branches for each variable, allowing us to capture the effects of the interactions. We discuss the result of the obtained decision-tree model in detail at the end of this section. Finally, we compare the classification accuracies of the decision-tree models. A model that uses variables that significantly contribute to classifying a response is likely to have high classification accuracy. To validate the decision-tree model, we used five-fold cross-validation. We used grouped cross-validation to ensure that data from the same subject were not included in either the training or the validation sets. We used the Python package Lightgbm (Ke et al., 2017) to implement the decision-tree models and the GroupKFold function from the scikit-learn package (Pedregosa et al., 2011) for cross-validation.

Table 6 lists the classification accuracy for the validation data of the analysis of the six datasets. Both KIF and CCKIF yielded the same accuracy in settings 1 and 2 because they selected the same variables and yielded similar results for settings 3, 4, and 5. However, the proposed method yielded a higher accuracy than KIF in setting 6. Table 7 lists the classification accuracy of the validation data for each label in setting 6. The results of the two methods differ significantly in the accuracy of the label "laying," which has a smaller sample size than the other labels. KIF mostly fails to classify "laying" correctly, suggesting that KIF incorrectly calculates the importance scores of important variables for identifying the minor class. In contrast, the CCKIF calculates the importance scores of the variables more accurately. CCKIF has also significantly improved the accuracy of the "sitting" label.

Table 8 shows the confusion matrix representing the number of true labels and the number of labels predicted by the classification models using the variables selected by CCKIF and KIF in setting 6. From this table, we can see that the six labels can be divided into three groups: ("walking," "walking upstairs," "walking downstairs"), ("sitting," "standing"), and ("laying"). A possible reason why the accuracy of both methods did not differ significantly for settings 1 through 5 is that the imbalanced labels

Table 7: Classification accuracy for each label in setting 6 in real dataset

		True labels					
		walking	walking upstairs	walking downstairs	sitting	standing	laying
Sample size		1722	1544	1406	1777	1906	219
Accuracy	KIF	0.560	0.642	0.530	0.542	0.706	0.018
	CCKIF	0.501	0.758	0.649	0.831	0.820	0.566

Table 8: Confusion matrix in setting 6 in real dataset

	True labels					
	walking	walking upstairs	walking downstairs	sitting	standing	laying
Predicted by KIF						
walking	965	231	346	0	2	0
walking upstairs	276	992	313	0	2	0
walking downstairs	480	320	746	1	1	0
sitting	0	0	0	964	548	125
standing	1	1	1	777	1346	90
laying	0	0	0	35	7	4
Predicted by CCKIF						
walking	862	216	270	0	0	0
walking upstairs	333	1170	222	0	9	0
walking downstairs	525	156	913	1	1	0
sitting	1	0	0	1477	301	58
standing	1	2	1	262	1562	37
laying	0	0	0	37	33	124

were included in a group, and the size of the group was sufficiently large. Additionally, for settings 6, both the KIF and CCKIF incorrectly classified "laying" as either "sitting" or "standing," indicating that these three labels are similar. CCKIF selected variables to specifically identify "laying" and improved the accuracy of this label. The variables selected by CCKIF also helped to classify between "sitting" and "standing," which improved the accuracy for both labels.

Figure 1 shows an example of the decision-tree for the class "laying" in setting 6, using the variables selected by CCKIF. The tree sorts each individual according to the values of the selected variables and assigns a score based on the final leaf reached. This score represents the relative degree of belonging to the "laying" class. For example, the internal structure of this tree shows that when an individual observation satisfies "fBodyGyrosma" ≤ -0.73 and "tGravityAccenergyY" ≤ -0.96, it receives a score of -0.08 in leaf 0. Here, "fBodyGyrosma" is the signal magnitude area obtained by applying a fast Fourier transform to the gyroscope sensor data, and "tGravityAccenergyY" is the sum of the squares of Y-axis gravity acceleration data divided by the sample size. The decision-tree model consists of a series of nodes and edges based on various variables that can capture and interpret the effects of interactions. In the practical prediction process, the scores for all six classes were computed and the class with the highest score was the predicted class.

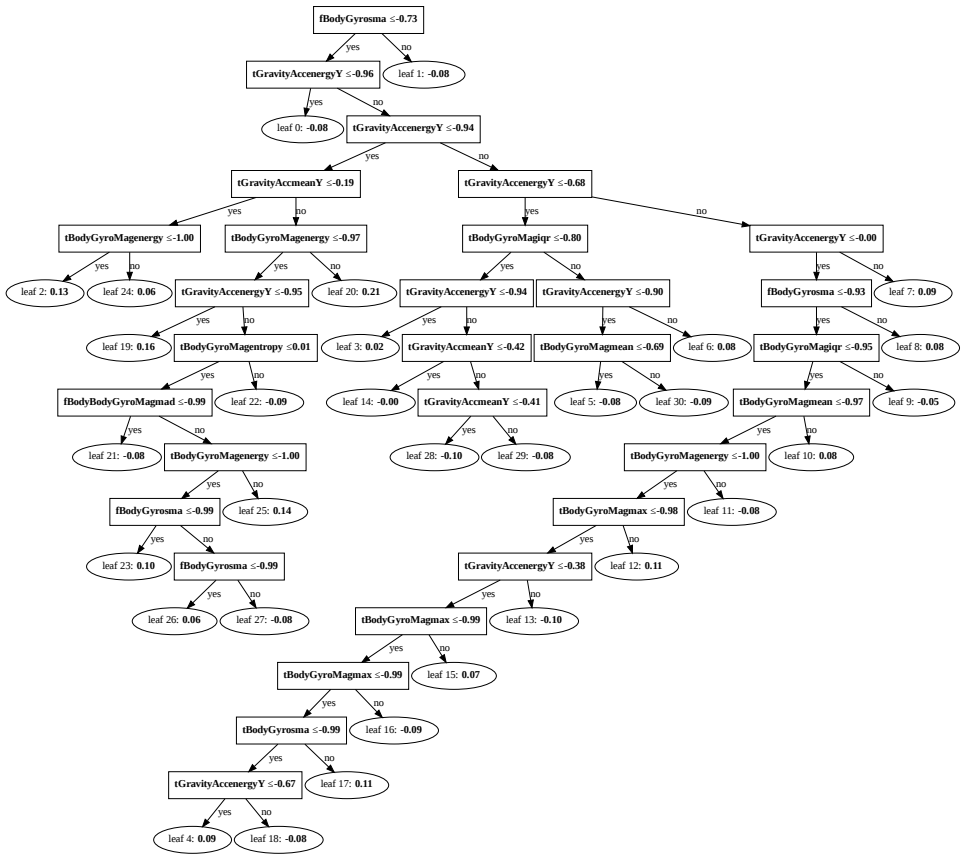

Figure 1: An example of the decision-trees created in setting 6 in real dataset

5. Discussion

We propose an interaction screening method, CCKIF, for the multiclass classification of imbalanced data. CCKIF enhances variable selection performance using the difference between Kendall's rank correlation for each class and averaged sample size ratios. We showed that the proposed method has the sure screening property with a faster convergence rate than the existing method and satisfies the ranking consistency property under some conditions. In the analysis of the simulated and real datasets, CCKIF selected important interactions more frequently than the existing methods in many cases, particularly those related to the classification of minor classes. This result suggests that CCKIF can correctly select interactions that are not considered important by the existing methods.

CCKIF sometimes performs worse in variable selection when the sample sizes of the minor classes are too small. To avoid the strong influence of a particular label, CCKIF calculates importance scores using the averaged sample size ratio. We used the arithmetic average for simulated and real data analyses. However, this approach was not always optimal. Exploring mechanisms for setting optimal average weights or incorporating additional information about data characteristics, such as the variance of predictors, may further improve the variable selection performance.

We can apply CCKIF not only to multiclass classification problems but also to regression problems with continuous responses by splitting the response into arbitrary values and then treating them as multiclass labels. For example, the response regarding income data could be segmented at thresholds where progressive tax rates change or the response regarding time data could be divided into early, mid, and late periods. Split values can be determined using business knowledge or statistical studies on the optimal choice of cut points.

In this paper, we introduced two methods for selecting important interactions using CCKIF: one that selects interactions with an importance score above a certain threshold and another that selects the top $[n/\log n]$ interactions based on their importance scores. In the simulation studies and real data analysis conducted in this work, the threshold for selecting important interactions is determined by existing methods or a given number. The exploration of more efficient and objective methods for determining the number of variables to be selected remains a topic for future work.

Acknowledgement

The authors are grateful to the anonymous reviewer for valuable comments and suggestions that improve the quality of this paper. This work was supported by JSPS KAKENHI Grant Numbers 19K11858 and 23K11005.

References

Anzarmou, Y., Mkhadri, A., and Oualkacha, K. (2023). The Kendall interaction filter for variable interaction screening in high dimensional classification problems. *Journal of Applied Statistics*, 50(7):1496–1514.

Cui, H., Li, R., and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.

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

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

Fan, Y., Kong, Y., Li, D., and Lv, J. (2016). Interaction pursuit with feature screening and selection. *arXiv preprint arXiv:1605.08933*.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301.

Huang, D., Li, R., and Wang, H. (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business & Economic Statistics*, 32(2):237–244.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3149–3157.

Kou, G., Xu, Y., Peng, Y., Shen, F., Chen, Y., Chang, K., and Kou, S. (2021). Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. *Decision Support Systems*, 140:113429.

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

Li, R., Zhong, W., and Zhu, L. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Li, Y. and Liu, J. S. (2019). Robust variable and interaction selection for logistic regression and general index models. *Journal of the American Statistical Association*, 114(525):271–286.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2019a). A generic sure independence screening procedure. *Journal of the American Statistical Association*, 114(526):928.

Pan, W., Wang, X., Zhang, H., Zhu, H., and Zhu, J. (2019b). Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Reese, R. D. (2018). *Feature screening of ultrahigh dimensional feature spaces with applications in interaction screening*. PhD thesis, Utah State University.

Reyes-Ortiz, J., Anguita, D., Ghio, A., Oneto, L., and Parra, X. (2012). Human Activity Recognition Using Smartphones. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C54S4K.

Valentini, M., dos Santos, G. B., and Muller Vieira, B. (2021). Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul—Brazil. *SN Applied Sciences*, 3:1–11.

Wu, D., Zhou, D., Zhang, J., and Chen, M. (2020). Multimode process monitoring based on fault dependent variable selection and moving window-negative log likelihood probability. *Computers & Chemical Engineering*, 136:106787.

Received: June 24, 2024
Revised: August 29, 2024
Accept: September 8, 2024

Appendix: Proofs

Lemma 1 is used as proof of Theorem 1. Lemma 1 states that Kendall's rank correlation τ_k has unbiasedness and consistency.

LEMMA 1. *For $k, m \in \{1, \ldots, K\}$ and $j, l \in \{1, \ldots, p\}$, τ_k verifies the following*

$$(i) \quad \mathbb{E}(\hat{\pi}_{k,m}\hat{\tau}_k(X_j, X_l) \mid Y) = \hat{\pi}_{k,m}\tau_k(X_j, X_l); \tag{2}$$

$$(ii) \quad \mathbb{P}(\hat{\pi}_{k,m}\,|\hat{\tau}_k(X_j, X_l)) - \tau_k(X_j, X_l)| > \varepsilon \mid Y) \leq 2\exp\left(\frac{-\hat{\pi}_k^2 n\varepsilon^2}{8\hat{\pi}_{k,m}^2}\right), \text{ for all } \varepsilon > 0. \tag{3}$$

Proof of Lemma 1

From the definitions of τ_k and $\hat{\tau}_k$, (i) holds. We provide the proof for (ii). For $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathbb{R}^3$ and $k \in \{1, \ldots, K\}$, we consider the following function:

$$\begin{aligned} &g_{k,m}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \\ &\quad = \frac{4\hat{\pi}_{k,m}}{n_k(n_k-1)} \sum_{i=1}^{n-1} \sum_{t=i+1}^{n} \mathbb{1}\left\{(v_{i1} - v_{t1})(v_{i2} - v_{t2}) > 0, v_{i3} = k, v_{t3} = k\right\} - \hat{\pi}_{k,m}, \end{aligned} \tag{4}$$

where n_k is the sample size of class k. Let $g_{k,m}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q)$ be a function that replaces the vector $\boldsymbol{v}_n$ in (4) with $\boldsymbol{v}_q$. The absolute value of the difference between the two functions is given by

$$\begin{aligned} &|g_{k,m}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) - g_{k,m}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q)| \\ &= \left| \frac{4\hat{\pi}_{k,m}}{n_k(n_k-1)} \left[\sum_{i=1}^{n-2} \sum_{t=i+1}^{n-1} \mathbb{1}\left\{(v_{i1} - v_{t1})(v_{i2} - v_{t2}) > 0, v_{i3} = k, v_{t3} = k\right\} \right.\right. \\ &\qquad \left. + \sum_{i=1}^{n-1} \mathbb{1}\left\{(v_{i1} - v_{n1})(v_{i2} - v_{n2}) > 0, v_{i3} = k, v_{n3} = k\right\} \right] \\ &\quad - \frac{4\hat{\pi}_{k,m}}{n_k(n_k-1)} \left[\sum_{i=1}^{n-2} \sum_{t=i+1}^{n-1} \mathbb{1}\left\{(v_{i1} - v_{t1})(v_{i2} - v_{t2}) > 0, v_{i3} = k, v_{t3} = k\right\} \right. \\ &\qquad \left.\left. + \sum_{i=1}^{n-1} \mathbb{1}\left\{(v_{i1} - v_{q1})(v_{i2} - v_{q2}) > 0, v_{i3} = k, v_{q3} = k\right\} \right] \right| \\ &= \left| \frac{4\hat{\pi}_{k,m}}{n_k(n_k-1)} \sum_{i=1}^{n-1} \left[\mathbb{1}\left\{(v_{i1} - v_{n1})(v_{i2} - v_{n2}) > 0, v_{i3} = k, v_{n3} = k\right\} \right.\right. \\ &\qquad \left.\left. - \mathbb{1}\left\{(v_{i1} - v_{q1})(v_{i2} - v_{q2}) > 0, v_{i3} = k, v_{q3} = k\right\} \right] \right| \\ &\leq \frac{4\hat{\pi}_{k,m}}{n_k} \\ &= \frac{4\hat{\pi}_{k,m}}{\hat{\pi}_k n}. \end{aligned}$$

Let $\left\{\boldsymbol{V}_i = (V_{i1}, V_{i2}, V_{i3})^\top, 1 \leq i \leq n\right\}$ be a set of i.i.d. random vectors and $Y = (V_{13}, \ldots, V_{n3})^\top$. Applying McDiarmid's inequality to $g_{k,m}$, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left| g_{k,m}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n) - \mathbb{E}\Big(g_{k,m}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n) \mid Y \Big) \right| > \varepsilon \mid Y \right) \\
&\leq 2 \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n \left(4\hat{\pi}_{k,m}/\hat{\pi}_k n\right)^2} \right) \\
&= 2 \exp\left(\frac{-\hat{\pi}_k^2 n \varepsilon^2}{8\hat{\pi}_{k,m}^2} \right).
\end{aligned}$$

Let $X = (V_{11}, \ldots, V_{n1})^\top$, $X' = (V_{12}, \ldots, V_{n2})^\top$, then we have $g_{k,m}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n) = \hat{\pi}_{k,m}\hat{\tau}_k(X, X')$. Using (2), $\mathbb{E}(g_{k,m}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n) \mid Y) = \mathbb{E}(\hat{\pi}_{k,m}\hat{\tau}_k(X, X')) \mid Y) = \hat{\pi}_{k,m}\tau_k(X, X')$. Using them, we have

$$\begin{aligned}
&\mathbb{P}\left(\left| g_{k,m}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n) - \mathbb{E}\Big(g_{k,m}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n) \mid Y \Big) \right| > \varepsilon \mid Y \right) \\
&= \mathbb{P}\left(\left| \hat{\pi}_{k,m}\hat{\tau}_k(X, X') - \hat{\pi}_{k,m}\tau_k(X, X') \right| > \varepsilon \mid Y \right) \\
&= \mathbb{P}\left(\hat{\pi}_{k,m} \left| \hat{\tau}_k(X, X') - \tau_k(X, X') \right| > \varepsilon \mid Y \right) \\
&\leq 2 \exp\left(\frac{-\hat{\pi}_k^2 n \varepsilon^2}{8\hat{\pi}_{k,m}^2} \right).
\end{aligned}$$

□

Proof of Theorem 1

First, we prove (i). For $j, l \in \{1, \ldots, p\}$, we obtain

$$
\begin{aligned}
&|\hat{w}_{j,l} - w_{j,l}| \\
&= \frac{1}{K^2}\left| \sum_{k=1}^{K}\sum_{m=1}^{K} \hat{\pi}_{k,m} \left|\hat{\tau}_k(X_j, X_l) - \hat{\tau}_m(X_j, X_l)\right| - \sum_{k=1}^{K}\sum_{m=1}^{K} \pi_{k,m} \left|\tau_k(X_j, X_l) - \tau_m(X_j, X_l)\right| \right| \\
&= \frac{1}{K^2}\left| \sum_{k=1}^{K}\sum_{m=1}^{K} \hat{\pi}_{k,m} \left|\hat{\tau}_k(X_j, X_l) - \hat{\tau}_m(X_j, X_l)\right| - \sum_{k=1}^{K}\sum_{m=1}^{K} \hat{\pi}_{k,m} \left|\tau_k(X_j, X_l) - \tau_m(X_j, X_l)\right| \right. \\
&\qquad \left. + \sum_{k=1}^{K}\sum_{m=1}^{K} \hat{\pi}_{k,m} \left|\tau_k(X_j, X_l) - \tau_m(X_j, X_l)\right| - \sum_{k=1}^{K}\sum_{m=1}^{K} \pi_{k,m} \left|\tau_k(X_j, X_l) - \tau_m(X_j, X_l)\right| \right| \\
&\leq \frac{1}{K^2}\left\{ \sum_{k=1}^{K}\sum_{m=1}^{K} \hat{\pi}_{k,m} \left| \left|\hat{\tau}_k(X_j, X_l) - \hat{\tau}_m(X_j, X_l)\right| - \left|\tau_k(X_j, X_l) - \tau_m(X_j, X_l)\right| \right| \right. \\
&\qquad \left. + \sum_{k=1}^{K}\sum_{m=1}^{K} \left|\hat{\pi}_{k,m} - \pi_{k,m}\right| \left|\tau_k(X_j, X_l) - \tau_m(X_j, X_l)\right| \right\} \\
&\leq \frac{1}{K^2}\left\{ \sum_{k=1}^{K}\sum_{m=1}^{K} \hat{\pi}_{k,m} \left|\left(\hat{\tau}_k(X_j, X_l) - \tau_k(X_j, X_l)\right) - \left(\hat{\tau}_m(X_j, X_l) - \tau_m(X_j, X_l)\right)\right| \right. \\
&\qquad \left. + \sum_{k=1}^{K}\sum_{m=1}^{K} \left|\hat{\pi}_{k,m} - \pi_{k,m}\right| \left|\tau_k(X_j, X_l) - \tau_m(X_j, X_l)\right| \right\} \\
&\leq \frac{1}{K^2}\left\{ \sum_{k=1}^{K}\sum_{m=1}^{K} \hat{\pi}_{k,m} \left|\hat{\tau}_k(X_j, X_l) - \tau_k(X_j, X_l)\right| + \sum_{k=1}^{K}\sum_{m=1}^{K} \hat{\pi}_{k,m} \left|\hat{\tau}_m(X_j, X_l) - \tau_m(X_j, X_l)\right| \right. \\
&\qquad \left. + \sum_{k=1}^{K}\sum_{m=1}^{K} \left|\hat{\pi}_{k,m} - \pi_{k,m}\right| \left|\tau_k(X_j, X_l) - \tau_m(X_j, X_l)\right| \right\} \\
&\leq \frac{1}{K^2}\left\{ \sum_{k=1}^{K}\sum_{m=1}^{K} \hat{\pi}_{k,m} \left|\hat{\tau}_k(X_j, X_l) - \tau_k(X_j, X_l)\right| + \sum_{k=1}^{K}\sum_{m=1}^{K} \hat{\pi}_{k,m} \left|\hat{\tau}_m(X_j, X_l) - \tau_m(X_j, X_l)\right| \right. \\
&\qquad \left. + 2\sum_{k=1}^{K}\sum_{m=1}^{K} \left|\hat{\pi}_{k,m} - \pi_{k,m}\right| \right\}.
\end{aligned}
$$

Then

$$
\begin{aligned}
&\mathbb{P}(|\hat{w}_{j,l} - w_{j,l}| > \varepsilon) \\
&\quad \leq \mathbb{P}\left(\frac{1}{K^2} \left\{ \sum_{k=1}^{K} \sum_{m=1}^{K} \hat{\pi}_{k,m} \, |\hat{\tau}_k(X_j, X_l) - \tau_k(X_j, X_l)| \right.\right. \\
&\qquad \left.\left. + \sum_{k=1}^{K} \sum_{m=1}^{K} \hat{\pi}_{k,m} \, |\hat{\tau}_m(X_j, X_l) - \tau_m(X_j, X_l)| + 2 \sum_{k=1}^{K} \sum_{m=1}^{K} |\hat{\pi}_{k,m} - \pi_{k,m}| \right\} > \varepsilon \right) \\
&\quad \leq \mathbb{P}\left(\frac{1}{K^2} \sum_{k=1}^{K} \sum_{m=1}^{K} \hat{\pi}_{k,m} \, |\hat{\tau}_k(X_j, X_l) - \tau_k(X_j, X_l)| > \frac{\varepsilon}{3} \right) \\
&\qquad + \mathbb{P}\left(\frac{1}{K^2} \sum_{k=1}^{K} \sum_{m=1}^{K} \hat{\pi}_{k,m} \, |\hat{\tau}_m(X_j, X_l) - \tau_m(X_j, X_l)| > \frac{\varepsilon}{3} \right) \\
&\qquad + \mathbb{P}\left(\frac{2}{K^2} \sum_{k=1}^{K} \sum_{m=1}^{K} |\hat{\pi}_{k,m} - \pi_{k,m}| > \frac{\varepsilon}{3} \right) \\
&\quad = 2\mathbb{P}\left(\frac{1}{K^2} \sum_{k=1}^{K} \sum_{m=1}^{K} \hat{\pi}_{k,m} \, |\hat{\tau}_k(X_j, X_l) - \tau_k(X_j, X_l)| > \frac{\varepsilon}{3} \right) \\
&\qquad + \mathbb{P}\left(\frac{2}{K^2} \sum_{k=1}^{K} \sum_{m=1}^{K} |\hat{\pi}_{k,m} - \pi_{k,m}| > \frac{\varepsilon}{3} \right). \qquad (5)
\end{aligned}
$$

Using subadditivity, Hoeffding's inequality, and (3) in Lemma 1 (ii), we obtain

$$
\begin{aligned}
&2\mathbb{P}\left(\frac{1}{K^2} \sum_{k=1}^{K} \sum_{m=1}^{K} \hat{\pi}_{k,m} \, |\hat{\tau}_k(X_j, X_l) - \tau_k(X_j, X_l)| > \frac{\varepsilon}{3} \right) \\
&\quad \leq 2 \sum_{k=1}^{K} \sum_{m=1}^{K} \mathbb{P}\left(\hat{\pi}_{k,m} \, |\hat{\tau}_k(X_j, X_l) - \tau_k(X_j, X_l)| > \frac{\varepsilon}{3} \right) \\
&\quad \leq 2 \sum_{k=1}^{K} \sum_{m=1}^{K} \mathbb{E}_Y \left[\mathbb{P}\left(\hat{\pi}_{k,m} \, |\hat{\tau}_k(X_j, X_l) - \tau_k(X_j, X_l)| > \frac{\varepsilon}{3} \,\middle|\, Y \right) \right] \\
&\quad \leq 2 \sum_{k=1}^{K} \sum_{m=1}^{K} 2 \exp\left(\frac{-\hat{\pi}_k^2 n (\varepsilon/3)^2}{8 \hat{\pi}_{k,m}^2} \right) \\
&\quad = 4 \sum_{k=1}^{K} \sum_{m=1}^{K} \exp\left(\frac{-\hat{\pi}_k^2 n \varepsilon^2}{72 \hat{\pi}_{k,m}^2} \right), \qquad (6)
\end{aligned}
$$

and

$$\begin{aligned}
\mathbb{P}\left(\frac{2}{K^2}\sum_{k=1}^{K}\sum_{m=1}^{K}|\hat{\pi}_{k,m}-\pi_{k,m}|>\frac{\varepsilon}{3}\right) &\leq \mathbb{P}\left(2\max_{k,m}|\hat{\pi}_{k,m}-\pi_{k,m}|>\frac{\varepsilon}{3}\right)\\
&\leq \mathbb{P}\left(\bigcup_{k,m}\left\{|\hat{\pi}_{k,m}-\pi_{k,m}|>\frac{\varepsilon}{6}\right\}\right)\\
&\leq \sum_{k=1}^{K}\sum_{m=1}^{K}\mathbb{P}\left(|\hat{\pi}_{k,m}-\pi_{k,m}|>\frac{\varepsilon}{6}\right)\\
&\leq 2\sum_{k=1}^{K}\sum_{m=1}^{K}\exp\left(-2n\left(\frac{\varepsilon}{6}\right)^2\right)\\
&\leq 2K^2\exp\left(\frac{-n\varepsilon^2}{18}\right). \qquad (7)
\end{aligned}$$

Using (5), (6), and (7), we have that

$$\mathbb{P}\left(|\hat{w}_{j,l}-w_{j,l}|>\varepsilon\right)\leq 4\sum_{k=1}^{K}\sum_{m=1}^{K}\exp\left(\frac{-\hat{\pi}_k^2 n\varepsilon^2}{72\hat{\pi}_{k,m}^2}\right)+2K^2\exp\left(\frac{-n\varepsilon^2}{18}\right).$$

As the values of j and l are at most p,

$$\begin{aligned}
\mathbb{P}\left(\max_{1\leq j,l\leq p}|\hat{w}_{j,l}-w_{j,l}|>\varepsilon\right) &\leq \sum_{j=1}^{p}\sum_{l=1}^{p}\mathbb{P}\left(|\hat{w}_{j,l}-w_{j,l}|>\varepsilon\right)\\
&\leq 4p^2\sum_{k=1}^{K}\sum_{m=1}^{K}\exp\left(\frac{-\hat{\pi}_k^2 n\varepsilon^2}{72\hat{\pi}_{k,m}^2}\right)+2p^2K^2\exp\left(\frac{-n\varepsilon^2}{18}\right). \qquad (8)
\end{aligned}$$

Using conditions (C1)–(C5), we obtain

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq j,l\leq p}|\hat{w}_{j,l}-w_{j,l}|\leq cn^{-r}\right)\\
&=1-\mathbb{P}\left(\max_{1\leq j,l\leq p}|\hat{w}_{j,l}-w_{j,l}|>cn^{-r}\right)\\
&\geq 1-4p^2\sum_{k=1}^{K}\sum_{m=1}^{K}\exp\left(\frac{-\hat{\pi}_k^2 n(cn^{-r})^2}{72\hat{\pi}_{k,m}^2}\right)+2p^2K^2\exp\left(\frac{-n(cn^{-r})^2}{18}\right)\\
&\geq 1-O\left(p^2n^{2d}\exp\left(\frac{-c^2n^{1-2r}}{72b^2}\right)\right).
\end{aligned}$$

This completes the proof of (i).

Next, we demonstrate (ii). We assume $\boldsymbol{S}\not\subseteq\hat{\boldsymbol{S}}$, then there exists $(j,l)\in\boldsymbol{S}$ satisfying

$\hat{w}_{j,l} \leq cn^{-r}$. Under condition (C2), $|\hat{w}_{j,l} - w_{j,l}| > cn^{-r}$. Therefore, we have

$$\begin{aligned}\{\boldsymbol{S} \not\subseteq \hat{\boldsymbol{S}}\} &\subseteq \{|\hat{w}_{j,l} - w_{j,l}| > cn^{-r}, \text{ for a certain } (j,l) \in \boldsymbol{S}\} \\ &\subseteq \{\max_{(j,l)\in\boldsymbol{S}} |\hat{w}_{j,l} - w_{j,l}| > cn^{-r}\}.\end{aligned}$$

From (8) and $|\boldsymbol{S}| = s$, we obtain that

$$\begin{aligned}\mathbb{P}(\boldsymbol{S} \not\subseteq \hat{\boldsymbol{S}}) &\leq \mathbb{P}\left(\max_{(j,l)\in\boldsymbol{S}} |\hat{w}_{j,l} - w_{j,l}| > cn^{-r}\right) \\ &\leq 4s \sum_{k=1}^{K}\sum_{m=1}^{K} \exp\left(\frac{-\hat{\pi}_k^2 n(cn^{-r})^2}{72\hat{\pi}_{k,m}^2}\right) + 2sK^2 \exp\left(\frac{-n(cn^{-r})^2}{18}\right).\end{aligned}$$

Using conditions (C3)–(C5), we obtain

$$\begin{aligned}\mathbb{P}(\boldsymbol{S} \subseteq \hat{\boldsymbol{S}}) &= 1 - \mathbb{P}(\boldsymbol{S} \not\subseteq \hat{\boldsymbol{S}}) \\ &\geq 1 - O\left(sn^{2d} \exp\left(\frac{-c^2 n^{1-2r}}{72b^2}\right)\right).\end{aligned}$$

This completes the proof of (ii).

□

Proof of Theorem 2

Under condition (C6) and (8), we have that

$$
\begin{aligned}
&\mathbb{P}\left[\left(\min_{(j,l)\in\boldsymbol{S}}\hat{w}_{j,l}-\max_{(j,l)\in\boldsymbol{S}^c}\hat{w}_{j,l}\right)<\frac{c_3}{2}\right]\\
&\leq\mathbb{P}\left[\left(\min_{(j,l)\in\boldsymbol{S}}\hat{w}_{j,l}-\max_{(j,l)\in\boldsymbol{S}^c}\hat{w}_{j,l}\right)-\left(\min_{(j,l)\in\boldsymbol{S}}w_{j,l}-\max_{(j,l)\in\boldsymbol{S}^c}w_{j,l}\right)<-\frac{c_3}{2}\right]\\
&\leq\mathbb{P}\left[\left|\left(\min_{(j,l)\in\boldsymbol{S}}\hat{w}_{j,l}-\min_{(j,l)\in\boldsymbol{S}}w_{j,l}\right)-\left(\max_{(j,l)\in\boldsymbol{S}^c}\hat{w}_{j,l}-\max_{(j,l)\in\boldsymbol{S}^c}w_{j,l}\right)\right|>\frac{c_3}{2}\right]\\
&\leq\mathbb{P}\left[2\max_{1\leq j,l\leq p}|\hat{w}_{j,l}-w_{j,l}|>\frac{c_3}{2}\right]\\
&=\mathbb{P}\left[\max_{1\leq j,l\leq p}|\hat{w}_{j,l}-w_{j,l}|>\frac{c_3}{4}\right]\\
&\leq 4p^2\sum_{k=1}^{K}\sum_{m=1}^{K}\exp\left(\frac{-\hat{\pi}_k^2 n(c_3/4)^2}{72\hat{\pi}_{k,m}^2}\right)+2p^2K^2\exp\left(\frac{-n(c_3/4)^2}{18}\right)\\
&=4p^2\sum_{k=1}^{K}\sum_{m=1}^{K}\exp\left(\frac{-\hat{\pi}_k^2 nc_3^2}{1152\hat{\pi}_{k,m}^2}\right)+2p^2K^2\exp\left(\frac{-nc_3^2}{288}\right)\\
&\leq 4p^2K^2\exp\left(-\xi c_5 n\right)+2p^2K^2\exp(-4c_5 n),
\end{aligned}
\tag{9}
$$

where $\xi=\min(\hat{\pi}_k^2/\hat{\pi}_{k,m}^2)$ and $c_5=c_3^2/1152$.

Each term is transformed into the following equation:

$$
\begin{aligned}
4p^2K^2\exp\left(-\xi c_5 n\right)&=4\exp\left(2\log(p)+2\log(K)-\xi c_5 n\right),\\
2p^2K^2\exp\left(-4c_5 n\right)&=2\exp\left(2\log(p)+2\log(K)-4c_5 n\right).
\end{aligned}
$$

Under condition (C4), when n is large, $\log(p)<C_3n^{1-2r}<C_3n$ where C_3 is a constant. Furthermore, because $\log(n)=o(n)$, $\log(n)<C_4n$ where C_4 is a constant. Let $C_3=\xi c_5/4$ and $C_4=\xi c_5/8$, then we have $2\log(p)<\xi c_5 n/2$ and $4\log(n)<\xi c_5 n/2$. Using the above results and $K<n$, we obtain

$$
\begin{aligned}
4p^2K^2\exp\left(-\xi c_5 n\right)&=4\exp\left(2\log(p)+2\log(K)-\xi c_5 n\right)\\
&<4\exp\left(\frac{\xi c_5 n}{2}+2\log(n)-\xi c_5 n\right)\\
&=4\exp\left(2\log(n)-\frac{\xi c_5 n}{2}\right)\\
&<4\exp\left(2\log(n)-4\log(n)\right)\\
&=\frac{4}{n^2}.
\end{aligned}
$$

Let $C_3=c_5$ and $C_4=c_5/2$, then $2\log(p)<2c_5 n$ and $4\log(n)<2c_5 n$. Therefore, we

obtain

$$\begin{aligned} 2p^2K^2\exp\left(-4c_5n\right) &= 2\exp\left(2\log(p)+2\log(K)-4c_5n\right) \\ &< 2\exp\left(2c_5n+2\log(n)-4c_5n\right) \\ &= 2\exp\left(2\log(n)-2c_5n\right) \\ &= 2\exp\left(2\log(n)-4\log(n)\right) \\ &= \frac{2}{n^2}. \end{aligned}$$

From (9), we obtain

$$\begin{aligned} \mathbb{P}\left\{\left(\min_{(j,l)\in\boldsymbol{S}}\hat{w}_{j,l}-\max_{(j,l)\in\boldsymbol{S}^c}\hat{w}_{j,l}\right)<\frac{c_3}{2}\right\} &\leq 4p^2K^2\exp\left(-\xi c_5n\right)+2p^2K^2\exp(-4c_5n) \\ &< \frac{4}{n^2}+\frac{2}{n^2} \\ &= \frac{6}{n^2}. \end{aligned}$$

For some n_0, using the result of the Basel problem, we have

$$\sum_{n=n_0}^{\infty}\mathbb{P}\left\{\left(\min_{(j,l)\in\boldsymbol{S}}\hat{w}_{j,l}-\max_{(j,l)\in\boldsymbol{S}^c}\hat{w}_{j,l}\right)<\frac{c_3}{2}\right\}<6\sum_{n=1}^{\infty}\frac{1}{n^2}<\infty.$$

Consequently, using the Borel-Cantelli's lemma, we obtain

$$\liminf_{n\to\infty}\left\{\min_{(j,l)\in\boldsymbol{S}}\hat{w}_{j,l}-\max_{(j,l)\in\boldsymbol{S}^c}\hat{w}_{j,l}\right\}>0 \text{ a.s.}$$

□