Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Efficient Method for Large-Sample Model-Free Feature Screening via Multi-Armed Bandits

Xiaxue Ouyang
Renmin University of China
Mengyu Li
Tsinghua University
Jun Yu
Beijing Institute of Technology
Cheng Meng
chengmeng@ruc.edu.cn

Renmin University of China

Research Article

Keywords: Big data, Variable selection, Sure independence screening, Chatterjee's rank

Posted Date: April 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8766750/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Efficient Method for Large-Sample Model-Free Feature Screening via Multi-Armed Bandits

Xiaxue Ouyang[1], Mengyu Li[1], 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

Abstract

We consider the model-free feature screening in large-scale ultrahigh-dimensional data analysis. Existing feature screening methods often face substantial computational challenges when dealing with large sample sizes. To alleviate the computational burden, we propose a rank-based model-free sure independence screening method (CR-SIS) and its efficient variant, BanditCR-SIS. The CR-SIS method, based on Chatterjee's rank correlation, is as straightforward to implement as the sure independence screening (SIS) method based on Pearson correlation introduced by Fan and Lv (2008), but it is significantly more powerful in detecting nonlinear relationships between variables. Motivated by the multi-armed bandit (MAB) problem, we reformulate the feature screening procedure to significantly reduce the computational complexity of CR-SIS. For a predictor matrix of size $n \times p$, the computational cost of CR-SIS is $O(n \log(n)p)$, while BanditCR-SIS further reduces this to $O(\sqrt{n} \log(n)p + n \log(n))$. Theoretically, we establish the sure screening property for both CR-SIS and BanditCR-SIS under mild regularity conditions. Furthermore, we demonstrate the effectiveness of our methods through extensive experimental studies on both synthetic and real-world datasets. The results highlight their superior performance compared to classical screening methods, requiring significantly less computational time.

Keywords: Big data; Variable selection; Sure independence screening; Chatterjee's rank

1 Introduction

In recent years, rapid advances in science and technology have led to explosive growth in high-dimensional data, which has become increasingly common in various academic and industrial domains. Some fields have pushed this trend even further, giving rise to ultrahigh-dimensional data, particularly in areas such as bioinformatics, medical imaging, computer vision, and social network analysis (Fan and Ren, 2006; Weinberger et al., 2009; Pang et al., 2018; Ghorbani et al., 2020). In such datasets, the number of features often far exceeds the available sample size, with dimensionality reaching thousands or even millions-several magnitudes larger. For example, thousands of gene expression profiles are used for disease classification (Chen et al., 2020), while millions of single nucleotide polymorphisms (SNP) are analyzed in genome-wide association studies to explore the relationship between genotypes and phenotypes (dbGaP Study Accession: phs000368.v1.p1). While these ultrahigh-dimensional datasets offer unprecedented opportunities for discovery, they also present significant challenges for traditional statistical and machine learning algorithms. Analyzing and modeling functional relationships in the presence of redundant and noisy features can be quite challenging due to computational burden, memory cost, statistical accuracy, and algorithmic stability (Fan et al., 2009; Hall and Miller, 2009; Lv and Liu, 2014).

To address the challenges posed by ultrahigh dimensionality, Fan and Lv (2008) highlighted the importance of feature screening and introduced a widely adopted method known as sure independence screening (SIS) within the framework of the linear regression model. Feature screening aims to use some marginal utility measures to efficiently reduce the dimensionality of data, scaling it down from an extremely high dimension to a moderate size. For example, the SIS method eliminates redundant features in the linear regression model by ranking the absolute values of marginal Pearson correlations with the response and attempts to retain all the important features at the same time. This is referred to as the sure screening property, which states that the selected features will contain all the informative ones with probability approaching one when the sample size goes to infinity.

SIS has demonstrated a significant impact in the study of ultrahigh-dimensional data. However, it relies on the linear regression model, which poses substantial limitations. To address these, SIS has been extended beyond the simple linear model to more diverse model settings, including generalized linear models (Fan and Song, 2010), nonparametric regression models such as additive models (Fan et al., 2011), and varying coefficient models (Fan et al., 2014), as well as applications to the categorical response (Fan and Fan, 2008). Although these model-based methods are effective under the assumption that the specified model closely approximates the true underlying relationship, they can perform poorly when this assumption is violated, especially in cases where the relationship between variables is complex

or unknown. These limitations have prompted researchers to explore model-free approaches. Examples include sure independent ranking and screening (SIRS, Zhu et al., 2011), distance correlation-based screening (DC-SIS, Li et al., 2012), and projection correlation-based screening (PC-SIS, Xu et al., 2020), among others. In situations where the number of features is large and little prior information is available about the true model, model-free methods become particularly valuable.

Although these model-free methods have already shown great performance and enjoy strong theoretical properties, they may suffer from huge computational costs in practice. For a predictor matrix of size $n \times p$, the computational cost of SIRS is at the order of $O(n \log(n)p)$. The computational costs of the other two methods, DC-SIS and PC-SIS, are even higher, with complexities of $O(n^2p)$ and $O(n^3p)$, respectively. Such high computational costs limit the practicality of model-free methods for large-scale datasets, when both the sample size n and the number of features p are considerable, especially in the case where n has several millions.

In this paper, to alleviate the computational burden commonly encountered in model-free feature screening, we first consider a simple yet effective dependence measure, namely Chatterjee's rank correlation (Chatterjee, 2021). This recently proposed rank-based marginal utility has attracted considerable attention due to its appealing theoretical properties and practical performance. Like distance correlation, Chatterjee's rank correlation measures the dependence between the predictor and the response, taking a value of zero if and only if the variables are independent. However, applying Chatterjee's rank correlation for feature screening also comes with a computational cost of $O(n \log(n)p)$, which can be burdensome for large datasets. To further mitigate this computational challenge for a large-sample case, we propose a more efficient sampling framework based on CR-SIS. Inspired by the multi-arm bandit problem (Even-Dar et al., 2006; Bubeck et al., 2009; Audibert and Bubeck, 2010; Jamieson and Nowak, 2014), we reformulate the feature screening process as a best-arm identification task. In the standard best-arm identification problem, a multi-arm bandit machine is equipped with p arms, where pulling a given arm yields a reward drawn from an unknown corresponding distribution. One goal is to identify the arm with the highest expected reward within a specified accuracy and probability. Analogously, in our framework, the p features are treated as p arms, the marginal Chatterjee's rank correlation serves as the rewards, and the task is to identify the top d most important features (the best d arms). This approach aims to efficiently eliminate features with the smallest marginal utilities while retaining the most significant ones with an overwhelming probability. Using this bandit-inspired methodology, our proposed method significantly reduces computational overhead, making it suitable for large-scale feature screening tasks.

Our major contributions are summarized as follows. First, we introduce a simple and effective rank-based marginal utility measure for feature screening (CR-SIS) that operates without relying on any specific regression model or the distributions of the variables. To address the challenge of computational efficiency, we develop a randomized and adaptive subsampling strategy, termed as BanditCR-SIS, which significantly accelerates CR-SIS by reducing computational complexity from $O(n\log(n)p)$ to $O(\sqrt{n}\log(n)p + n\log(n))$. From a theoretical perspective, we demonstrate that both CR-SIS and BanditCR-SIS satisfy the well-known sure screening property under some mild regularity conditions. To assess the empirical performance of our methods, we conduct extensive simulation studies and a real data analysis, comparing them against two mainstream competitors. The results demonstrate that CR-SIS exhibits strong capability in identifying informative features, while BanditCR-SIS achieves comparable screening accuracy with substantially reduced computational cost. This reduction in CPU time highlights the effectiveness of its refined, strategically guided sampling mechanism, which is particularly advantageous in large-scale settings where computational resources are limited.

The remainder of this article is organized as follows. Section 2 introduces the definition of Chatterjee's rank correlation and provides a brief overview of the multi-armed bandit problem. In Section 3, we employ Chatterjee's rank correlation for feature screening and propose a more efficient version based on CR-SIS. We formulate the screening procedure as a best-arm identification problem and present the corresponding algorithm. The theoretical properties of the two methods are detailed in Section 4. Section 5 presents the performance of our proposed methods, comparing them with two existing model-free methods through various simulations and a real-data analysis. Finally, Section 6 concludes the paper, with technical proofs provided in the Supplementary Material.

2 Preliminaries

2.1 Chatterjee's Rank Correlation

Chatterjee (2021) proposed a novel rank-based correlation coefficient to measure the dependence between two random variables. Concretely, consider a pair of random variables (X, Y), where Y is not a constant. The new correlation coefficient between X and Y is defined as

$$\mathrm{CR}(X,Y) := \frac{\int \mathrm{var}(\mathbb{E}(\mathbb{I}_{\{Y\geq t\}}|X))d\mu(t)}{\int \mathrm{var}(\mathbb{I}_{\{Y\geq t\}})d\mu(t)}, \tag{1}$$

where μ represents the distribution function of Y, defined as $\mu(t) = \mathbb{P}(Y \leq t)$. We can note that the value of $\mathrm{CR}(X, Y)$ lies in the interval $[0, 1]$.

Chatterjee (2021) also provided the corresponding empirical estimator. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be n i.i.d. pairs with the same distribution as (X, Y). First, rearrange the data as $(x_{(1)}, y_{(1)}), \ldots, (x_{(n)}, y_{(n)})$ so that $x_{(1)} \leq \ldots \leq x_{(n)}$. If the x_i's do not have ties, this rearrangement is unique; otherwise, the ties are broken uniformly at random to produce a strictly increasing order. Let r_i denote the rank of $y_{(i)}$, defined as the number of k such that $y_{(k)} \leq y_{(i)}$. Additionally, define l_i as the number of k such that $y_{(k)} \geq y_{(i)}$. The empirical estimator of Chatterjee's rank correlation is given by

$$\widehat{\mathrm{CR}}(X, Y) := 1 - \frac{n \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{2 \sum_{i=1}^{n} l_i (n - l_i)}. \tag{2}$$

Chatterjee's rank correlation offers a simple formulation as the most widely used statistical measure currently and excels in detecting the strength of relationships between variables, even when the relationships are not monotonic or in the presence of noise. In the following, we provide a brief overview of its properties (Chatterjee, 2021).

(1) The value of $\mathrm{CR}(X, Y)$ belongs to the interval $[0, 1]$ since $\mathrm{var}(\mathbb{I}_{\{Y \geq t\}}) \geq \mathrm{var}(\mathbb{E}(\mathbb{I}_{\{Y \geq t\}} | X))$ for any t. If X and Y are independent, then $\mathbb{E}(\mathbb{I}_{\{Y \geq t\}} | X) = \mathbb{E}(\mathbb{I}_{\{Y \geq t\}})$, which is a constant. In this case, $\mathrm{CR}(X, Y) = 0$. Conversely, if Y is a measurable function of X, we have $\mathbb{E}(\mathbb{I}_{\{Y \geq t\}} | X) = \mathbb{I}_{\{Y \geq t\}}$, which leads to $\mathrm{CR}(X, Y) = 1$.

(2) By the definition of $\mathrm{CR}(X, Y)$, it is not symmetric with respect to X and Y. This asymmetry is intentional, as the primary objective is to assess whether Y is a measurable function of X, rather than simply to determine whether one variable is a function of the other. However, if symmetry is required, one can consider the maximum of $\mathrm{CR}(X, Y)$ and $\mathrm{CR}(Y, X)$. This combined measure equals zero if and only if X and Y are independent and equals one if and only if at least one of X or Y is a measurable function of the other.

(3) If Y is not almost surely a constant, then $\widehat{\mathrm{CR}}(X, Y)$ converges to $\mathrm{CR}(X, Y)$ almost surely when n goes to infinity. The convergence of $\widehat{\mathrm{CR}}(X, Y)$ does not depend on the specific distribution of (X, Y), apart from the requirement that Y is not a constant. This correlation measure remains robust and effective regardless of whether X and Y are discrete, continuous, light-tailed, or heavy-tailed.

(4) The coefficient $\widehat{\mathrm{CR}}(X, Y)$ is invariant under strictly increasing transformations of X and Y because it is rank-based. It is robust and applicable to data that exhibit oscillatory patterns.

(5) The computational complexity of the coefficient $\widehat{\mathrm{CR}}(X, Y)$ is $O(n \log(n))$.

2.2 Multi-Arm Bandit Problem

The multi-arm bandit (MAB) problem is a classic framework for sequential decision-making. It captures the challenge of balancing exploration (gathering information about the environment to discover potentially profitable actions) and exploitation (maximizing immediate rewards by choosing the empirically best action based on current knowledge). The MAB problem has broad application in areas such as clinical trials (Durand et al., 2018), recommender systems in temporal settings (Mary et al., 2015), and anomaly detection (Ding et al., 2019).

Consider an MAB with p arms, each associated with an unknown distribution of rewards. When an arm is pulled, it generates a reward that is independently and randomly sampled from its corresponding distribution. As discussed in Chen et al. (2014) and Kaufmann et al. (2016), the MAB problem typically focuses on one of two distinct objectives: either maximizing the cumulative reward over a limited number of pulls or determining the arm with the highest expected reward (the best arm) with high confidence using the fewest possible pulls required. This paper is concerned with the latter goal, specifically focusing on the best arm identification problem, which generally has two different settings.

- *Fixed budget setting*. In the fixed budget setting, the total number of pulls allocated is predetermined. The goal is to identify the best arm within this limited budget while maximizing the confidence level δ. Algorithms such as Successive Halving (Karnin et al., 2013; Jamieson and Talwalkar, 2016) and Successive Rejects (Audibert and Bubeck, 2010) operate by dynamically allocating the budget across arms. They progressively eliminate suboptimal arms round by round until only one arm remains. These methods are designed to fully utilize the available budget to ensure that the selected arm approximates the true best arm as closely as possible.

- *Fixed confidence setting*. The fixed confidence setting aims to devise a sampling procedure that, for a given confidence level δ, identifies the best arm with a probability of at least $1 - \delta$, regardless of the reward distributions. In this setting, an algorithm must dynamically adapt its sampling strategy to account for cases where the reward expectations of the best and second-best arms are arbitrarily close, ensuring reliable identification under challenging situations. These algorithms include Successive Elimination (Even-Dar et al., 2006), Exponential Gap Elimination (Karnin et al., 2013), Lower Upper Confidence Bound algorithms (Kalyanakrishnan et al., 2012).

Motivated by the MAB problem, we treat the candidate features as the arms and the informative features as the best arms. Each pull of an arm corresponds to randomly drawing subsamples associated with a specific feature to compute its correlation coefficient. The reward returned from this interaction is

represented by the computed correlation coefficient based on the subsamples drawn currently. Using the intermediate results observed so far, we must dynamically determine how many additional calculations to allocate to each feature to refine the correlation estimates. The second setting aligns with our objective: to minimize sample complexity compared to the original screening procedure while ensuring that the returned results are within a pre-specified proximity of the optimal set. In this way, we can accelerate the original screening method. In the following section, we provide a detailed description of our proposed method.

3 Method

In this section, we first integrate a modified Chatterjee's rank dependency into feature screening. This rank-based screening procedure is fully model-free, as it allows for an arbitrary regression relationship between the response and the features, whether linear or nonlinear. To address computational time limitations, we accelerate this method by drawing inspiration from the MAB problem. As a result, we successfully reduce the computational cost from $O(n\log(n)p)$ to $O(\sqrt{n}\log(n)p + n\log(n))$, while preserving the well-known sure screening property.

3.1 Chatterjee's Rank-Based Feature Screening

Let $Y = (y_1, y_2, \ldots, y_n)^\top$ be the response vector and $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top = (X_1, X_2, \ldots, X_p)$ be an $n \times p$ random design matrix, where $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$. We focus on the ultrahigh-dimensional setting where the dimensionality p greatly exceeds the sample size n. In such cases, it is natural to assume that only a small subset of features contributes to the response. Following the notations in the literature by Li et al. (2012) and Zhu et al. (2011), without specifying a regression model, we define the index set of the active features as

$$\mathcal{M} := \{j : F(Y|\boldsymbol{X}) \text{ functionally depends on } X_j, 1 \le j \le p\}, \tag{3}$$

where $F(Y|\boldsymbol{X})$ is the conditional distribution function of Y given $\boldsymbol{X}$. This means that when $j \in \mathcal{M}^c$, X_j is statistically independent of Y given all active features $X_{\mathcal{M}}$. This setup abstracts a large number of sparse regression problems, including both linear and nonlinear models. For example, Li et al. (2012) considered a case where $F(Y|\boldsymbol{X})$ depends on $\boldsymbol{X}$ only through $\boldsymbol{\beta}^\top \boldsymbol{X}_{\mathcal{M}}$, that is, $F(Y|\boldsymbol{X}) = F_0(Y|\boldsymbol{\beta}^\top \boldsymbol{X}_{\mathcal{M}})$, where $F_0(\cdot|\boldsymbol{\beta}^\top \boldsymbol{X}_{\mathcal{M}})$ is an unknown distribution function for a given $\boldsymbol{\beta}^\top \boldsymbol{X}_{\mathcal{M}}$. This framework is common in a large variety of parametric and semiparametric models where the response

Y depends on the predictors $\boldsymbol{X}$ through linear combinations of the active set $\boldsymbol{X}_{\mathcal{M}}$. Examples of such models include the partially linear model (Härdle et al., 2000), the proportional hazards model (Cox, 1972), and the generalized partially linear single-index model (Carroll et al., 1997). Now we gives the general model framework illustrate by (3) as follows,

$$F(Y|\boldsymbol{X}) = g(\boldsymbol{X}_{\mathcal{M}}),$$

where $g(\cdot)$ is an unknown measurable function.

Recall that the Chatterjee's rank correlation $\mathrm{CR}(X_j, Y)$ equals zero if and only if X_j and Y are independent and equals one if and only if Y is a measurable function of X_j. This motivates us to screen out features X_j with relatively small values of $\mathrm{CR}(X_j, Y)$, which selects those with large $\mathrm{CR}(X_j, Y)$ as active features. We observe that the denominator of the correlation $\mathrm{CR}(X_j, Y)$ does not depend on the predictor X_j. Consequently, ranking each feature by this correlation is equivalent to ranking the corresponding numerator:

$$\omega_j = \int \mathrm{var}(\mathbb{E}(\mathbb{I}_{\{Y \geq t\}}|X_j))d\mu(t).$$

Chatterjee (2021) derived a rank-based consistent estimator of ω_j. To be precise, first rearrange the data pairs $(x_{1j}, y_1), \ldots, (x_{nj}, y_n)$ into $(x_{(1)j}, y_{(1)}), \ldots, (x_{(n)j}, y_{(n)})$ such that $x_{(1)j} \leq \ldots \leq x_{(n)j}$. The sample estimator of ω_j is given by

$$\widehat{\omega}_j = \frac{1}{n^3}\sum_{i=1}^{n} l_i(n - l_i) - \frac{1}{2n^2}\sum_{i=1}^{n-1}|r_{i+1} - r_i| + \frac{r_n - r_1}{2n^2}, \tag{4}$$

where $r_i = \#\{1 \leq k \leq n : y_{(k)} \leq y_{(i)}\}$, and $l_i = \#\{1 \leq k \leq n : y_{(k)} \geq y_{(i)}\}$. This modified correlation serves as an adaptation of Chatterjee's rank correlation estimator, as defined in Equation (2), optimized for theoretical simplicity when screening. Empirically, we propose to estimate the active set $\mathcal{M}$ as

$$\widehat{\mathcal{M}} := \{j : \widehat{\omega}_j \geq cn^{-\kappa}, 1 \leq j \leq p\},$$

where c and κ are prespecified threshold values, to be defined in Condition (2) in the subsequent Section 4. This screening procedure is called Chatterjee's rank-based screening (CR-SIS), summarized in Algorithm 1.

Algorithm 1 CR-SIS

Input: Data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, threshold values c, κ

for $j = 1 : p$ **do**

  Sort the data pairs $\{(x_{ij}, y_i)\}_{i=1}^n$ into $\{(x_{(i)j}, y_{(i)}\}_{i=1}^n$ in ascending order of $\{x_{ij}\}_{i=1}^n$

  Calculate $r_i = \#\{1 \le k \le n : y_{(k)} \le y_{(i)}\}$, $l_i = \#\{1 \le k \le n : y_{(k)} \ge y_{(i)}\}$

  Calculate $\widehat{\omega}_j$ using Equality (4)

end for

$\widehat{\mathcal{M}} = \left\{j : \widehat{\omega}_j \ge cn^{-\kappa}, 1 \le j \le p\right\}$

Output: $\widehat{\mathcal{M}}$

The term $cn^{-\kappa}$ serves as a benchmark to separate the active predictors from inactive ones. It corresponds to the soft thresholding rule widely used in the marginal screening literature (Li et al., 2012; Shao and Zhang, 2014; Wang et al., 2023). Alternatively, the hard thresholding rule proposed by Fan and Lv (2008) can be applied to estimate the active set $\mathcal{M}$ as follows:

$$\left\{j : \widehat{\omega}_j \text{ is among the first } d \text{ largest values of all}, 1 \le j \le p\right\}$$

where d is the prespecified number of selected features. Empirically, d is often chosen as $\lfloor n/\log(n) \rfloor$.

3.2 A More Efficient Screening Method

We now present a more efficient screening method for the large-sample CR-SIS method. Just as existing MAB algorithms do, our method iteratively discards non-informative features and samples batches of observations to update current correlation coefficients. Instead of computing the correlation coefficient for each feature X_j using all available observations, our method estimates the coefficients using a subset of subsamples. Since a larger sample size gives higher estimation accuracy, our method adaptively allocates more samples to the more promising features, allowing it to better identify the most informative features.

Denote the number of remaining features in the lth round as p_l and the size of the selected model as d. We define a shrinking parameter $0 < \delta_l < 1$ such that $p_l = \lfloor \delta_l p_{l-1} \rfloor$, which controls the number of features left in each round. To determine the value of δ_l, we apply the classic median elimination strategy in the best arm identification problem (Even-Dar et al., 2002; Liu et al., 2019). The basic idea behind the median elimination strategy is to discard the least performing half of the candidates in each iteration. By quickly eliminating poorly performing candidates, the algorithm focuses more resources on

the promising ones, thus avoiding unnecessary exploration and improving overall efficiency. Adapted to our setting, we extend this strategy to obtain the top d features with the largest correlation coefficients, rather than just the best one. Specifically, at the end of the lth iteration, we remove the $\lceil (p_{l-1} - d)/2 \rceil$ features with the lowest correlation values $\widehat{\omega}_j$, retaining the top $\lfloor (p_{l-1} + d)/2 \rfloor$ features in the current set. Equivalently, this can be expressed as the setting $\delta_l = (p_{l-1} + d)/(2p_{l-1})$.

Next, we discuss the process of determining the number of selected subsamples n_l in the lth round. Our objective is to develop an efficient variant of the CR-SIS method that maintains the sure screening property while incorporating the median elimination strategy. To achieve this, we propose a subsample function, defined as follows:

$$t(n, \alpha) := \frac{n(\alpha^2 + 1)}{\alpha^2 \sqrt{n} + 1}. \tag{5}$$

Denote $\alpha_0 = \alpha$. In the lth round, the total number of selected subsamples n_l is determined by the subsample function $t(n, \alpha_{l-1})$, where $\alpha_l = \alpha_0/\eta^{l-1}$ for $l = 1, 2, \ldots, k$, with k representing the total number of iterations. The hyper-parameter α balances between screening efficiency and accuracy, while η regulates the rate at which α is scaled down in each round. Based on numerical experiments, we empirically set $\eta = 1.4$, achieving an appropriate growth rate for subsamples under the median elimination strategy and the chosen subsample function. A more detailed discussion of the selection of η is provided in the Supplementary Material. This approach not only provides insights into the theoretical behavior of our method, but also offers a flexible mechanism for balancing error and computational efficiency in practice. Specifically, it helps us understand how the accuracy of the solution improves as more computational resources are allocated. This approach aligns with the fixed confidence setting. The proposed functional form $t(n, \alpha)$ offers the following properties: (1) it is compatible with the median elimination strategy, (2) it preserves the sure screening property, and (3) it significantly lowers computational complexity compared to the original screening procedure. These properties facilitate both the practical execution of our algorithm and the theoretical analysis. We will demonstrate these results in the next section.

Intuitively, the efficient screening algorithm, referred to as BanditCR-SIS, can be summarized in the following steps:

- We start by initializing $\widehat{\mathcal{M}}_0$ as the full model $\{1, 2, \ldots, p\}$, and setting the initial model size $p_0 = p$.
- First, we compute the Chatterjee's rank correlation coefficients using n_1 randomly selected observations for each feature X_j, where $j \in \widehat{\mathcal{M}}_0$, and obtain a reduced submodel $\widehat{\mathcal{M}}_1$ with size $p_1 = \lfloor \delta_1 p_0 \rfloor$.
- We then apply a similar procedure to the model $\widehat{\mathcal{M}}_1$ using n_2 randomly selected observations,

resulting in a submodel $\widehat{\mathcal{M}}_2$ with size $p_2 = \lfloor \delta_2 p_1 \rfloor$.

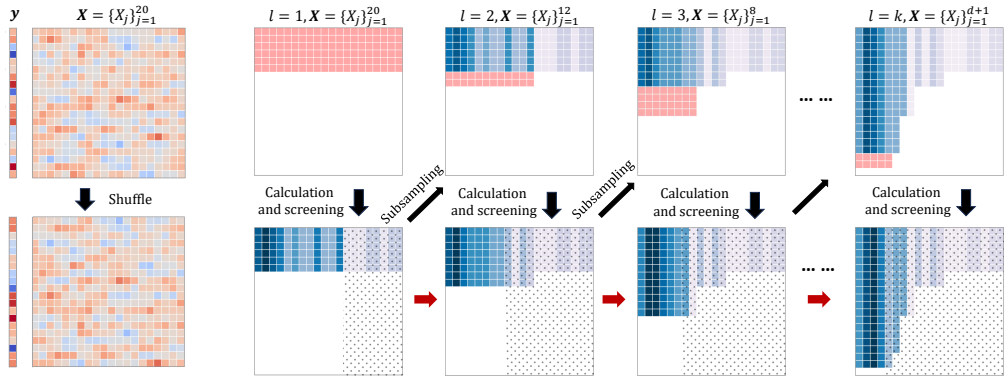
- These process is repeated iteratively until the kth round, when the model size reaches the target size d, i.e., $|\widehat{\mathcal{M}}_{k-1}| = p_{k-1} > d$, and $|\widehat{\mathcal{M}}_k| = p_k = d$.

Each n_l is determined by the subsample function in Equation (5). For practical implementation, rather than randomly selecting subsamples in each round, we begin by shuffling the entire dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, which includes the response and the corresponding predictors. Subsamples are then sequentially selected in each round from this shuffled dataset. The BanditCR-SIS procedure is outlined in Algorithm 2 and visually represented in Figure 1.

Algorithm 2 BanditCR-SIS

Input: Data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, model size d, hyper-parameter α
Initialize: $\widehat{\mathcal{M}}_B = \{1, 2, \ldots, p\}$, $n_0 = 0$, $l = 0$.
Randomly shuffle the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$
while $|\widehat{\mathcal{M}}_B| > d$ **do**
    Calculate $n_{l+1} = t(n, \alpha)$ and select $n_{l+1} - n_l$ subsamples sequentially from feature X_j, where $j \in \widehat{\mathcal{M}}_B$
    Update the correlation coefficient $\widehat{\omega}_j$ with the selected subsamples
    $\widehat{\mathcal{M}}_B \leftarrow \left\{ j \in \widehat{\mathcal{M}}_B : \hat{\omega}_j \text{ is among the first } \left\lfloor \frac{|\widehat{\mathcal{M}}_B| + d}{2} \right\rfloor \right.$ $\left.\text{largest values of all} \right\}$
    $l \leftarrow l + 1$; $\alpha \leftarrow \alpha / 1.4$
end while
Output: $\widehat{\mathcal{M}}_B$

The overall computational complexity of BanditCR-SIS is given by $O(\sqrt{n}\log(n)p + n\log(n))$. As illustrated in Figure 1, the complexity of the first step, which involves shuffling the full sample (including the response and corresponding predictor features), has a complexity of $O(n \log n)$. The second step alternates between discarding features and updating the correlation coefficients based on the selected subsamples in each iteration, with a complexity of $O(\sqrt{n}\log(n)p)$. Compared to the original CR-SIS method, BanditCR-SIS optimizes computational efficiency by iteratively eliminating less informative features, focusing calculations on more relevant coefficients, and avoiding redundant operations. In addition, although our algorithm alternates between discarding features and updating coefficients, we demonstrate that, with the median elimination strategy, the overall complexity remains independent of



(a) Random shuffle. (b) Main algorithm of BanditCR-SIS.

Figure 1: An illustration of the proposed BanditCR-SIS algorithm. (a) Randomly shuffle the entire dataset, including the response the corresponding predictors. (b) The light red cells indicate the subsamples (rows) chosen for the current iteration. Within the predictor matrix, selected entries are marked in blue, with deeper shades of blue signifying higher values of Chatterjee's rank correlation. The shaded areas denote predictor features (columns) that have been excluded in both the current and earlier rounds. The proposed algorithm alternates between discarding features with low Chatterjee's rank correlation and updating the correlation with selected observations in each round.

the total number of iterations k, which is of the order of $\log(p)$. Detailed proof of this computational complexity is available in the Supplementary Material.

4 Theoretical Analysis

We now examine the theoretical properties of the proposed methods. Under mild regularity conditions, we demonstrate that the CR-SIS method possesses both the sure screening property (Fan and Lv, 2008; Li et al., 2012) and the consistency in ranking property (Zhu et al., 2011). The sure screening property ensures that, with asymptotic probability one, all active features will be included in the selected feature set. The ranking consistency property guarantees that the method selects all active predictors and excludes all inactive ones with high probability. Furthermore, the proposed efficient version preserves the sure screening property within our framework, yielding favorable results. To facilitate the technical proofs, we impose the following conditions, although they may not be the weakest possible.

Condition.

(1) $p > n$ *and* $\log(p) = O(n^{\xi})$ *for* $\xi \in (0, 1/2 - 2\kappa)$, *where* κ *is given by Condition (2).*

(2) The minimum ω_j of active features satisfies

$$\min_{j\in\mathcal{M}} \omega_j \geq 2cn^{-\kappa}$$

for some constants $c > 0$ and $\kappa > 0$.

(3) For some constant $c_0 > 0$,

$$\liminf_{p\to\infty} \left\{ \min_{j\in\mathcal{M}} \omega_j - \max_{j\in\mathcal{M}^c} \omega_j \right\} \geq 2c_0.$$

(4) In each round l,

$$\left\{ j : \widehat{\omega}_j^{n_l} \geq cn^{-\kappa}, j \in \widehat{\mathcal{M}}_{l-1} \right\} \subseteq \widehat{\mathcal{M}}_l,$$

where $\widehat{\omega}_j^{n_l}$ denotes the correlation coefficient $\widehat{\omega}_j$ calculated using n_l samples.

These conditions are milder than those for the SIS (Fan and Lv, 2008) and DC-SIS (Li et al., 2012), because we do not require the regression function of Y onto X to be linear and it needs little assumption on the moments of predictors. Condition (1) corresponds to Condition 1 of Fan and Lv (2008), which indicates that we focus on the NP-dimensionality case where the dimension p grows exponentially with the sample size n. Condition (2) is commonly encountered in the marginal screening literature (Fan and Lv, 2008; Li et al., 2012). Intuitively, it requires that the dependency between the active predictors and the response dose not decay too quickly to zero as n increases. Condition (3) parallels Condition (C3) in Cui et al. (2015) and Condition (B2) in Li et al. (2023), which ensures a clear signal separation between the active and inactive features. This condition is an additional assumption for independence screening and is typically imposed to ensure ranking consistency, as studied in Zhu et al. (2011), Cui et al. (2015), and Li et al. (2023). Condition (4) is typically introduced for the technical analysis of the BanditCR-SIS method. It implies that in each round, the number of retained features, p_l, is larger than what would be selected using the soft thresholding rule. For example, in the final round k, the model size $d = |\widehat{\mathcal{M}}_k| = |\widehat{\mathcal{M}}_B| > |\widehat{\mathcal{M}}|$.

Theorem 1 below provides an explicit lower bound for the sure screening probability of the proposed CR-SIS method. Allowing the dimensionality p to grow at an exponential rate, all active predictors can be included in the reduced model $\widehat{\mathcal{M}}$ with the probability approaching one exponentially fast as n approaches infinity, as established in Corollary 1.

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

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

Under Condition (2), we have that

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

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

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

Another interesting property for independence screening is ranking consistency property (Zhu et al., 2011). Theorem 2 below demonstrates a stronger theoretical result than sure screening property. It tells us that, almost surely, the sample correlation values $\widehat{\omega}_j$ rank active predictors above the inactive ones with high probability, guaranteeing a clear separation between active and inactive predictors.

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

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

We now provide our main theorem that the proposed efficient version of CR-SIS method preserves the sure screening property under our settings.

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

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

where n_l is given by the subsample function (5) for $l = 1, 2, \ldots k$.

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

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

Theorem 3 highlights two key points. Firstly, the probability that the selected set $\widehat{\mathcal{M}}_B$ contains all active features is closely related to the choice of n_l. Clearly, a larger value of n_l leads to a higher screening accuracy. When n_l is chosen to be the full sample size n, the BanditCR-SIS method reduces to the original screening method, and so does the screening accuracy converge towards that of the original method as n_l increases to n. Recall that n_l is determined by the tuning parameter α. Thus, by varying α, we can control the accuracy of the results, depending on the available computational budget. Secondly, the overall probability is the product of the probabilities that the active features set $\mathcal{M}$ is contained in $\mathcal{M}_l$ during each round. In particular, the median elimination strategy implies that $k = O(\log(p)) = O(n^{\xi})$, leading to k approaching infinity as n diverges. Corollary 2 shows that the increase in k is negligible compared to the exponential convergence of the total probability towards one, since n_l also increases with n. A more detailed discussion is provided in the Supplementary Material.

5 Numerical Studies

In this section, we evaluate the empirical performance and computational time of the proposed CR-SIS and BanditCR-SIS methods through Monte Carlo simulations. The simulation studies were implemented using Python. We further demonstrate the effectiveness of these screening procedures with an empirical analysis of a real-world dataset. The performance of the proposed methods is compared with that of classic sure independence screening (Fan and Lv, 2008) and distance correlation-based screening (Li et al., 2012). Overall, the results consistently demonstrate that CR-SIS exhibits superior capability in identifying informative features. Moreover, BanditCR-SIS effectively preserves this strong screening performance, making it particularly advantageous in large-scale practical applications where computational resources are limited.

5.1 Simulations

We generate the predictors $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ from a multivariate distribution with zero mean and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. The i.i.d. error term ϵ is drawn from a specified distribution. For each example, we set the sample size to $n = 1500$ and the number of predictors to $p = 2000$. The number of true active predictors is denoted by dt. Each experiment is replicated 200 times, and the performance is evaluated using the following two criteria:

1. $\mathcal{S}$: the 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size that includes all active predictors.

2. $\mathcal{P}$: the proportion that all active predictors are selected for a given model size d_i over 200 replicates, where $d_1 = \lfloor n/\log(n) \rfloor$, $d_2 = 2\lfloor n/\log(n) \rfloor$, $d_3 = 3\lfloor n/\log(n) \rfloor$.

A method performs better when its results frequently align with or closely approximate the actual number of truly active predictors dt according to the criteria $\mathcal{S}$. Similarly, according to criteria $\mathcal{P}$, superior performance is indicated by values equal to or close to 100%, reflecting a very high accuracy in both ranking and selection.

5.1.1 Linear and Poisson Models

Let Y be the response variable and X_j be the jth feature of $\boldsymbol{X}$. Following Liu et al. (2022) and Li et al. (2023), we consider the following data generation settings.

- **Model 1.a:** $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \epsilon$, where $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\epsilon \sim N(0, 1)$.

- **Model 1.b:** $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \epsilon$, where $\boldsymbol{X} \sim t_1(\mathbf{0}, \boldsymbol{\Sigma})$ and $\epsilon \sim t_1(0)$.
- **Model 1.c:** $Y = \exp\{2(X_1 + X_2 + X_3 + X_4 + X_5)\} + \epsilon$, where $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\epsilon \sim N(0, 1)$.
- **Model 1.d:** $Y \sim$ Poisson$(\exp\{2(X_1 + X_2 + X_3 + X_4 + X_5) + \epsilon\})$, where $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.

The number of active features in these models, dt, is set to 5. Model 1.a and Model 1.b set a linear model, with Model 1.b addressing a heavy-tailed scenario where both $\boldsymbol{X}$ and ϵ follow a t-distribution with zero mean and one degree of freedom. Model 1.c considers the continuous response, while the response in Model 1.d is discrete under Poisson regression.

Table 1 presents the ranking criterion $\mathcal{S}$ and the selection proportion criterion $\mathcal{P}$. For the simple linear regression model, where both the predictors and the error follow a normal distribution, all methods demonstrate strong performance. However, in scenarios with heavy-tailed predictors and errors, CR-SIS significantly outperforms other methods. This superior performance can be attributed to Chatterjee's rank correlation, which only relies on the ranks of the observed response values, making it more robust in such cases. In the Poisson Models 1.c and 1.d, CR-SIS effectively recovers the true active set with a model size close to 5, whereas SIS and DC-SIS need much larger model sizes to recover the active set in most of the time. Another noteworthy observation is that BanditCR-SIS achieves results comparable to CR-SIS, even as the tuning parameter α increases. This indicates that the proposed BanditCR-SIS method is a highly efficient variant of CR-SIS, achieving strong performance while utilizing fewer samples. This is because BanditCR-SIS adaptively adjusts the structure of the screening procedure and substantially reduces unnecessary computations related to non-informative features.

5.1.2 Nonlinear Models

Consider the following four nonlinear data-generating models:

- **Model 2.a:** $Y = 5X_1 + 2\sin(\pi X_2/2) + 2X_3\mathbb{I}_{\{X_3>0\}} + 2\exp\{5X_4\} + \epsilon$;
- **Model 2.b:** $Y = 2X_2^{-2} + 4X_2^3 + 3\cos(X_3) + 10\mathbb{I}_{\{X_4>0\}} + \epsilon$;
- **Model 2.c:** $Y = 1 - 5(X_2 + X_3)^3\exp\{-5(X_1 + X_4^3)\} + \epsilon$;
- **Model 2.d:** $Y = 1 - 5(X_2 + X_3)^{-3}\exp\{1 + 10\sin(\pi X_1/2) + 5X_4\} + \epsilon$;
- **Model 2.e:** $Y = \cos(X_1) + X_2 + \exp\{2X_2 + 2X_3 + \epsilon\}$;

where $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\epsilon \sim N(0, 1)$.

In Models 2, we examine nonlinear scenarios with the true model size set to $dt = 4$. Models 2.a and 2.b are based on simple additive structures. Models 2.c and 2.d incorporate more complex and

Table 1: The quantiles of minimum model size $\mathcal{S}$ and the proportion $\mathcal{P}$ with $d_1 = 205, d_2 = 410, d_3 = 615$ for Models 1 over 200 replications. The hyper-parameters in BanditCR-SIS are set to 0.35, 0.70, 1.00 and 1.20.

				Model 1.a				
	5%	25%	50%	75%	95%	$\mathcal{P}_{d_1}(\%)$	$\mathcal{P}_{d_2}(\%)$	$\mathcal{P}_{d_3}(\%)$
SIS	5.0	5.0	5.0	5.0	5.0	100.0	100.0	100.0
DC-SIS	5.0	5.0	5.0	5.0	5.0	100.0	100.0	100.0
CR-SIS	5.0	5.0	5.0	5.0	5.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.35$)	5.0	5.0	5.0	5.0	5.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.70$)	5.0	5.0	5.0	5.0	5.0	99.5	100.0	100.0
BanditCR-SIS ($\alpha = 1.00$)	5.0	5.0	5.0	5.0	5.0	97.0	99.0	100.0
BanditCR-SIS ($\alpha = 1.20$)	5.0	5.0	5.0	5.0	5.0	97.0	97.5	98.5
				Model 1.b				
	5%	25%	50%	75%	95%	$\mathcal{P}_{d_1}(\%)$	$\mathcal{P}_{d_2}(\%)$	$\mathcal{P}_{d_3}(\%)$
SIS	185.8	927.5	1443.0	1841.3	1985.1	0.0	5.0	8.5
DC-SIS	5.0	17.8	498.5	1400.0	1876.6	40.0	46.5	54.5
CR-SIS	5.0	5.0	5.0	5.0	5.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.35$)	5.0	5.0	5.0	5.0	5.0	99.5	99.5	99.5
BanditCR-SIS ($\alpha = 0.70$)	5.0	5.0	5.0	5.0	1003.0	89.0	91.5	93.5
BanditCR-SIS ($\alpha = 1.00$)	5.0	5.0	5.0	36.3	1003.0	84.0	88.5	91.0
BanditCR-SIS ($\alpha = 1.20$)	5.0	5.0	5.0	317.3	1003.0	78.0	84.5	88.0
				Model 1.c				
	5%	25%	50%	75%	95%	$\mathcal{P}_{d_1}(\%)$	$\mathcal{P}_{d_2}(\%)$	$\mathcal{P}_{d_3}(\%)$
SIS	5.0	17.8	68.5	175.5	472.3	76.5	90.0	96.0
DC-SIS	5.0	5.0	7.5	38.0	377.1	89.0	95.5	98.5
CR-SIS	5.0	5.0	5.0	5.0	5.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.35$)	5.0	5.0	5.0	5.0	5.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.70$)	5.0	5.0	5.0	5.0	5.0	98.5	99.5	99.5
BanditCR-SIS ($\alpha = 1.00$)	5.0	5.0	5.0	5.0	504.0	95.5	98.0	98.5
BanditCR-SIS ($\alpha = 1.20$)	5.0	5.0	5.0	5.0	1003.0	91.5	95.0	95.5
				Model 1.d				
	5%	25%	50%	75%	95%	$\mathcal{P}_{d_1}(\%)$	$\mathcal{P}_{d_2}(\%)$	$\mathcal{P}_{d_3}(\%)$
SIS	6.0	25.0	87.0	206.5	569.5	75.0	90.5	95.5
DC-SIS	5.0	5.0	12.0	68.3	388.6	88.5	95.5	97.0
CR-SIS	5.0	5.0	5.0	5.0	5.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.35$)	5.0	5.0	5.0	5.0	5.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.70$)	5.0	5.0	5.0	5.0	5.0	99.0	99.0	99.5
BanditCR-SIS ($\alpha = 1.00$)	5.0	5.0	5.0	5.0	504.0	94.0	95.5	97.0
BanditCR-SIS ($\alpha = 1.20$)	5.0	5.0	5.0	5.0	1003.0	90.0	93.5	97.0

challenging nonlinear relationships. Additionally, Model 2.e addresses a case with heteroscedastic errors combined with nonlinear regression.

In Table 2, we observe that CR-SIS performs consistently well, while other methods fail to effectively screen out inactive features in complex nonlinear scenarios across 200 replications. This highlights the remarkable ability of Chatterjee's rank correlation to capture nonlinear relationships between variables. Additionally, CR-SIS demonstrates consistently superior performance in handling heteroscedastic variance, further underscoring its robustness in diverse settings. In addition, it is also observed that the hyper-parameter α plays a critical role in balancing the accuracy and computational efficiency. As α decreases, both $\mathcal{S}$ and $\mathcal{P}$ for BanditCR-SIS gradually approach those of CR-SIS. When α is large, BanditCR-SIS trades a certain degree of accuracy for improved efficiency. This trade-off remains acceptable, as BanditCR-SIS successfully identifies all the active features in nearly 80% of the linear settings (Models 1.a–d). In more challenging nonlinear settings (Models 2.a–e), BanditCR-SIS with $d = d_1$ identifies all active features in roughly 60% of the replications, while the version with $d = d_3$ achieves about 80% accuracy, substantially outperforming SIS and DC-SIS, which often fail to detect all important features under these scenarios. A detailed analysis of how computational time varies with α is provided in the next subsection.

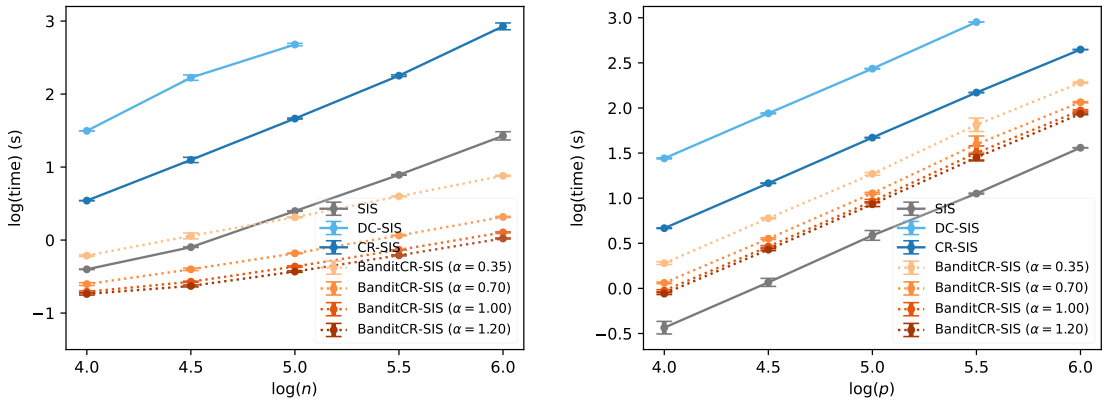
5.2 Computational Time

We compare the average CPU time (in seconds) versus the sample size n and the feature number p for the proposed CR-SIS and BanditCR-SIS methods against competitors SIS and DC-SIS. Here, DC-SIS is replaced by an accelerated version in the univariate setting (Huo and Székely, 2016), achieving a computational complexity of $O(n \log(n))$ through the use of AVL trees. In the first scenario, we fix the number of features at $p = 2000$, and in the second scenario, we fix the sample size at $n = 2000$. In both cases, the number of retained features is set to $d = 20$, and the experiments are repeated over 200 times. The hyper-parameter α is tested with values $0.35, 0.70, 1.00$, and 1.20, consistent with the previous section. The results are presented in Figure 2. The slope of the folded line in the graph of $\log(\text{time})$ versus $\log(n)$ (on the left) or $\log(p)$ (on the right) provides an indication of how quickly the computational time of the algorithms scales with the sample size n or the feature number p, respectively.

As illustrated in Figure 2(a), DC-SIS requires significantly more CPU time, making it impractical for large datasets. In contrast, CR-SIS exhibits a slope magnitude similar to that of DC-SIS, indicating comparable scalability with sample size, but it requires substantially less CPU time. All of these methods with their computational time grow sharply as the sample size increases. These observations

Table 2: The quantiles of minimum model size $\mathcal{S}$ and the proportion $\mathcal{P}$ with $d_1 = 205, d_2 = 410, d_3 = 615$ for Models 2 over 200 replications. The hyper-parameters in BanditCR-SIS are set to 0.35, 0.70, 1.00 and 1.20.

Model 2.a								
	5%	25%	50%	75%	95%	$\mathcal{P}_{d_1}(\%)$	$\mathcal{P}_{d_2}(\%)$	$\mathcal{P}_{d_3}(\%)$
SIS	83.4	339.3	855.0	1457.5	1888.1	15.5	30.0	41.0
DC-SIS	26.9	205.5	449.5	942.0	1632.5	25.0	45.5	63.5
CR-SIS	4.0	4.0	4.0	4.0	5.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.35$)	4.0	4.0	4.0	4.0	5.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.70$)	4.0	4.0	4.0	4.0	254.0	96.5	97.0	98.0
BanditCR-SIS ($\alpha = 1.00$)	4.0	4.0	4.0	5.0	1003.0	90.0	93.5	94.5
BanditCR-SIS ($\alpha = 1.20$)	4.0	4.0	4.0	5.0	1003.0	84.5	86.0	91.0
Model 2.b								
	5%	25%	50%	75%	95%	$\mathcal{P}_{d_1}(\%)$	$\mathcal{P}_{d_2}(\%)$	$\mathcal{P}_{d_3}(\%)$
SIS	1869.9	1923.0	1960.5	1981.0	1997.1	0.0	0.0	0.0
DC-SIS	341.5	830.0	1028.0	1582.8	1853.0	2.0	8.0	15.5
CR-SIS	4.0	4.0	4.0	4.0	10.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.35$)	4.0	4.0	4.0	4.0	504.0	95.0	97.0	97.5
BanditCR-SIS ($\alpha = 0.70$)	4.0	4.0	4.0	254.0	1003.0	84.0	88.0	90.5
BanditCR-SIS ($\alpha = 1.00$)	4.0	4.0	7.0	504.0	1003.0	67.0	78.0	82.5
BanditCR-SIS ($\alpha = 1.20$)	4.0	4.0	254.0	504.0	1003.0	59.0	70.5	82.5
Model 2.c								
	5%	25%	50%	75%	95%	$\mathcal{P}_{d_1}(\%)$	$\mathcal{P}_{d_2}(\%)$	$\mathcal{P}_{d_3}(\%)$
SIS	142.8	744.0	1253.0	1765.0	1969.1	6.5	11.5	20.0
DC-SIS	142.8	474.3	1325.5	1803.0	1992.2	6.0	12.0	19.5
CR-SIS	4.0	4.0	4.0	4.0	4.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.35$)	4.0	4.0	4.0	4.0	4.0	99.5	99.5	99.5
BanditCR-SIS ($\alpha = 0.70$)	4.0	4.0	4.0	4.0	1003.0	86.0	91.5	93.0
BanditCR-SIS ($\alpha = 1.00$)	4.0	4.0	4.0	504.0	1003.0	72.0	77.0	81.0
BanditCR-SIS ($\alpha = 1.20$)	4.0	4.0	98.0	1003.0	1003.0	60.5	70.0	75.0
Model 2.d								
	5%	25%	50%	75%	95%	$\mathcal{P}_{d_1}(\%)$	$\mathcal{P}_{d_2}(\%)$	$\mathcal{P}_{d_3}(\%)$
SIS	905.6	1246.5	1586.5	1810.8	1965.2	0.0	0.0	1.0
DC-SIS	714.4	1124.0	1483.5	1700.5	1963.1	0.5	0.5	3.0
CR-SIS	4.0	4.0	4.0	4.0	5.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.35$)	4.0	4.0	4.0	4.0	14.1	96.5	98.0	98.0
BanditCR-SIS ($\alpha = 0.70$)	4.0	4.0	4.0	254.0	1003.0	79.5	87.5	89.0
BanditCR-SIS ($\alpha = 1.00$)	4.0	4.0	254.0	1003.0	1003.0	60.5	71.5	78.0
BanditCR-SIS ($\alpha = 1.20$)	4.0	4.0	504.0	1003.0	1003.0	51.5	64.5	72.0
Model 2.e								
	5%	25%	50%	75%	95%	$\mathcal{P}_{d_1}(\%)$	$\mathcal{P}_{d_2}(\%)$	$\mathcal{P}_{d_3}(\%)$
SIS	5.0	23.8	176.5	542.3	1694.5	56.0	72.0	78.5
DC-SIS	5.0	5.0	7.0	35.5	286.7	92.0	95.5	96.5
CR-SIS	4.0	4.0	5.0	5.0	7.0	100.0	100.0	100.0
BanditCR-SIS ($\alpha = 0.35$)	4.0	4.0	5.0	5.0	504.0	95.0	96.0	98.0
BanditCR-SIS ($\alpha = 0.70$)	4.0	4.0	5.0	13.5	1003.0	82.5	87.0	89.5
BanditCR-SIS ($\alpha = 1.00$)	4.0	4.0	5.0	254.0	1003.0	78.5	83.5	89.5
BanditCR-SIS ($\alpha = 1.20$)	4.0	4.0	5.0	504.0	1003.0	67.0	77.0	82.5



(a) Sample size n vs. CPU time.

(b) Feature number p vs. CPU time.

Figure 2: (a) The sample size n versus the CPU time (in seconds) of SIS, DC-SIS, CR-SIS and BanditCR-SIS for different hyper-parameters ($\alpha = 0.35, 0.70, 1.00$ and 1.20) with $p = 2000, d = 20$. (b) The number of features p versus the CPU time (in seconds) of SIS, DC-SIS, CR-SIS and BanditCR-SIS for the same hyper-parameters ($\alpha = 0.35, 0.70, 1.00$ and 1.20) with $n = 2000, d = 20$. Vertical bars are the standard errors with respect to 200 replicates.

are expected since both DC-SIS and CR-SIS have a computational complexity of $O(n\log(n)p)$, while SIS has a lower complexity of $O(np)$. However, CR-SIS achieves a better efficiency in screening non-informative features compared to DC-SIS, especially when the sample size n is large. Moreover, for the proposed BanditCR-SIS method, the computational time grows approximately sublinearly with respect to the sample size n. This behavior is desirable, as BanditCR-SIS, being an efficient variant of CR-SIS, achieves this scalability due to its computational complexity of $O(\sqrt{n}\log(n)p + n\log(n))$. It maintains a satisfactory level of screening accuracy while significantly reducing the computational workload. Notably, when the hyper-parameter α is relatively large (e.g., greater than 0.70), the computational time of BanditCR-SIS becomes even lower than that of SIS, demonstrating its high efficiency in screening out noninformative features. Furthermore, we observe that BanditCR-SIS requires more CPU time as the value of α decreases. This phenomenon arises because smaller values α correspond to a larger number of subsamples selected in each iteration, leading to a more precise estimation of the correlation coefficient. In essence, the hyper-parameter α balances the computational efficiency and performance of the BanditCR-SIS method, serving as a trade-off between speed and accuracy.

Figure 2(b) shows that all screening methods exhibit similar slopes, suggesting comparable computational complexity with respect to p. Moreover, the performance of BanditCR-SIS becomes closer to that of CR-SIS as the value of α decreases. This behavior can be attributed to the subsampling mechanism

described in Section 3.

5.3 Real Data Example

We consider a regression task on a dataset named Relative Location of CT Slices on Axial (Graf et al., 2011) from the UCI Machine Learning Repository. This dataset was derived from a collection of 53,500 CT images obtained from 74 different patients. Each CT slice is characterized by two histograms in polar space. The first histogram represents the location of bone structures within the image, while the second captures the location of air inclusions inside the body. These two histograms are concatenated to create the final feature vectors. The response variable corresponds to the relative location of the CT slice along the axial plane, ranging from 0 to 180, where 0 represents the top of the head and 180 indicates the soles of the feet. The objective of this dataset is to predict the relative location of a CT slice based on its extracted features.

After removing missing values, this dataset consists of $n = 53500$ samples and $p = 384$ features. We randomly partition the data, using 80% as a training set and the remaining 20% as the testing set. During the screening procedures, we retain $d = 50$ features and use the selected features to fit a decision tree regression model with a maximum depth of 30 after the screening. Finally, we compare the mean absolute percentage errors of predictions based on SIS, DC-SIS, CR-SIS, and BanditCR-SIS under varying hyper-parameters over 100 replicates.

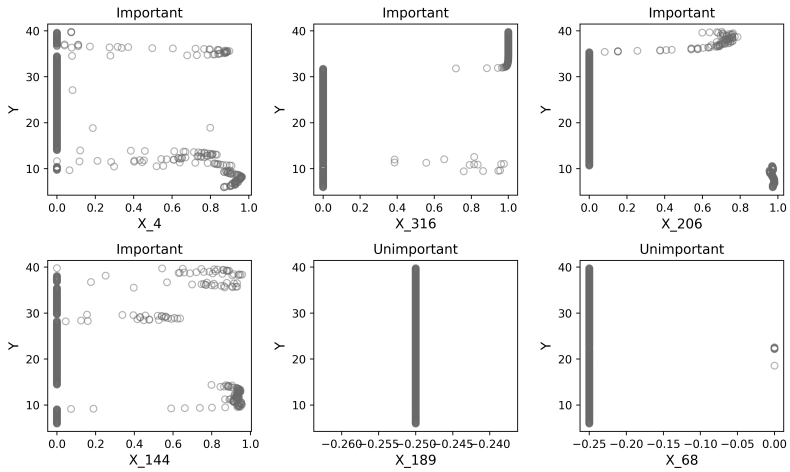


Figure 3: Visualization of the screening results given by CR-SIS. The scatter plots show the relationship between the response against the four important and two unimportant features identified for the first patient.

Figure 3 illustrates the screening results of CR-SIS, showing the response in relation to the four

important and two unimportant features identified for the first patient. A clear and strong relationship is observed between the response and the important features, while the unimportant ones display only a trivial constant value, underscoring their weak explanatory power. The fitting results, summarized in Table 3, indicate that the prediction errors of CR-SIS and BanditCR-SIS are consistently lower than those of SIS and DC-SIS, highlighting the effectiveness of Chatterjee's rank correlation in capturing complex dependence structures in this dataset. However, DC-SIS demands significantly more computational time for feature screening compared to the other methods. Notably, the proposed BanditCR-SIS method achieves prediction errors comparable to CR-SIS while requiring significantly less CPU time. Notably, when the tuning parameter α is moderately large (e.g., $\alpha \geq 0.70$), the computational time of BanditCR-SIS becomes even less than that of SIS, which is applicable only in linear settings.

Table 3: Mean absolute percentage errors (MAPE) of predictions and time consumption (seconds) based on SIS, (Fast) DC-SIS, CR-SIS and BanditCR-SIS with different hyper-parameters ($\alpha = 0.70$ and 1.20) for 80% training set over 100 replicates.

	SIS	**DC-SIS**	**CR-SIS**	**BanditCR-SIS** $\alpha = 0.70$	**BanditCR-SIS** $\alpha = 1.20$
MAPE	13.360 ($\pm$0.331)	11.802 ($\pm$0.273)	6.034 ($\pm$0.345)	6.477 ($\pm$0.411)	6.481 ($\pm$0.579)
Time (s)	0.347 ($\pm$0.109)	42.369 ($\pm$3.039)	2.925 ($\pm$0.284)	0.750 ($\pm$0.053)	0.363 ($\pm$0.022)

6 Conclusion

In this paper, we introduce a rank-based correlation, named Chatterjee's rank, into the screening procedure, that is, CR-SIS. To further alleviate the corresponding computational burden, we further develop a more efficient variant named BanditCR-SIS, inspired by multi-armed bandit algorithms. We establish the sure screening property for both methods when the number of predictors grows exponentially with the sample size. The proposed CR-SIS method is particularly effective in detecting informative features, especially when the variables exhibit complex nonlinear relationships. Additionally, we accelerate the computational complexity of the CR-SIS method from $O(n\log(n)p)$ to $O(\sqrt{n}\log(n)p + n\log(n))$, which is a sublinear order with respect to sample size n. This improvement makes BanditCR-SIS significantly more efficient for large-scale datasets compared to existing model-free screening methods, while preserving the strong screening performance of CR-SIS. Our numerical studies, conducted on both synthetic and real datasets, demonstrate the superior performance of the proposed methods in terms of both

computational efficiency and screening accuracy.

Like most independence screening methods, CR-SIS also struggles to identify truly important predictors that are marginally independent of, but jointly correlated with, the response variable. To overcome this issue, some iterative screening methods have been proposed, such as iterative SIS (Fan and Lv, 2008) and iterative DC-SIS (Zhong and Zhu, 2015). However, these methods often suffer from significantly higher computational costs. The iterative nature of these procedures aligns well with the BanditCR-SIS algorithm, as both iteratively eliminate noninformative features in batches. Therefore, developing more efficient methods to identify important predictors that are marginally independent of the response, while also addressing the computational burden, is a critical topic for our future research.

Supplementary Material

Supplementary material includes technical details of the theoretical results and additional numerical results.

References

Audibert, J.-Y. and S. Bubeck (2010). Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory*.

Bubeck, S., R. Munos, and G. Stoltz (2009). Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pp. 23–37. Springer.

Carroll, R. J., J. Fan, I. Gijbels, and M. P. Wand (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association 92*(438), 477–489.

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

Chen, R., L. Yang, S. Goodison, and Y. Sun (2020). Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics 36*(5), 1476–1483.

Chen, S., T. Lin, I. King, M. R. Lyu, and W. Chen (2014). Combinatorial pure exploration of multi-armed bandits. *Advances in Neural Information Processing Systems 27*, 379–387.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*(2), 187–202.

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

Ding, K., J. Li, and H. Liu (2019). Interactive anomaly detection on attributed networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 357–365.

Durand, A., C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau (2018). Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine Learning for Healthcare Conference*, pp. 67–82. PMLR.

Even-Dar, E., S. Mannor, and Y. Mansour (2002). PAC bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory: 15th Annual Conference on Computational Learning Theory, COLT 2002 Sydney, Australia, July 8–10, 2002 Proceedings 15*, pp. 255–270. Springer.

Even-Dar, E., S. Mannor, and Y. Mansour (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research 7*, 1079–1105.

Fan, J. and Y. Fan (2008). High dimensional classification using features annealed independence rules. *Annals of Statistics 36*(6), 2605.

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

Fan, J. and J. Lv (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., Y. Ma, and W. Dai (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association 109*(507), 1270–1284.

Fan, J. and Y. Ren (2006). Statistical analysis of DNA microarray data in cancer research. *Clinical Cancer Research 12*(15), 4469–4473.

Fan, J., R. Samworth, and Y. Wu (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research 10*, 2013–2038.

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

Ghorbani, A., D. Ouyang, A. Abid, B. He, J. H. Chen, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou (2020). Deep learning interpretation of echocardiograms. *NPJ Digital Medicine 3*(1), 10.

Graf, F., H. Kriegel, M. Schubert, S. Poelsterl, and A. Cavallaro (2011). Relative location of CT slices on axial axis data set. *UCI Machine Learning Repository*.

Hall, P. and H. Miller (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics 18*(3), 533–550.

Härdle, W., H. Liang, and J. Gao (2000). *Partially linear models*. Springer Science & Business Media.

Huo, X. and G. J. Székely (2016). Fast computing for distance covariance. *Technometrics 58*(4), 435–447.

Jamieson, K. and R. Nowak (2014). Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual conference on Information Sciences and Systems*, pp. 1–6. IEEE.

Jamieson, K. and A. Talwalkar (2016). Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, pp. 240–248. PMLR.

Kalyanakrishnan, S., A. Tewari, P. Auer, and P. Stone (2012). PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, Volume 12, pp. 655–662.

Karnin, Z., T. Koren, and O. Somekh (2013). Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 1238–1246. PMLR.

Kaufmann, E., O. Cappé, and A. Garivier (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research 17*(1), 1–42.

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

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

Li, T., J. Yu, and C. Meng (2023). Scalable model-free feature screening via sliced-Wasserstein dependency. *Journal of Computational and Graphical Statistics 32*(4), 1501–1511.

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.

Liu, W., Y. Ke, J. Liu, and R. Li (2022). Model-free feature screening and FDR control with knockoff features. *Journal of the American Statistical Association 117*(537), 428–443.

Lv, J. and J. S. Liu (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*(1), 141–167.

Mary, J., R. Gaudel, and P. Preux (2015). Bandits and recommender systems. In *First International Workshop on Machine Learning, Optimization, and Big Data*, pp. 325–336. Springer.

Pang, G., L. Cao, L. Chen, and H. Liu (2018). Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2041–2050.

Shao, X. and J. Zhang (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association 109*(507), 1302–1318.

Wang, H., J. Yan, and X. Yan (2023). Spearman rank correlation screening for ultrahigh-dimensional censored data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 37, pp. 10104–10112.

Weinberger, K., A. Dasgupta, J. Langford, A. Smola, and J. Attenberg (2009). Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1113–1120.

Xu, K., Z. Shen, X. Huang, and Q. Cheng (2020). Projection correlation between scalar and vector variables and its use in feature screening with multi-response data. *Journal of Statistical Computation and Simulation 90*(11), 1923–1942.

Zhong, W. and L. Zhu (2015). An iterative approach to distance correlation-based sure independence screening. *Journal of Statistical Computation and Simulation 85*(11), 2331–2345.

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

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- appendix.pdf