Supplementary Material

Xiaxue Ouyang[1], Mengyu Li[2], Jun Yu[3], and Cheng Meng[4,*]

[1]Institute of Statistics and Big Data, Renmin University of China
[2]Department of Statistics and Data Science, Tsinghua University
[3]School of Mathematics and Statistics, Beijing Institute of Technology
[4]Center for Applied Statistics, Institute of Statistics and Big Data,
Renmin University of China
[*]Corresponding author: Cheng Meng, chengmeng@ruc.edu.cn

1 Preliminaries

Suppose that $(X_1, Y_1), \ldots (X_n, Y_n)$ are the pairs with the same distribution as (X, Y). Rearrange the data as $(X_{(1)}, Y_{(1)}), \ldots (X_{(n)}, Y_{(n)})$ such that $X_{(1)} \leq \cdots \leq X_{(n)}$. Define

$$\omega = \int \text{var}(\mathbb{E}(\mathbb{I}_{\{Y \geq t\}}|X))d\mu(t), \tag{1}$$

$$\widehat{\omega} = \frac{1}{n^3}\sum_{i=1}^{n} l_i(n - l_i) - \frac{1}{2n^2}\sum_{i=1}^{n-1}|r_{i+1} - r_i| + \frac{r_n - r_1}{2n^2}, \tag{2}$$

where $\mu(t)$ is the law of Y, r_i is the number of j such that $Y_{(j)} \leq Y_{(i)}$ and l_i is the number of j such that $Y_{(j)} \geq Y_{(i)}$. Here we introduce the following notations to simplify our subsequent expression. Let

$$Q := \int \text{var}(\mathbb{E}(\mathbb{I}_{\{\mathrm{Y} \geq \mathrm{t}\}}|\mathrm{X}))\mathrm{d}\mu(\mathrm{t}), \tag{3}$$

$$Q_n := \frac{1}{n}\sum_{i=1}^{n} \min\{F_n(Y_i), F_n(Y_{N(i)})\} - \frac{1}{n}\sum_{i=1}^{n} G_n^2(Y_i), \tag{4}$$

$$\widetilde{Q_n} := \frac{1}{n}\sum_{i=1}^{n} \min\{F(Y_i), F(Y_{N(i)})\} - \frac{1}{n}\sum_{i=1}^{n} G^2(Y_i), \tag{5}$$

where $\pi(i)$ is the rank of X_i,

$$N(i) := \begin{cases} \pi^{-1}(\pi(i) + 1) & \text{if} \quad \pi(i) < \mathrm{n} \\ i & \text{if} \quad \pi(i) = \mathrm{n}, \end{cases}$$

and for any $t \in \mathbb{R}$,

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Y_i \le t\}}, \quad G_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{Y_i \ge t\}},$$
$$F(t) := \mathbb{P}\{Y \le t\}, \quad G(t) := \mathbb{P}\{Y \ge t\}.$$

Now, We can derive the following lemmas to help us build our proof.

Lemma 1. *For any sequences $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$, where $a_i, b_i \in \mathbb{R}$, we have*

$$\left|\min_{1\le i\le n} a_i - \min_{1\le i\le n} b_i\right| \le \max_{1\le i\le n} |a_i - b_i|,$$

and symmetrically,

$$\left|\max_{1\le i\le n} a_i - \max_{1\le i\le n} b_i\right| \le \max_{1\le i\le n} |a_i - b_i|.$$

Proof. For the first part, we can only consider the terms that minimize $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ respectively. Suppose $a_k = \min_{1\le i\le n} a_i$ and $b_l = \min_{1\le i\le n} b_i$, where $1 \le k, l \le n$. If $l = k$, the claim holds immediately. If $l \ne k$, we can only need to prove that

$$|a_k - b_l| \le \max\{|a_k - b_k|, |a_l - b_l|\}.$$

Note that $a_k \le a_l$ and $b_l \le b_k$. There are only six cases for the order of a_k, a_l, b_k, b_l as follows: when (1) $a_k \le a_l \le b_l \le b_k$, (2) $a_k \le b_l \le a_l \le b_k$ and (3) $a_k \le b_l \le b_k \le a_l$, then $|a_k - b_l| \le |a_k - b_k|$; when (4) $b_l \le a_k \le a_l \le b_k$, (5) $b_l \le a_k \le b_k \le a_l$, and (6) $b_l \le b_k \le a_k \le a_l$, then $|a_k - b_l| \le |a_l - b_l|$. Thus we have the claim. The second part is obvious if we let $a_i = -a_i$ and $b_i = -b_i$.

□

Lemma 2. *(Chatterjee, 2021) As $n \to \infty$, $Q_n \to Q$ almost surely. Furthermore, we have*

$$Q_n = \frac{1}{n^2}\sum_{i=1}^{n} r_i - \frac{1}{2n^2}\sum_{i=1}^{n-1} |r_{i+1} - r_i| + \frac{r_n - r_1}{2n^2} - \frac{1}{n^3}\sum_{i=1}^{n} l_i^2.$$

Then, $\widehat{\omega} \to \omega$ almost surely as $n \to \infty$. That's to say $\widehat{\omega}$ is a consistent estimator of ω.

Lemma 3. $\mathbb{E}(\widetilde{Q_n}) = Q$.

Proof. First, note that for any $1 \le i \le n$,

$$\begin{aligned}
\min\{F(Y_i), F(Y_{N(i)})\} &= \min\{\mathbb{P}(Y \le Y_i), \mathbb{P}(Y \le Y_{N(i)})\} \\
&= \mathbb{P}(Y \le \min\{Y_i, Y_{N(i)}\}) \\
&= \int \mathbb{I}_{\{\min\{Y_i, Y_{N(i)}\} \ge t\}} d\mu(t) \\
&= \int \mathbb{I}_{\{Y_i \ge t\}} \mathbb{I}_{\{Y_{N(i)} \ge t\}} d\mu(t).
\end{aligned}$$

Thus,

$$\mathbb{E}\big[\min\{F(Y_i), F(Y_{N(i)})\}\big] = \mathbb{E}\Big[\int \mathbb{I}_{\{Y_i\geq t\}}\mathbb{I}_{\{Y_{N(i)}\geq t\}} d\mu(t)\Big] = \int \mathbb{E}\big(\mathbb{I}_{\{Y_i\geq t\}}\mathbb{I}_{\{Y_{N(i)}\geq t\}}\big) d\mu(t).$$

Let $\mathcal{F}$ be the σ-algebra generated by the X_i's and the randomness used for breaking ties in the selection of π. Then for any t,

$$\mathbb{E}\big(\mathbb{I}_{\{Y_i\geq t\}}\mathbb{I}_{\{Y_{N(i)}\geq t\}}\big|\mathcal{F}\big) = \mathbb{E}\big(\mathbb{I}_{\{Y_i\geq t\}}\big|X_i\big)\mathbb{E}\big(\mathbb{I}_{\{Y_{N(i)}\geq t\}}\big|X_{N(i)}\big) = \mathbb{E}^2\big(\mathbb{I}_{\{Y\geq t\}}\big|X\big).$$

Therefore, by double expectation theorem,

$$\begin{aligned}
\mathbb{E}(\widetilde{Q_n}) &= \int \mathbb{E}\big(\mathbb{I}_{\{Y_1\geq t\}}\mathbb{I}_{\{Y_{N(1)}\geq t\}}\big) d\mu(t) - \int G^2(t) d\mu(t) \\
&= \int \mathbb{E}\big[\mathbb{E}\big(\mathbb{I}_{\{Y_i\geq t\}}\mathbb{I}_{\{Y_{N(i)}\geq t\}}\big|\mathcal{F}\big)\big] d\mu(t) - \int G^2(t) d\mu(t) \\
&= \int \mathbb{E}[\mathbb{E}^2(\mathbb{I}_{\{Y\geq t\}}|X)] d\mu(t) - \int G^2(t) d\mu(t).
\end{aligned}$$

At the same time,

$$\begin{aligned}
Q &= \int \Big\{\mathbb{E}[\mathbb{E}^2(\mathbb{I}_{\{Y\geq t\}}|X)] - \mathbb{E}^2[\mathbb{E}(\mathbb{I}_{\{Y\geq t\}}|X)]\Big\} d\mu(t) \\
&= \int \Big\{\mathbb{E}[\mathbb{E}^2(\mathbb{I}_{\{Y\geq t\}}|X)] - \mathbb{E}^2(\mathbb{I}_{\{Y\geq t\}})\Big\} d\mu(t) \\
&= \int \mathbb{E}[\mathbb{E}^2(\mathbb{I}_{\{Y\geq t\}}|X)] d\mu(t) - \int G^2(t) d\mu(t).
\end{aligned}$$

Hence, $\mathbb{E}(\widetilde{Q_n}) = Q$. □

Lemma 4. *Q_n and $\widetilde{Q_n}$ are defined as before, we have*

$$|Q_n - \widetilde{Q_n}| \leq \sup_{y\in\mathbb{R}} |F_n(y) - F(y)| + 2\sup_{y\in\mathbb{R}} |G_n(y) - G(y)|.$$

Proof. By the first conclusion in Lemma 1, we can note that for any $1 \leq i \leq n$,

$$\begin{aligned}
\big|\min\{F_n(Y_i), F_n(Y_{N(i)})\} - \min\{F(Y_i), F(Y_{N(i)})\}\big| &\leq \max_{y=Y_i, Y_{N(i)}} |F_n(y) - F(y)| \\
&\leq \sup_{y\in\mathbb{R}} |F_n(y) - F(y)|.
\end{aligned}$$

Then by the fact that $0 \leq G_n(Y) \leq 1$, $0 \leq G(Y) \leq 1$,

$$\begin{aligned}
|Q_n - \widetilde{Q_n}| &\leq \frac{1}{n}\sum_{i=1}^{n} \big|\min\{F_n(Y_i), F_n(Y_{N(i)})\} - \min\{F(Y_i), F(Y_{N(i)})\}\big| \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \Big|\big(G_n(Y_i) + G(Y_i)\big)\big(G_n(Y_i) - G(Y_i)\big)\Big| \\
&\leq \sup_{y\in\mathbb{R}} |F_n(y) - F(y)| + 2\sup_{y\in\mathbb{R}} |G_n(y) - G(y)|.
\end{aligned}$$

□

Lemma 5. *There exist some constants $C > 0$ such that for any n and any $\epsilon > 0$,*

$$\mathbb{P}\big\{\big|\widetilde{Q_n} - Q\big| > \epsilon\big\} \le 2\exp\Big\{-\frac{2n\epsilon^2}{C^2}\Big\}. \tag{6}$$

Proof. Firstly we prove the lemma under the case where X has a continuous distribution, so that no randomization is involved in the definitions of π and $N(i)$'s . Suppose that for some $i \le n$, (X_i, Y_i) is replaced by a different value (X_i', Y_i'). Then there are at most three indices j such that the value of $N(j)$ changes after the replacement. Those three are the indices which correspond to $\pi(i)$, $\pi(i) - 1$ and $\pi(i) + 1$ except for the tow cases where $\pi(i) = n$ and $\pi(i) = 1$. Therefore there are at most three entries change in $\sum_{i=1}^n \min\{F(Y_i), F(Y_{N(i)})\}$ and each term changes up to 1. This shows that the first term in $\widetilde{Q_n}$ changes by at most $\widetilde{C}/n$ due to this replacement. Similarly, up to one item has changed in $\sum_{i=1}^n G^2(Y_i)$ after the replacement and each change is controlled by 1. Let $h_Q(Y_1, Y_2, \ldots, Y_n) := \widetilde{Q_n}$. Combining the above argument, we can further yield the claim that for any $1 \le i \le n$ and some constant $C > 0$,

$$\sup_{Y_i'} \big|h_Q(Y_1, Y_2, \ldots, Y_i, \ldots, Y_n) - h_Q(Y_1, Y_2, \ldots, Y_i', \ldots, Y_n)\big| \le \frac{C}{n}$$

under continuous case. By McDiarmid's inequality, for any $\epsilon > 0$, we have

$$\mathbb{P}\{|\widetilde{Q_n} - \mathbb{E}(\widetilde{Q_n})| > \epsilon\} \le 2\exp\Big\{-\frac{2n\epsilon^2}{C^2}\Big\}.$$

Now consider a more general case where there are ties among X_i's. Let $U_1, \ldots, U_n$ be i.i.d Uniform$[0, 1]$ random variables, and define $X_i^\theta = X_i + \theta U_i$, where $\theta \le \theta^*$,

$$\theta^* := \frac{1}{2}\min\{|X_i - X_j| : 1 \le i, j \le n, X_i \ne X_j\}.$$

With such an operation, we randomly break the ties between X_i's without changing the order if $X_i < X_j$. That's to say, if $X_i < X_j$, we have that $X_i^\theta < X_j^\theta$ for all $\theta < \theta^*$. Similarly, we also use definition (5) to define $\widetilde{Q_n^\theta}$ corresponding to $(X_1^\theta, Y_1), \ldots, (X_n^\theta, Y_n)$. Notice that $\widetilde{Q_n^\theta} = \widetilde{Q_n}$ for all $\theta < \theta^*$ if the randomly disordered $X_1, \ldots, X_n$ have the same arrangement as $X_1^\theta, \ldots, X_n^\theta$ after being disordered. So we can always choose the same permutation π of $X_1, \ldots, X_n$ as $X_1^\theta, \ldots, X_n^\theta$. Then we can use the conclusion in the continuous situation which we have proved in the first part, i.e.

$$\mathbb{P}\{|\widetilde{Q_n} - \mathbb{E}(\widetilde{Q_n})| > \epsilon\} = \mathbb{P}\{|\widetilde{Q_n^\theta} - \mathbb{E}(\widetilde{Q_n^\theta})| > \epsilon\} \le 2\exp\Big\{-\frac{2n\epsilon^2}{C^2}\Big\}.$$

Lemma 3 gives that $\mathbb{E}(\widetilde{Q_n}) = Q$. This completes the proof of the lemma. □

Proposition 1. *For any $\delta > 0$, there exits a constant $K > 0$ such that*

$$\mathbb{P}\{|\widehat{\omega} - \omega| > \delta\} \leq 6\exp\{-Kn\delta^2\}.$$

Proof. By the triangle inequality, it is obvious that for any $\delta > 0$, we have

$$\mathbb{P}\{|Q_n - Q| > \delta\} \leq \mathbb{P}\{|Q_n - \widetilde{Q_n}| > \frac{\delta}{2}\} + \mathbb{P}\{|\widetilde{Q_n} - Q| > \frac{\delta}{2}\}.$$

By Lemma 4, the first part in the above decomposition yields that for any $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\Big\{|Q_n - \widetilde{Q_n}| > \frac{\delta}{2}\Big\} &\leq \mathbb{P}\Big\{\sup_{y\in\mathbb{R}} |F_n(y) - F(y)| > \frac{\delta}{4}\Big\} + \mathbb{P}\Big\{2\sup_{y\in\mathbb{R}} |G_n(y) - G(y)| > \frac{\delta}{4}\Big\} \\
&\leq 2\exp\Big\{-\frac{n\delta^2}{8}\Big\} + 2\exp\Big\{-\frac{n\delta^2}{32}\Big\} \\
&\leq 4\exp\Big\{-\frac{n\delta^2}{32}\Big\}.
\end{aligned}$$

The second inequality is given by Dvoretzky–Kiefer–Wolfowitz inequality. Thus, combining the this result and Lemma 5, we have

$$\begin{aligned}
\mathbb{P}\{|Q_n - Q| > \delta\} &\leq 4\exp\left\{-\frac{n\delta^2}{32}\right\} + 2\exp\left\{-\frac{n\delta^2}{2C^2}\right\} \\
&\leq 6\exp\left\{-Kn\delta^2\right\},
\end{aligned}$$

where $K = \min\{-1/32, -1/(2C^2)\}$. Since $\widehat{\omega} = Q_n$ and $\omega = Q$, we establish the final result. □

2 Proof of Theorem 1

Theorem 1. *For any $\delta > 0$, threshold values $c > 0$ and $1/4 > \kappa > 0$, there exits a constant $K > 0$ such that*

$$\mathbb{P}\left\{\max_{1\leq j\leq p} |\widehat{\omega}_j - \omega_j| > cn^{-\kappa}\right\} \leq O\left(p\exp\left\{-Kc^2n^{1-2\kappa}\right\}\right).$$

Under Condition (2), we have that

$$\mathbb{P}\big\{\mathcal{M} \subseteq \widehat{\mathcal{M}}\big\} \geq 1 - |\mathcal{M}|O\left(\exp\left\{-Kc^2n^{1-2\kappa}\right\}\right).$$

Proof. Let $\delta = cn^{-\kappa}$. By Proposition 1, for any $\widehat{\omega}_j$ and ω_j defined by the corresponding feature X_j, $1 \leq j \leq p$, we have

$$\mathbb{P}\{|\widehat{\omega}_j - \omega_j| > cn^{-\kappa}\} \leq 6\exp\{-Kc^2n^{1-2\kappa}\}.$$

Then

$$\begin{aligned}\mathbb{P}\big\{\max_{1\le j\le p}|\widehat{\omega}_j-\omega_j|>cn^{-\kappa}\big\}&\le p\max_{1\le j\le p}\mathbb{P}\big\{|\widehat{\omega}_j-\omega_j|>cn^{-\kappa}\big\}\\&\le O(p\exp\{-Kc^2n^{1-2\kappa}\})\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\big\{\max_{j\in\mathcal{M}}|\widehat{\omega}_j-\omega_j|>cn^{-\kappa}\big\}&\le|\mathcal{M}|\max_{j\in\mathcal{M}}\mathbb{P}\big\{|\widehat{\omega}_j-\omega_j|>cn^{-\kappa}\big\}\\&\le|\mathcal{M}|O(\exp\{-Kc^2n^{1-2\kappa}\}).\end{aligned}\tag{7}$$

We know that if $\mathcal{M}\nsubseteq\widehat{\mathcal{M}}$, then there exists some $j\in\mathcal{M}$ such that $\widehat{\omega}_j<cn^{-\kappa}$. By Condition 2, we have $|\widehat{\omega}_j-\omega_j|>cn^{-\kappa}$ for some $j\in\mathcal{M}$. Hence, the event satisfies that $\big\{\mathcal{M}\nsubseteq\widehat{\mathcal{M}}\big\}\subseteq\big\{|\widehat{\omega}_j-\omega_j|>cn^{-\kappa}$, for some $j\in\mathcal{M}\big\}$, and thus $\big\{\max_{j\in\mathcal{M}}|\widehat{\omega}_j-\omega_j|\le cn^{-\kappa}\big\}\subseteq\big\{\mathcal{M}\subseteq\widehat{\mathcal{M}}\big\}$. Consequently,

$$\begin{aligned}\mathbb{P}\big\{\mathcal{M}\subseteq\widehat{\mathcal{M}}\big\}&\ge1-\mathbb{P}\big\{\max_{j\in\mathcal{M}}|\widehat{\omega}_j-\omega_j|>cn^{-\kappa}\big\}\\&\ge1-|\mathcal{M}|O(\exp\{-Kc^2n^{1-2\kappa}\}).\end{aligned}$$

□

Corollary 1 (Sure screening)**.** *Under Conditions (1) and (2), we have the sure screening property*

$$\mathbb{P}\big\{\mathcal{M}\subseteq\widehat{\mathcal{M}}\big\}\to1\quad\textit{as}\ n\to\infty.$$

Proof. Under Condition (2), the inequality (7) implies that

$$\begin{aligned}\mathbb{P}\big\{\max_{j\in\mathcal{M}}|\widehat{\omega}_j-\omega_j|>cn^{-\kappa}\big\}&\le O(p\exp\{-Kc^2n^{1-2\kappa}\})\\&=O\big(\exp\big\{n^{\xi}-Kc^2n^{1-2\kappa}\big\}\big)\\&\to0\quad\text{as}\ n\to\infty.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}\big\{\mathcal{M}\subseteq\widehat{\mathcal{M}}\big\}&\ge1-\mathbb{P}\big\{\max_{j\in\mathcal{M}}|\widehat{\omega}_j-\omega_j|>cn^{-\kappa}\big\}\\&\to1\quad\text{as}\ n\to\infty.\end{aligned}$$

□

3 Proof of Theorem 2

Theorem 2 (Ranking consistency)**.** *Under Condition (1) and Condition (3), we have that almost surely*

$$\liminf_{n\to\infty}\left(\min_{j\in\mathcal{M}}\widehat{\omega}_j-\min_{j\in\mathcal{M}^c}\widehat{\omega}_j\right)>0.$$

Proof. For any given constant c_0,

$$\begin{aligned}
\mathbb{P}\left\{\min_{j\in\mathcal{M}}\widehat{\omega}_j - \max_{j\in\mathcal{M}^c}\widehat{\omega}_j < c_0\right\} &\le \mathbb{P}\left\{(\min_{j\in\mathcal{M}}\widehat{\omega}_j - \max_{j\in\mathcal{M}^c}\widehat{\omega}_j) - (\min_{j\in\mathcal{M}}\omega_j - \max_{j\in\mathcal{M}^c}\omega_j) < -c_0\right\} \\
&\le \mathbb{P}\left\{\left|(\min_{j\in\mathcal{M}}\widehat{\omega}_j - \max_{j\in\mathcal{M}^c}\widehat{\omega}_j) - (\min_{j\in\mathcal{M}}\omega_j - \max_{j\in\mathcal{M}^c}\omega_j)\right| > c_0\right\} \\
&\le \mathbb{P}\left\{\left|\min_{j\in\mathcal{M}}\widehat{\omega}_j - \min_{j\in\mathcal{M}}\omega_j\right| + \left|\min_{j\in\mathcal{M}^c}\widehat{\omega}_j - \min_{j\in\mathcal{M}^c}\omega_j\right| > c_0\right\} \\
&\le \mathbb{P}\left\{\max_{j\in\mathcal{M}}|\widehat{\omega}_j - \omega_j| + \max_{j\in\mathcal{M}^c}|\widehat{\omega}_j - \omega_j| > c_0\right\} \\
&\le \mathbb{P}\left\{2\max_{1\le j\le p}|\widehat{\omega}_j - \omega_j| > c_0\right\} \\
&\le 6p\exp\{-Kc_0^2 n/4\}.
\end{aligned}$$

The forth inequality is given by Lemma 1. Note that

$$\sum_{n=1}^{\infty} 6p\exp\{-Kc_0^2 n/4\} = \sum_{n=1}^{\infty} O(\exp\{n^\xi - Kc_0^2 n/4\}) < \infty.$$

Therefore, by Borel-Contelli Lemma, we obtain that

$$\mathbb{P}\left(\liminf_{n\to\infty}\left\{\min_{j\in\mathcal{M}}\widehat{\omega}_j - \max_{j\in\mathcal{M}^c}\widehat{\omega}_j \ge c_0\right\}\right) = 1.$$

That is to say, almost surely

$$\liminf_{n\to\infty}\left\{\min_{j\in\mathcal{M}}\widehat{\omega}_j - \max_{j\in\mathcal{M}^c}\widehat{\omega}_j\right\} > 0.$$

□

4 Proof of Theorem 3

Theorem 3. *Under Conditions (2) and (4), there exists a constant $K > 0$ such that*

$$\mathbb{P}\{\mathcal{M}\subseteq\widehat{\mathcal{M}}_B\} \ge \prod_{l=1}^{k}\Big[1 - |\mathcal{M}|O\big(\exp\{-Kc^2 n_l n^{-2\kappa}\}\big)\Big].$$

where n_l is given by the subsample function for $l = 1, 2, \ldots k$.

Proof. Setting $\delta = cn^{-\kappa}$ in Proposition 1 and applying the similar argument as in inequality (7), we obtain

$$\begin{aligned}
\mathbb{P}\big\{\max_{j\in\mathcal{M}}|\widehat{\omega}_j^{n_l} - \omega_j| > cn^{-\kappa} \mid \mathcal{M}\subseteq\widehat{\mathcal{M}}_{l-1}\big\} &\le |\mathcal{M}|\max_{j\in\mathcal{M}}\mathbb{P}\big\{|\widehat{\omega}_j^{n_l} - \omega_j| > cn^{-\kappa}\big\} \\
&\le |\mathcal{M}|O\big(\exp\{-Kc^2 n_l n^{-2\kappa}\}\big),
\end{aligned}$$

where $\widehat{\omega}_j^{n_l}$ denotes empirical Chatterjee's variant $\widehat{\omega}_j$ using the n_l samples in the l-th round. Here, n_l is the number of samples actually used in the l-th round, whereas n appears through the choice of δ. Note that, compared with Proposition 1, the above inequality contains an additional conditional probability term $\mathcal{M} \subseteq \widehat{\mathcal{M}}_{l-1}$ while Proposition 1 is stated in an unconditional form. Indeed, Proposition 1 (equivalently Theorem 1) applies to a single screening step. By taking $\widehat{\mathcal{M}}$ as the initial set $\widehat{\mathcal{M}}_0 = 1, 2, \ldots, p$, Proposition 1 equals

$$\mathbb{P}\{|\widehat{\omega}^{n_1} - \omega| > \delta | \mathcal{M} \subseteq \widehat{\mathcal{M}}_0\} \leq 6\exp\{-Kn_1\delta^2\}.$$

By Condition (4), we have $\widehat{\omega}_j^{n_l} < cn^{-\kappa}$. Then by Condition (2), we know that $\big\{ \max_{j\in\mathcal{M}} |\widehat{\omega}_j^{n_l} - \omega_j| > cn^{-\kappa} \mid \mathcal{M} \subseteq \widehat{\mathcal{M}}_{l-1}\big\} \supseteq \big\{\mathcal{M} \nsubseteq \widehat{\mathcal{M}}_l \mid \mathcal{M} \subseteq \widehat{\mathcal{M}}_{l-1}\big\}$. For any $1 \leq l \leq k$, this leads to

$$\begin{aligned} \mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_l \mid \mathcal{M} \subseteq \widehat{\mathcal{M}}_{l-1}\} &\geq 1 - \mathbb{P}\big\{ \max_{j\in\mathcal{M}} |\widehat{\omega}_j^{n_l} - \omega_j| > cn^{-\kappa} \mid \mathcal{M} \subseteq \widehat{\mathcal{M}}_{l-1}\big\} \\ &\geq 1 - |\mathcal{M}| O\big(\exp\{-Kc^2 n_l n^{1-2\kappa}\}\big). \end{aligned} \tag{8}$$

It's obvious that $\widehat{\mathcal{M}}_l \subseteq \widehat{\mathcal{M}}_{l-1}$. Then

$$\begin{aligned} \mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_l\} &= \mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_l \text{ and } \mathcal{M} \subseteq \widehat{\mathcal{M}}_{l-1}\} \\ &= \mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_{l-1}\}\mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_l \mid \mathcal{M} \subseteq \widehat{\mathcal{M}}_{l-1}\}. \end{aligned} \tag{9}$$

Thus, applying the equality (9) recursively, we have the final conclusion that

$$\begin{aligned} \mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_B\} &= \mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_k\} \\ &= \prod_{l=1}^{k} \mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_l \mid \mathcal{M} \subseteq \widehat{\mathcal{M}}_{l-1}\} \\ &\geq \prod_{l=1}^{k} \Big[1 - |\mathcal{M}| O\big(\exp\{-Kc^2 n_l n^{-2\kappa}\}\big)\Big]. \end{aligned}$$

□

Corollary 2 (Sure screening). *Under Conditions (1), (2) and (4), we have the sure screening property*

$$\mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_B\} \to 1 \quad \textit{as } n \to \infty.$$

Proof. With Theorem 3, we show that

$$\begin{aligned}\mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_B\} &\geq \prod_{l=1}^{k}\Big[1 - |\mathcal{M}|O\big(\exp\{-Kc^2 n_l n^{-2\kappa}\}\big)\Big] \\ &= 1 - \sum_{l=1}^{k}|\mathcal{M}|O\big(\exp\{-Kc^2 n_l n^{-2\kappa}\}\big) \\ &+ \sum_{m=2}^{k}\sum_{\substack{\{l_1,\ldots,l_m\}\\ \subset\{1,\ldots,k\}}}(-1)^m \prod_{l_i\in\{l_1,\ldots,l_m\}}|\mathcal{M}|O\big(\exp\{-Kc^2 n_{l_i} n^{-2\kappa}\}\big).\end{aligned} \tag{10}$$

From the expression of subsample function, it is obvious that

$$\sqrt{n} \leq n_l = \frac{n(\alpha_{l-1}^2+1)}{\alpha_{l-1}^2\sqrt{n}+1} \leq n,\ 1 \leq l \leq k.$$

It follows from Liu et al. (2019) that the total number of total iterations for the median elimination strategy $k = O(\log(p))$, which is of order $O(n^{\xi})$ by Condition 1. Thus, we have

$$\begin{aligned}\sum_{l=1}^{k}|\mathcal{M}|O(\exp\{-Kc^2 n_l n^{-2\kappa}\}) &\leq kO(p\exp\{-Kc^2 n^{\frac{1}{2}-2\kappa}\}) \\ &= O(\exp\{n^{\xi}+\xi\log(n) - Kc^2 n^{\frac{1}{2}-2\kappa}\}) \\ &\to 0,\ \text{as } n\to\infty.\end{aligned}$$

Finally, since $\mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_B\}$ is dominated by the term $\sum_{l=1}^{k}|\mathcal{M}|O(\exp\{-Kc^2 n_l n^{-2\kappa}\})$ in Equality (10), we conclude that

$$\mathbb{P}\{\mathcal{M} \subseteq \widehat{\mathcal{M}}_B\} \to 1 \quad \text{as } n\to\infty.$$

□

5 Computational Complexity

Theorem 4. *The computational complexity of BanditCR-SIS is* $O(\sqrt{n}\log(n)p + n\log(n))$.

Proof. The first step in BanditCR-SIS has the computational complexity at the order of $O(n\log(n))$. We just need to prove that the computational complexity of the second step is $O(\sqrt{n}\log(n)p)$. We set $\eta = 1.1$ and define $n_0 = 0$. From the expression of subsample function , we have when $l = 1$, $p_0 = p$, $t_1 = n_1$, $\alpha_1 = \alpha_0$ and

$$n_1 = \frac{n(\alpha_0^2+1)}{\alpha_0^2\sqrt{n}+1};$$

when $l = 2$, then $p_1 = \lfloor (p_0 + d)/2 \rfloor$, $t_2 = n_2 - n_1$, $\alpha_2 = \alpha_0/\eta$, and

$$n_2 = \frac{n[(\alpha_0/\eta)^2 + 1]}{\sqrt{n}(\alpha_0/\eta)^2 + 1}.$$

So in the l-th round, $\alpha_l = \alpha_0/\eta^{l-1}$,

$$p_{l-1} = \left\lfloor \frac{p_{l-2} + d}{2} \right\rfloor \leq \frac{p + d(2^{l-1} - 1)}{2^{l-1}},$$

and

$$t_l = \frac{n\alpha_0^2(\eta^{2(l-1)} - \eta^{2(l-2)})(\sqrt{n} - 1)}{[\alpha_0^2\sqrt{n} + \eta^{2(l-1)}][\alpha_0^2\sqrt{n} + \eta^{2(l-2)}]},$$

and the corresponding computational complexity equals to $O(t_l \log(t_l) p_{l-1})$. The total computational complexity is

$$\begin{aligned}
\sum_{l=1}^{k} O(t_l \log(t_l) p_{l-1}) &= \sum_{l=1}^{k} O\left(t_l \log(t_l) \frac{p + d(2^{l-1} - 1)}{2^{l-1}} \right) \\
&\leq \sum_{l=1}^{k} O\left(\frac{t_l}{2^{l-1}} \log(n) p \right) + \sum_{l=1}^{k} O\left(t_l \log(n) d \right).
\end{aligned}$$

The second inequality given by the fact that $\log(t_l) \leq n_l \leq n$. Note that $\eta < \sqrt{2}$, then

$$\begin{aligned}
\sum_{l=1}^{k} \frac{t_l}{2^{l-1}} &= \frac{n(\alpha_0^2 + 1)}{\alpha_0^2\sqrt{n} + 1} + \sum_{l=2}^{k} \frac{n\alpha_0^2(\sqrt{n} - 1)(\eta^{2(l-1)} - \eta^{2(l-2)})}{2^{l-1}[\alpha_0^2\sqrt{n} + \eta^{2(l-1)}][\alpha_0^2\sqrt{n} + \eta^{2(l-2)}]} \\
&\leq \frac{n(\alpha_0^2 + 1)}{\alpha_0^2\sqrt{n} + 1} + \frac{\sqrt{n}}{\alpha_0^2} \sum_{l=2}^{k} \left[\left(\frac{\eta}{\sqrt{2}} \right)^{2(l-1)} - \frac{1}{2} \left(\frac{\eta}{\sqrt{2}} \right)^{2(l-2)} \right] \\
&\leq \sqrt{n} \left\{ 1 + \frac{1}{\alpha_0^2} \sum_{l=2}^{k} \left[\left(\frac{\eta}{\sqrt{2}} \right)^{2(l-1)} - \frac{1}{2} \left(\frac{\eta}{\sqrt{2}} \right)^{2(l-2)} \right] \right\} \\
&= O(\sqrt{n}),
\end{aligned}$$

and similarly,

$$\sum_{l=1}^{k} t_l = n_k = \frac{n(\alpha_0^2 + \eta^{2(k-1)})}{\sqrt{n}\alpha_0^2 + \eta^{2(k-1)}} = O(\sqrt{n}),$$

In each round, only the top $\lfloor (p_{l-1} + d)/2 \rfloor$ features out of p_{l-1} features will remain. It can be observed that the number of remaining features satisfy that the following inequality:

$$\lfloor (p_{l-1} + d)/2 \rfloor \leq (p_{l-1} + d)/2 \leq p/2^l + d,$$

and in a specific round m, if $p/2^m < 1$ and $p/2^{m-1} > 1$, the iterations will stop before mth round. Therefore, $k \leq m$, which is on the order of $\log(p)$. Finally, by the inequality in (5), we obtain the final result

$$\begin{aligned}\sum_{l=1}^{k} O(t_l \log(t_l) p_{l-1}) &\leq O\left(\log(n) p \sum_{l=1}^{k} \frac{t_l}{2^{l-1}}\right) + O\left(\log(n) d \sum_{l=1}^{k} t_l\right) \\ &\leq O(\sqrt{n}\log(n) p).\end{aligned}$$

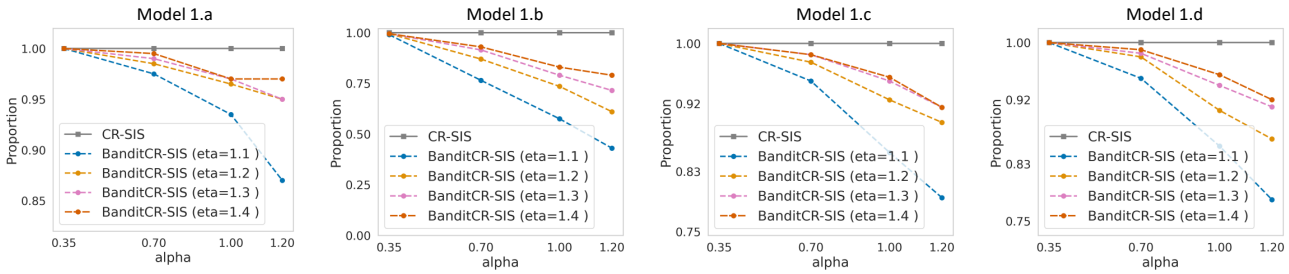
□

6 Additional Experiments

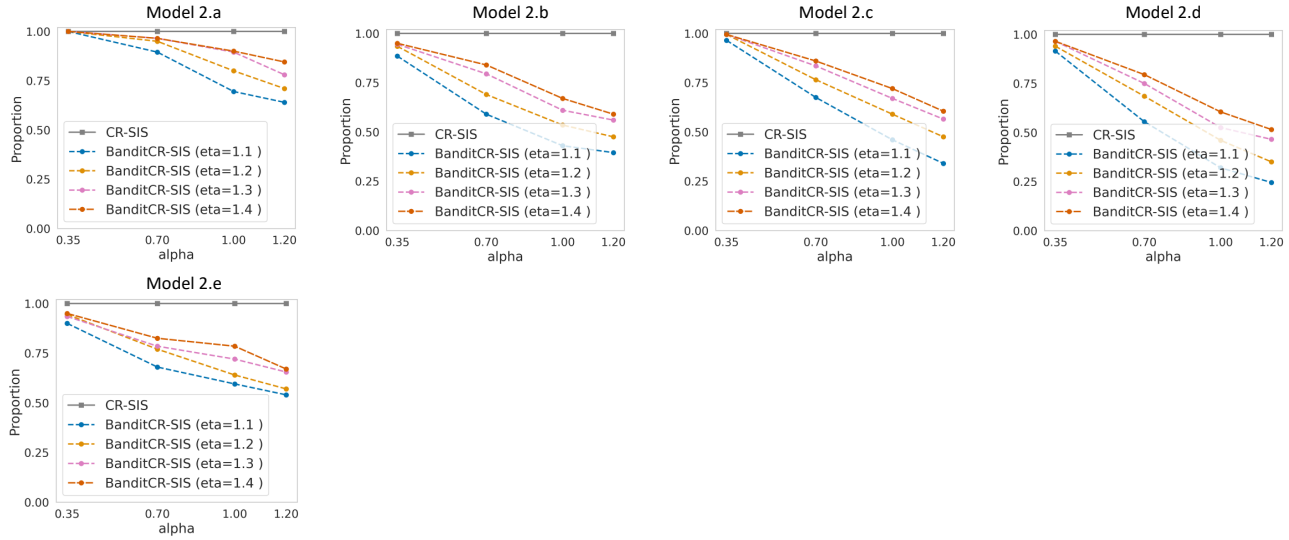
6.1 Sensitivity analysis

We conduct a sensitivity analysis with respect to another hyper-parameter, η, which determines the rate at which α is scaled down in each iteration and thus influences the number of subsamples used throughout the BanditCR-SIS procedure. We empirically examine how different choices of η affect the performance of BanditCR-SIS under various initial values of $\alpha = 0.35, 0.70, 1.00,$ and 1.20. For this analysis, we consider the linear and Poisson settings in Models 1.a–d and the nonlinear settings in Models 2.a–e, with the sample size n and dimensionality p fixed at 1500 and 2000, respectively, consistent with previous simulations. Based on both theoretical considerations of computational complexity and practical interpretability, η should lie within the range $1 < \eta < \sqrt{2}$. Accordingly, we evaluate the effect of $\eta = 1.1, 1.2, 1.3,$ and 1.4 on the screening accuracy of BanditCR-SIS under the criterion $\mathcal{P}$, that is, the proportion of correctly selected all active features for a model size $d = \lfloor n/\log(n) \rfloor$ across 200 replications, as shown in Figure 1.

From Figure 1, we observe that the performance of BanditCR-SIS with $\eta = 1.1$ is highly sensitive to the initial choice of α, exhibiting a sharp decline as α increases, corresponding to a reduction in the number of subsamples used. In contrast, the performance of BanditCR-SIS with $\eta = 1.3$ or 1.4 remains comparatively stable across all model settings. This robustness indicates that these values of η induce a more moderate and well-balanced subsample growth rate under the median elimination strategy and the associated subsampling function. From an MAB perspective, the parameter α controls the overall computational budget allocated to the algorithm, while η governs how this budget is distributed across iterations. An appropriately chosen η helps allocate computational resources in a balanced and effective



(a) Impact of the hyper-parameter η in the linear and Poisson settings (Models 1.a–d).



(b) Impact of the hyper-parameter η in the nonlinear settings (Models 2.a–e).

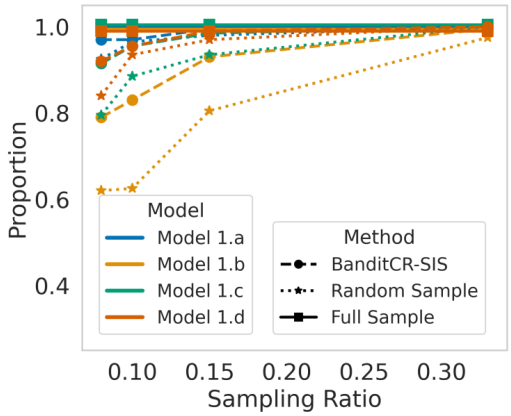
Figure 1: Impact of the hyper-parameter η on the performance of BanditCR-SIS, evaluated under the criterion $\mathcal{P}$, for different initial values of $\alpha = 0.35, 0.70, 1.00,$ and 1.20 across the linear and Poisson settings (Models 1.a–d) and the nonlinear settings (Models 2.a–e).

manner over the course of the algorithm. This prevents under-allocation of resources in the early stages, when identifying important variables is most critical, and avoids overspending computational effort in later stages, when additional resources yield diminishing returns. For simplicity and practical effectiveness, we set $\eta = 1.4$ throughout this paper, as it provides a favorable balance between screening accuracy and computational efficiency.

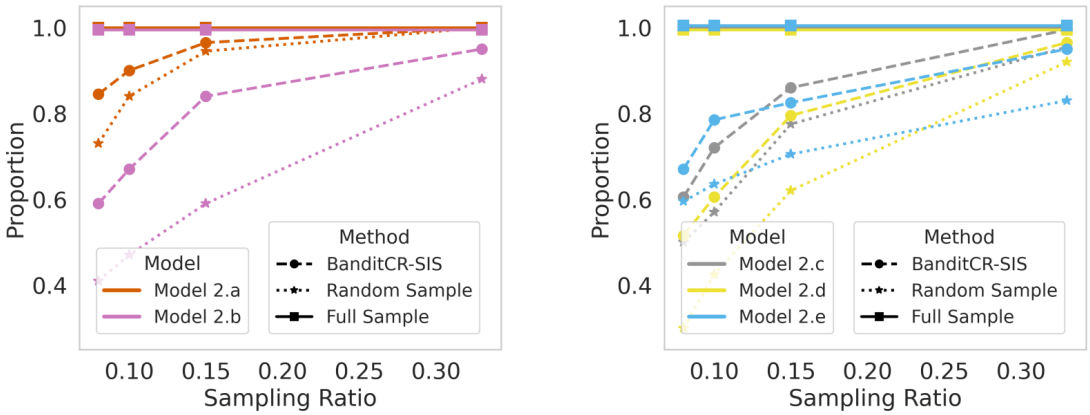
6.2 Compared with random sampling

We compare the performance of the proposed BanditCR-SIS algorithm with that of CR-SIS with equal-probability random sampling across all model settings, to more clearly evaluate the performance gains

attributable to the MAB strategy. Figure 2 illustrates the relationship between the sampling ratio (corresponding to different values of α) and the screening proportion $\mathcal{P}$ over 200 replications for the various methods and models. From left to right, the sampling ratios (0.08, 0.10, 0.15, and 0.33) correspond to $\alpha = 1.20, 1.00, 0.70$, and 0.35, respectively. The experimental configuration follows the setup described earlier, with the selected feature size $d = \lfloor n/\log(n) \rfloor$, sample size $n = 1500$, and dimensionality $p = 2000$. For reference, the results of CR-SIS using the full sample are also included in the figure.



(a) The comparison of BanditCR-SIS and random sampling in the linear and Poisson settings (Models 1.a–d).



(b) The comparison of BanditCR-SIS and random sampling in the linear and Poisson settings (Models 2.a–e).

Figure 2: Screening performance of BanditCR-SIS and CR-SIS under different sampling ratios, evaluated using the proportion $\mathcal{P}$ with model size $d = d_1$ across 200 repetitions. The four sampling ratios (ordered from left to right) correspond to BanditCR-SIS with $\alpha = 1.20, 1.00, 0.70$, and 0.35. For comparison, the results of CR-SIS applied to the full sample are plotted as horizontal reference lines, with color variations representing different model settings.

As shown in Figure 2, BanditCR-SIS consistently outperforms random sampling, particularly when the sampling ratio is small (i.e., when α is large). This performance gain is especially pronounced in more challenging scenarios, such as Models 1.b and 2.b–e, where model behavior is more sensitive to the effective sample size and therefore benefits more from a refined, strategically guided sampling mechanism under limited computational budget. Moreover, as the sampling ratio increases (i.e., as α decreases), the results of both BanditCR-SIS and random sampling converge toward those obtained using the full sample, and the performance gap between the two methods gradually narrows. The advantage of BanditCR-SIS naturally diminishes when a larger fraction of the data is utilized, since the benefit of adaptive sample allocation becomes less critical in this regime. These findings are be attributed to the fact that Chatterjee's rank correlation can effectively capture variable relationships even with relatively small subsamples in some scenarios, whereas Pearson and distance correlation generally lack such robustness. Overall, the performance improvement achieved by BanditCR-SIS highlights the strength of the proposed MAB-inspired adaptive sampling strategy, particularly in large-scale problems where computational constraints make efficient and accurate dimensionality reduction essential.

References

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association 116*(536), 2009–2022.

Liu, R., T. Wu, and B. Mozafari (2019). A bandit approach to maximum inner product search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, pp. 4376–4383.