Journal of Machine Learning Research 21 (2020) 1-32 Submitted 6/19; Revised 12/19; Published 02/20

Distributed Feature Screening via Componentwise Debiasing

Xingxiang Li XLI396@UOTTAWA.CA
School of Mathematics and Statistics
Xi'an Jiaotong University, China
Department of Mathematics and Statistics
University of Ottawa, Canada

Runze Li RZLI@PSU.EDU
Department of Statistics and The Methodology Center
The Pennsylvania State University, USA

Zhiming Xia STATXZM@NWU.EDU.CN
School of Mathematics
Northwest University, China

Chen Xu CX3@UOTTAWA.CA
Department of Mathematics and Statistics
University of Ottawa, Canada

Editor: Garvesh Raskutti

Abstract

Feature screening is a powerful tool in processing high-dimensional data. When the sample size N and the number of features p are both large, the implementation of classic screening methods can be numerically challenging. In this paper, we propose a distributed screening framework for big data setup. In the spirit of "divide-and-conquer", the proposed framework expresses a correlation measure as a function of several component parameters, each of which can be distributively estimated using a natural U-statistic from data segments. With the component estimates aggregated, we obtain a final correlation estimate that can be readily used for screening features. This framework enables distributed storage and parallel computing and thus is computationally attractive. Due to the unbiased distributive estimation of the component parameters, the final aggregated estimate achieves a high accuracy that is insensitive to the number of data segments m. Under mild conditions, we show that the aggregated correlation estimator is as efficient as the centralized estimator in terms of the probability convergence bound and the mean squared error rate; the corresponding screening procedure enjoys sure screening property for a wide range of correlation measures. The promising performances of the new method are supported by extensive numerical examples.

Keywords: Feature screening, Big data, Divide-and-conquer, Aggregated correlation, Sure screening property

1. Introduction

With rapid development of data generation and acquisition, massive data with a huge number of features are frequently encountered in many scientific fields. High dimensionality poses simultaneous challenges of computational cost, statistical accuracy, and algorithmic

stability for classic statistical methods (Fan et al., 2009). To facilitate the computing process, one natural strategy is to screen most irrelevant features out before an elaborative analysis. This procedure is referred to as feature screening. With dimensionality reduced from high to low, analytical difficulties are reduced drastically. In the literature, plenty of works have been done in this area; in particular, the correlation-based screening methods have attracted a great deal of attention. These methods conduct screening based on a certain correlation measure between features and the response. Features with weak correlations are treated as irrelevant ones and are to be removed. This type of methods can be conveniently implemented without strong model assumptions (even model-free). Thus, they are commonly used for analyzing high-dimensional data with complex structures. For example, Fan and Lv (2008) proposed a sure independence screening (SIS) based on Pearson correlation. Zhu et al. (2011) proposed a sure independent ranking and screening (SIRS) based on a utility measure that is concerned with the entire conditional distribution of the response given the predictors. Li et al. (2012a) proposed a robust rank correlation screening (RRCS) based on the Kendall τ rank correlation. Li et al. (2012b) developed a model-free sure independence screening procedure based on the distance correlation (DC-SIS). Wu and Yin (2015) proposed a distribution function sure independence screening (DF-SIS) approach, which uses a measure to test the independence of two variables. Zhou et al. (2019) proposed a robust correlation measure to screen features containing extreme values.

Feature screening has been demonstrated to be an attractive strategy in many applications. Most existing methods are developed under the situation, where the number of features p is large but the sample size N is moderate. However, in modern scientific research, it is increasingly common that data analysts have to deal with big data sets, where p and N are both huge. For example, in modern genome wide genetic studies, millions of SNPs are genotyped on hundreds of thousands participants. In Internet studies, an antivirus software may scan tens of thousands keywords in millions of URLs per minute. When faced with large-p-large-N data, the direct implementation of classic screening methods can be numerically inefficient due to storage bottleneck and algorithmic feasibility. For example, for a data set with $N = p = 10,000$, the well-known DC-SIS needs about 60 hours to conduct a full screening on a computer with 3.2 GHz CPU and 32 GB memory. Developing computationally convenient methods for big data screening is therefore desirable in practice.

When a data set is too huge to be processed on a single computer, it is natural to consider using a "divide-and-conquer" strategy. In such a strategy, a large problem is first divided into smaller manageable subproblems and the final output is obtained by combining the corresponding sub-outputs. In this spirit, many machine learning and statistical methods have been rebuilt for processing big data (e.g., Chen and Xie, 2014; Xu et al., 2016; Jordan et al., 2019; Shi et al., 2018; Banerjee et al., 2019). These inspiring works motivate us to explore the feasibility of using this promising strategy for feature screening with big data.

In this paper, we propose a distributed feature screening framework based on aggregated correlation measures, and refer to it as aggregated correlation screening (ACS). In ACS, we express a correlation measure as a function of several component parameters, each of which can be distributively estimated using a natural U-statistic from data segments. With the unbiased component estimates combined together, we obtain an aggregated correlation estimate, which can be readily used for feature screening. In the proposed ACS framework, a massive data set is split into and processed in m manageable segments, which can be

stored in multiple computers and the corresponding local estimations can be done by parallel computing. It thus provides a computationally attractive route for feature screening with large-p-large-N data. This framework is also suitable for the setup, where data are naturally stored in different locations (e.g., medical data at hospital level). The U-statistic estimation of the component parameters serves as an effective and convenient debasing technique, which ensures the high accuracy of the aggregated correlation estimator and the reliability of the corresponding screening procedure. Under mild conditions, we show that the aggregated correlation estimator is as efficient as the classic centralized estimator in terms of the probability convergence bound and the mean squared error (MSE) rate. Such a full efficiency is insensitive to the choice of m, which may be specified by the problem itself or to be determined by the users. For a wide range of correlation measures, we further show that ACS enjoys the sure screening property without the need of specifying a parametric model (model-free).

The proposed ACS has its roots in component-wise estimation. In the literature, this idea has been used to distributively recover a centralized estimator defined by smooth estimating equations that are separatable in data segments (e.g., Chen et al., 2006; Lin and Xi, 2011). Unfortunately, these works are not directly applicable to the correlation-based screening, as the centralized estimators of many commonly-used correlation measures are not typically defined by estimating equations and are non-separatable in data segments (e.g., SIRS and DC). The proposed ACS follows from the natural composition of a centralized correlation estimator; it does not seek to fully recover the centralized estimator but leads to an effective and computationally affordable alternative of it. Our results in this paper provide a theoretical support of using this natural strategy for distributed feature screening. We demonstrate the computational advantages and promising screening accuracy of ACS in a series of numerical examples.

The rest of this paper is organized as follows. In Section 2, we formulate the research problem and introduce the ACS framework. In Section 3, we investigate the theoretical properties of ACS. In Section 4, we demonstrate the promising performance of ACS by Monte Carlo simulations and a real data example. Concluding remarks are given in Section 5 and the proofs of theorems are provided in the Appendix.

2. Methodology

2.1. Feature screening with big data

Let $\mathcal{D} = \{(Y_i, \mathbf{X}_i)\}_{i=1}^N$ be N independently and identically distributed (i.i.d) copies of $\{Y, \mathbf{X}\}$, where Y is a response variable with support Φ_y and $\mathbf{X} = (X_1, ..., X_p)^T$ is a p-dimensional covariate vector. We are interested in the situation, where p and N are both large. When a data set is massive and high-dimensional, it is often reasonable to assume that only a handful of covariates (features) are relevant to the response. Let $F(y|\mathbf{X})$ be the conditional distribution function of Y given $\mathbf{X}$. A feature X_j is considered to be relevant if $F(y|\mathbf{X})$ functionally depends on X_j for some $y \in \Phi_y$. We use $\mathcal{M}$ to denote the index set of the relevant features and define $\mathcal{M}^c = \{1, ..., p\} \setminus \mathcal{M}$. The goal of feature screening is to remove most irrelevant features X_js with $j \in \mathcal{M}^c$ before an elaborative analysis.

One commonly used strategy is to first estimate a marginal correlation measure between the response and each feature, and then remove the features with weak correlations.

Specifically, let $\omega_j \geq 0$ be a measure of correlation strength between Y and X_j. Let $\hat{\omega}_j$ be a centralized estimate of ω_j based on $\mathcal{D}$. With a pre-specified threshold $\gamma > 0$, one may retain the features in

$$\hat{\mathcal{M}} = \{j : \hat{\omega}_j \geq \gamma, j = 1, ..., p\},$$

and remove the others. This classic approach is effective when sample size N is moderate. However, when N and p are both huge, computing $\{\hat{\omega}_j\}_{j=1}^p$ based on the full data set $\mathcal{D}$ can be numerically costly.

2.2. Aggregated correlation screening

Motivated by the recent works in distributed learning, we consider adopting the idea of "divide-and-conquer" to tackle big data feature screening. Without loss of generality, suppose that the original full data set $\mathcal{D}$ is equally partitioned into m manageable segments $\{\mathcal{D}_l\}_{l=1}^m$, each of which contains $n = N/m$ observations. Depending on the computational environment, these segments can be distributively stored on and processed by multiple computers or can be sequentially processed by a single computer. Let $\hat{\omega}_{l,j}$ be the local correlation estimate between X_j and Y based on data segment $\mathcal{D}_l$. One simple screening strategy is to compute an averaged correlation estimate

$$\bar{\omega}_j = \frac{1}{m} \sum_{l=1}^{m} \hat{\omega}_{l,j} \tag{1}$$

for $1 \leq j \leq p$ and remove the features with small $\bar{\omega}_j$ values. This approach is referred to as simple average screening (SAS), which perhaps is the most straightforward way for distributed screening. To facilitate the computing process, using a relatively large number of small segments is often preferred in the analysis. However, when m is large, $\bar{\omega}_j$ may substantially differ from the centralized estimator $\hat{\omega}_j$ due to the cumulated bias inherited from the local estimators. As a result, its screening performance is often unstable in practice, as to be revealed in our numerical studies.

One way to improve SAS is to conduct debiasing on $\hat{\omega}_{l,j}$s before averaging them over. Unfortunately, this is not straightforward for many commonly-used correlation measures that are nonlinear. Our idea is to express a correlation measure ω_j as a function of several component parameters, and conduct the distributed unbiased estimation of the component parameters. By doing so, we carry out componentwise debasing on original $\hat{\omega}_{l,j}$s in an effective but much easier way. With the unbiased component estimates naturally combined together, we obtain an aggregated correlation estimate that can be readily used for feature screening.

To be more specific, suppose that a correlation measure between Y and X_j can be expressed as

$$\omega_j = g(\theta_{j,1}, ..., \theta_{j,s}), \tag{2}$$

where g is a pre-specified function and $\theta_{j,1}, ..., \theta_{j,s}$ are s component parameters from a compact space. For a given correlation measure, expression (2) may not be unique. We choose the form of g such that the corresponding component parameters can be conveniently estimated with no bias. For the ease of presentation, let $\hat{\theta}_{j,h}(Z_{i_1 j}, \ldots, Z_{i_{k_h} j})$ denote a basis unbiased estimator (kernel) of $\theta_{j,h}$ with the minimal k_h i.i.d copies of $Z_j = \{Y, X_j\}$ for

$h = 1, \ldots, s$. Without loss of generality, we assume that $\hat{\theta}_{j,h}$ is symmetric such that its value is invariant to the permutation of $\{Z_{i_1 j}, \ldots, Z_{i_{k_h} j}\}$.

Suppose that $\mathcal{D}$ is too big to be processed on a single computer and is equally partitioned into m segments $\{\mathcal{D}_l\}_{l=1}^m$. We use $\mathcal{S}_l$ to denote the index set of $\{Y, \mathbf{X}\}$ copies on $\mathcal{D}_l$. With a pre-specified correlation measure ω_j, we propose to distributively screen features in the following framework.

1. Express ω_j in the form of (2) with an appropriate g.
2. On each data segment, we estimate $\theta_{j,h}$ by a local U-statistic

$$U_{j,h}^l = \binom{n}{k_h}^{-1} \sum_{\{i_1, \ldots, i_{k_h}\} \in \mathcal{S}_l} \hat{\theta}_{j,h}(Z_{i_1 j}, ..., Z_{i_{k_h} j}), \tag{3}$$

   where the summation is over all $\{Z_{i_1 j}, ..., Z_{i_{k_h} j}\}$ combinations chosen from $\mathcal{D}_l$.
3. We compute an aggregated correlation estimate between Y and X_j by

$$\widetilde{\omega}_j = g(\bar{U}_{j,1}, ..., \bar{U}_{j,s}), \tag{4}$$

   where $\bar{U}_{j,h} = \frac{1}{m} \sum_{l=1}^m U_{j,h}^l$ for $h = 1, \ldots, s$.
4. With a user-specified threshold $\gamma > 0$, we retain the features in

$$\widetilde{\mathcal{M}} = \{j : \widetilde{\omega}_j \geq \gamma, j = 1, ..., p\},$$

   and remove the others.

We name the proposed screening framework as the aggregated correlation screening (ACS). It is seen that step 2 only requires information stored on the data segments, and thus it can be carried out by parallel or sequential processing. This makes ACS computationally suitable for the large-p-large-N situation. The use of U-statistics in step 2 helps to further reduce the variances of the local unbiased estimators on $\theta_{j,h}$s and helps to enhance the stability of the method. The computational complexity of (3) is $O(n^{k_h})$, which can be conveniently handled with an appropriate m such that the local sample size $n = N/m$ is moderate. Compared with SAS, ACS screens features based on a non-linear aggregation of unbiased component estimates. This way enables us to substantially reduce the bias of the final correlation estimate with a little sacrifice on the variance. The overall accuracy of the ω_j estimate is therefore improved; this in turn leads to a more reliable screening result in the distritbuted setup.

2.3. Examples and extension

2.3.1. Examples

The proposed ACS framework is suitable for many commonly used correlation measures. In this subsection, we provide a few concrete examples of using ACS. As a reference, we also list the corresponding expressions of $\hat{\omega}_{l,j}$ used in SAS in the Appendix for the interested readers. For the convenience of presentation, let X_{ij} denote the jth entry of $\mathbf{X}_i$ defined in Section 2.1 for $j = 1, \ldots, p$.

1. **Pearson correlation**

   Pearson correlation measures the strength of linear relationship between Y and X_j. Fan and Lv (2008) used it as a feature screening index for the linear model. When Pearson correlation is used in ACS, ω_j can be expressed in the form of (2) by

$$\omega_j = g(\theta_{j,1}, ..., \theta_{j,5}) = \left| \frac{E(X_jY) - E(X_j)E(Y)}{\sqrt{(EX_j^2 - E^2(X_j))(EY^2 - E^2(Y))}} \right|,$$

   where $\theta_{j,1} = E(X_jY)$, $\theta_{j,2} = E(X_j)$, $\theta_{j,3} = E(Y)$, $\theta_{j,4} = EX_j^2$, and $\theta_{j,5} = EY^2$. In step 2 of ACS, $U_{j,h}^l$ can be computed by (3) with $k_h = 1$ and

$$\hat{\theta}_{j,1} = X_{i_1 j}Y_{i_1}, \quad \hat{\theta}_{j,2} = X_{i_1 j}, \quad \hat{\theta}_{j,3} = Y_{i_1}, \quad \hat{\theta}_{j,4} = X_{i_1 j}^2, \quad \hat{\theta}_{j,5} = Y_{i_1}^2,$$

   for $i_1 \in \mathcal{S}_l$. It is seen that $\bar{U}_{j,h}$ in (4) coincides with classic moment estimates. When the data set is properly standardized, the expression of ω_j can be further simplified.

2. **Kendall τ rank correlation**

   Kendall τ rank correlation measures the ordinal association between Y and X_j. It was used in Li et al. (2012a) for feature screening in linear and transformation models. When this correlation measure is used in ACS, ω_j can be expressed by

$$\omega_j = g(\theta_{j,1}) = \left| E(I(X_j < X_j')I(Y < Y')) - 1/4 \right|,$$

   where $\{X_j', Y'\}$ is an independent copy of $\{X_j, Y\}$ and $\theta_{j,1} = E(I(X_j < X_j')I(Y < Y'))$. In step 2 of ACS, $U_{j,1}^l$ can be computed by (3) with $k_1 = 2$ and

$$\hat{\theta}_{j,1} = \frac{1}{2} \sum_{(i_1, i_2)} I(X_{i_1 j} < X_{i_2 j})I(Y_{i_1} < Y_{i_2}),$$

   where $\{i_1, i_2\} \in \mathcal{S}_l$ and the summation is over all permutations of (i_1, i_2).

3. **SIRS correlation**

   SIRS correlation can be used to detect nonlinear relationship between Y and X_j. It was proposed by Zhu et al. (2011) for feature screening in parametric and semiparametric models. When this correlation is used in ACS, ω_j can be expressed by

$$\omega_j = \theta_{j,1} = E_{Y'}\{E^2(X_jI(Y < Y'))\},$$

   where Y' is an independent copy of Y and feature X_j is assumed to have zero mean and unit variance. In step 2 of ACS, $U_{j,1}^l$ can be computed by (3) with $k_1 = 3$ and

$$\hat{\theta}_{j,1} = \frac{1}{6} \sum_{(i_1, i_2, i_3)} X_{i_1 j}X_{i_2 j}I(Y_{i_1} < Y_{i_3})I(Y_{i_2} < Y_{i_3}),$$

   where $\{i_1, i_2, i_3\} \in \mathcal{S}_l$ and the summation is over all permutations of (i_1, i_2, i_3).

4. **Distance correlation**

Distance correlation (DC) can be used to measure the dependence between Y and X_j. Li et al. (2012b) used it as a model-free screening index. When DC is used in ACS, ω_j can be expressed by

$$\omega_j = g(\theta_{j,1}, ..., \theta_{j,8}) = \frac{\theta_{j,1} + \theta_{j,2} \cdot \theta_{j,3} - 2\theta_{j,4}}{\sqrt{(\theta_{j,5} + \theta_{j,2}^2 - 2\theta_{j,6})(\theta_{j,7} + \theta_{j,3}^2 - 2\theta_{j,8})}}$$

with

$$\begin{aligned}
&\theta_{j,1} = E\{|Y - Y'| \cdot |X_j - X_j'|\},\\
&\theta_{j,2} = E\{|Y - Y'|\}, \quad \theta_{j,3} = E\{|X_j - X_j'|\},\\
&\theta_{j,4} = E\{E(|Y - Y'| \mid Y)E(|X_j - X_j'| \mid X_j)\},\\
&\theta_{j,5} = E\{|Y - Y'|^2\}, \quad \theta_{j,6} = E\{E^2(|Y - Y'| \mid Y)\},\\
&\theta_{j,7} = E\{|X_j - X_j'|^2\}, \quad \theta_{j,8} = E\{E^2(|X_j - X_j'| \mid X_j)\},
\end{aligned}$$

where (Y', X_j') is an independent copy of (Y, X_j). In step 2 of ACS, $U_{j,1}^l$, $U_{j,4}^l$ can be computed by (3) with $k_1 = 2$, $k_4 = 3$, and

$$\hat{\theta}_{j,1} = \frac{1}{2} \sum_{(i_1,i_2)} |Y_{i_1} - Y_{i_2}| \cdot |X_{i_1 j} - X_{i_2 j}|, \tag{5}$$

$$\hat{\theta}_{j,4} = \frac{1}{6} \sum_{(i_1,i_2,i_3)} |Y_{i_1} - Y_{i_3}| \cdot |X_{i_2 j} - X_{i_3 j}|. \tag{6}$$

The expression of $\hat{\theta}_{j,h}$ for $h = 2, 3, 5, 7$ is similar to (5); the expression of $\hat{\theta}_{j,h}$ for $h = 6, 8$ is similar to (6).

Remark: When Pearson correlation is used, the aggregated estimator $\widetilde{\omega}_j$ in (4) coincides with the centralized estimator $\hat{\omega}_j$; the proposed ACS leads to the same screening result of the classic SIS. For the correlations in Examples 2-4, the computational cost of $\widetilde{\omega}_j$ is substantially lower than that of $\hat{\omega}_j$. For DC, when the data segments are parallel processed, ACS drastically reduces the cost of centralized screening from $O(pN^3)$ down to $O(pN^3/m^3)$; this is also numerically cheaper than the feature-splitting-based screening, whose cost is $O(pN^3/m)$ with each local computer evaluating p/m features based on $\hat{\omega}_j$s.

The idea of componentwise debiasing in ACS provides a viable and effective route to estimate ω_j in a distributed manner. For commonly-used correlation measures, form (2) can be naturally constructed. The simplicity and compatibility of ACS make it a user-friendly approach in practice.

2.3.2. Extension

When data partition is manually done, one may further improve the stability of ACS with multiple partitions. Specifically, suppose that we repeat the random data partition R times.

For each partition, we conduct unbiased estimation of component parameters based on (3). We then carry out (4) with $\bar{U}_{j,h}$ replaced by

$$\breve{U}_{j,h}^{R} = \frac{1}{R}\sum_{r=1}^{R} \bar{U}_{j,h}^{r},$$

where $\bar{U}_{j,h}^{r}$ denotes the mean U-statistic for the rth partition. By averaging over R partitions, the variability of $\widetilde{\omega}_j$ is further reduced; this leads to a reinforced ACS that is more reliable for feature screening.

When a correlation measure with large k_hs is used and m is constricted to be small, one may further split a data segment $\mathcal{D}_l$ into smaller and manageable sub-segments. One can then conduct U-statistic estimation for the component parameters sequentially based on the sub-segments and use an averaged quantity to replace $U_{j,h}^{l}$ in Step 2 of ACS. If needed, a reinforced version of this strategy can also be used with multiple sub-partitions.

3. Theoretical Analysis

We now provide some theoretical justification of using ACS. Apparently, the screening performance of ACS relies on the accuracy of the aggregated correlation estimator $\widetilde{\omega}_j$ (4). We show that $\widetilde{\omega}_j$ is an effective and efficient tool to estimate ω_j; this serves as a theoretical foundation of ACS. Our theoretical investigation is based on the following technical conditions.

C1 There exist two constants κ_0 and D_0 such that, for any $0 \le \kappa \le \kappa_0$, $E\{\exp(\kappa|\hat{\theta}_{j,h}|)\} < D_0$ for all $h = 1, ..., s$, $j = 1, ..., p$.

C2 In (2), $g(\cdot)$ is formed by finite operations of addition, subtraction, multiplication, division, absolutization, and square root, where the division and square root are taken over a quantity uniformly bounded away from zero.

C3 There exist two constants $c > 0$ and $0 < \tau < 1/2$ such that $\min_{j \in \mathcal{M}} \omega_j \ge 2cN^{-\tau}$.

Condition C1 requires that $|\hat{\theta}_{j,h}|$ has a regular distribution, such that its moment generating function exists on $[0, \kappa_0]$. This is a mild condition for many correlation measures. For example, when Kendall τ correlation is used with ACS, $\hat{\theta}_{j,h}$ is bounded and thus C1 is naturally satisfied; when SIRS is used with ACS, C1 is implied if $E\{\exp(\xi X_j^2)\} < D_0'$ for some $\xi > 0$, $D_0' > 0$ and $1 \le j \le p$. Condition C2 is applicable to a variety of commonly used correlation measures, including the ones discussed in Section 2.3.1. We conjecture that ACS would still be effective with a more complicated $g(\cdot)$. However, the corresponding theoretical justification is likely to be lengthy. Here, we aim to provide some theoretical understanding of the proposed screening framework and do not intend to make this condition weakest possible. Condition C3 requires that the marginal correlation between any relevant feature and the response should not be too small. This is a natural feature identifiability requirement, which has been widely used in the literature; see, for example, Condition 3 of Fan and Lv (2008), Condition 2 of Li et al. (2012b), and Condition 6 of Wu and Yin (2015).

With the conditions above, we first show that the averaged local U-statistics $\bar{U}_{j,h}$s enjoy the properties stated in the following proposition.

Proposition 1 *Under Condition C1, we have*

$$\max_{1\le j\le p, 1\le h\le s} \mathrm{Var}(\bar{U}_{j,h}) = O(\frac{1}{N}) + O(\frac{m}{N^2}) + ... + O(\frac{m^{k_h-1}}{N^{k_h}}), \tag{7}$$

which is increasing in m.

By (7), we see that the largest $\mathrm{Var}(\bar{U}_{j,h})$ is in the order of $O(N^{-1})$ for both fixed m and diverging m. While a larger m leads to an increased variance of $\bar{U}_{j,h}$, such an information loss is insignificant in the sense that the asymptotic order of (7) remains unchanged. Since $\bar{U}_{j,h}$s are unbiased, the high precision implies the uniform second moment consistency, which echoes Theorem 2 of Lin and Xi (2010). Proposition 1 indicates that the component parameters can be effectively estimated by summarizing the corresponding local U-statistics from data segments. With the properties of $\bar{U}_{j,h}$, we show the effectiveness of $\widetilde{\omega}_j$ in the following two theorems.

Theorem 2 *Suppose that Conditions C1-C3 are satisfied and $k = \max\{k_h, h = 1, \ldots, s\} \le n$. There exists a constant $\eta > 0$ such that*

$$P\left(\max_{1\le j\le p} |\widetilde{\omega}_j - \omega_j| \ge cN^{-\tau}\right) \le \eta p(1 - N^{-2\tau}/\eta)^{m\lfloor n/k\rfloor},$$

where $\lfloor n/k\rfloor$ denotes the largest integer no larger than n/k.

Note that k is a constant depending on the choice of ω_j and $g(\cdot)$; thus, $m\lfloor n/k\rfloor$ is in the same order of N. Theorem 2 implies that the aggregated correlation estimators are uniformly consistent even when p grows exponentially with N^α for some $0 < \alpha < 1$. In the literature, it has been shown that the centralized estimator achieves convergence bound $|\hat{\omega}_j - \omega_j| = O_p(N^{-\tau})$ for $0 < \tau < 1/2$ (Li et al., 2012a,b; Cui et al., 2015; Wu and Yin, 2015). Theorem 2 indicates that $\widetilde{\omega}_j$ works as efficiently as the centralized estimator $\hat{\omega}_j$ in terms of the probability convergence bound.

For the correlation measures admitting a Lipschitz continuous $g(\cdot)$, we further justify the corresponding $\widetilde{\omega}_j$ in terms of MSE. To be specific, recall that we express a correlation measure $\omega_j = g(\boldsymbol{\theta}_j)$ with $\boldsymbol{\theta}_j = (\theta_{j,1}, ..., \theta_{j,s})$ from a compact space $\boldsymbol{\Theta} \subset R^s$. We say $g(\cdot)$ is Lipschitz continuous in $\boldsymbol{\theta}_j$, if the following condition is satisfied.

C2′ There exists a positive constant L such that $|g(\boldsymbol{\theta}_j) - g(\boldsymbol{\theta}'_j)| \le L\|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j\|$ for any $\boldsymbol{\theta}_j, \boldsymbol{\theta}'_j \in \boldsymbol{\Theta}$, where $\|\cdot\|$ denotes the Euclidean norm.

Condition C2′ can be naturally satisfied by many correlation measures. For example, as shown in Section 2.3.1, we have $g(\theta_{j,1}) = |\theta_{j,1} - 1/4|$ for Kendall τ correlation and $g(\theta_{j,1}) = \theta_{j,1}$ for SIRS, where both $g(\cdot)$s are Lipschitz continuous. When component parameters

appear in the denominator in $g(\cdot)$, Condition C2′ is often satisfied with mild requirements on $(Y, \mathbf{X})$ such that the denominator is uniformly bounded away from zero. For example, when Pearson correlation is used, Condition C2′ is satisfied when $\mathrm{Var}(Y) > 0$ and $\mathrm{Var}(X_j) > c$ for $j = 1, \ldots, p$ with some $c > 0$; when DC is used, Condition C2′ is satisfied when the distance variances of X_j and Y defined in Li et al. (2012b) are all bounded away from zero. In fact, when the parameter space $\mathbf{\Theta}$ is compact, Condition C2′ can be implied by Condition C2, the proof of which is similar to Lemma 6 in the Appendix. Based on this condition, we derive a uniform MSE bound for $\widetilde{\omega}_j$s in the following theorem.

Theorem 3 *Under Conditions C1 and C2', if $k \leq n$ and $\bar{U}_{j,h} \in \mathbf{\Theta}$, we have*

$$\max_{1 \leq j \leq p} \mathrm{MSE}(\widetilde{\omega}_j) = E(\widetilde{\omega}_j - \omega_j)^2 = O(1/N).$$

Theorem 3 shows that, when $g(\cdot)$ is smooth enough, the aggregated correlation estimator $\widetilde{\omega}_j$ matches the optimal MSE rate achievable by a centralized estimator having access to the entire data of size N. This result also indicates that using the aggregation of U-statistic component estimates is an effective way in reducing the bias of the final correlation estimator. Different from the existing debiasing techniques developed for the distributed M-estimation (e.g., Zhang et al., 2012; Battey et al., 2018), the idea of componentwise debiasing is built upon the natural composition of a centralized correlation estimator; instead of estimating the bias of each $\hat{\omega}_{l,j}$ in (1) based on $\mathcal{D}_l$, it addresses the debiasing task via improving the component estimators, where the distributed U-statistic can be conveniently used. Benefited from the high precision of $\bar{U}_{j,h}$, $\widetilde{\omega}_j$ enjoys a full estimation efficiency that is insensitive to the choice of m; this further leads to a reliable feature screening.

Admittedly, $\widetilde{\omega}_j$ can be still biased due to the non-linear aggregation in (4). For the commonly-used correlation measures, function g is smooth and the number of component parameters s is finite. Consequently, aggregating those $\hat{\omega}_{l,j}$s via g is unlikely to bring a significant bias to $\widetilde{\omega}_j$. In fact, Theorem 3 immediately implies that $\mathrm{Bias}(\widetilde{\omega}_j) = O(1/\sqrt{N})$ regardless of the choice of m. In comparison, when SAS is used in our setup, we have $\mathrm{Bias}(\bar{\omega}_j) = \mathrm{Bias}(\hat{\omega}_{l,j})$, the scale of which is mainly determined by the amount of data $n = N/m$ stored in segment $\mathcal{D}_l$. For example, when SIRS is used without local standardization, it can be shown that $\mathrm{Bias}(\hat{\omega}_{l,j}) = O(1/n) = O(m/N)$ (Zhu et al., 2011). When m is large, this bias can severely affect its accuracy. In particular, when $m/\sqrt{N} \to \infty$, $\mathrm{MSE}(\bar{\omega}_j)$ has a rate slower than $O(1/N)$.

Next, we justify the proposed ACS framework using the following theorem.

Theorem 4 *Under Conditions C1-C3, if $k \leq n$ and $\gamma = cN^{-\tau}$, then there exists a constant $\eta > 0$ such that*

$$P\{\mathcal{M} \subseteq \widetilde{\mathcal{M}}\} \geq 1 - \eta d(1 - N^{-2\tau}/\eta)^{m\lfloor n/k \rfloor},$$

where d is the cardinality of $\mathcal{M}$.

We show in the Appendix that, when $d = O(N)$, the probability bound in Theorem 4 goes to one as $N \to \infty$. Thus, the proposed ACS enjoys sure screening property in the sense

of Fan and Lv (2008), even when the number of relevant features d is diverging. That is, when N is large, ACS removes most irrelevant features and retains all relevant features with an overwhelming probability. It is a desired property for a good feature screening method. Note that the requirement $n = N/m \geq k$ is very mild in general; for many correlation measures, it can be naturally satisfied with a liberal choice of $m = O(N)$, which makes ACS a flexible and reliable approach. Although the asymptotic error bound of $\widetilde{\omega}_j$ is less affected by m, our empirical experience does show that a small m may help to improve the practical screening accuracy of ACS. However, an overly small m often leads to a high computational cost. In applications, one good strategy is to choose the smallest m for ACS within the computational budget.

4. Numerical Studies

We assess the finite sample performance of ACS via simulations and a real data example. In particular, we compare ACS with the naive SAS in terms of the screening accuracy and stability. All numerical experiments are conducted using software MATLAB on Windows computers with 3.2 GHz CPUs and 32 GB memory.

4.1. Simulations

4.1.1. Example 1

Apparently, an effective screening relies on the accurate estimates of the correlation strength ω_j. Our first experiment is to check whether the proposed aggregated correlation (AC) measure $\widetilde{\omega}_j$ in (4) is an effective estimator of ω_j. To this end, we generate $N = 2700$ independent copies from (Y, X), where Y and X are two independent random variables following $N(0, 1)$. Due to independence, the Kendall τ correlation, SIRS, and DC between Y and X are all zero. We randomly split the data into $m = 45, 90, 180$ equal-sized segments and use $\widetilde{\omega}_j$ specified in Section 2.3.1 (with $j = 1$) to estimate the three aforementioned correlations between Y and X. We repeat the procedure $T = 500$ times and measure the accuracy of $\widetilde{\omega}_j$ by root-mean-squared error (RMSE). Specifically, let $\widetilde{\omega}_j(t)$ denote the value of $\widetilde{\omega}_j$ for the tth repetition. RMSE is computed by

$$\mathrm{RMSE}(\widetilde{\omega}_j) = \left[\frac{1}{T}\sum_{t=1}^{T}(\widetilde{\omega}_j(t))^2\right]^{1/2}.$$

For comparison, we report the corresponding RMSEs of the simple averaging (SA) estimators $\bar{\omega}_j$ defined in (1) under the same m setup. Moreover, we check the performance of the reinforced $\widetilde{\omega}_j$ (rAC) using the multiple partition strategy with $R = 3$ as discussed in Section 2.3.2. As a benchmark, we also report the RMSEs of the centralized estimators with $m = 1$. The results are summarized in Figure 1 with the corresponding computational time (in seconds) given in Table 1.

For all the three tested correlations, we see that both $\widetilde{\omega}_j$ and $\bar{\omega}_j$ work well when m is small. As m increases, $\bar{\omega}_j$ becomes less accurate. As discussed, this is mainly due to the non-negligible biases of the segmental estimates. In comparison, $\widetilde{\omega}_j$ conducts componentwise debiasing and leads to a high estimation accuracy over a wide range of m. Compared

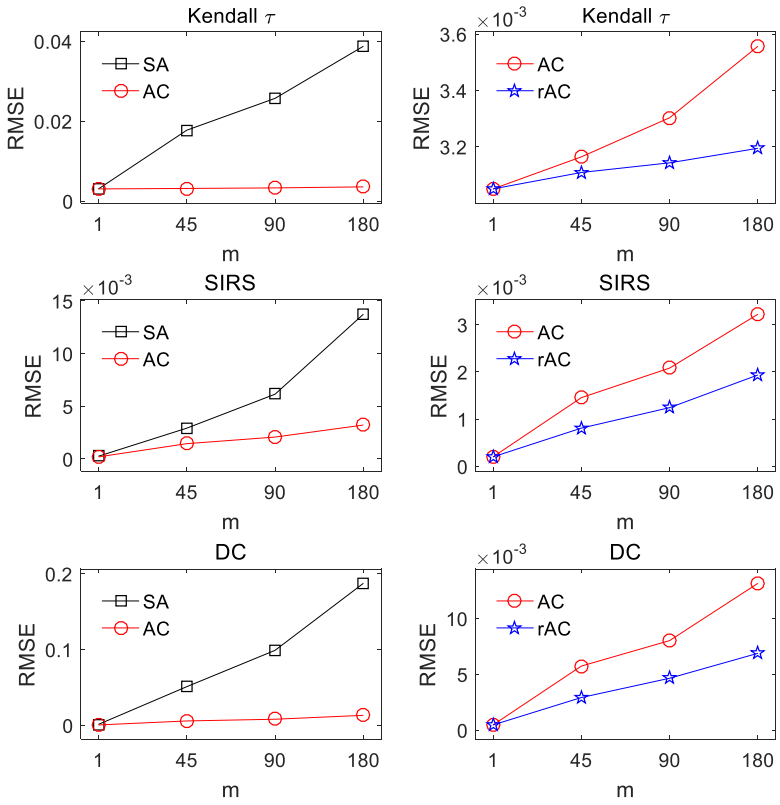

Figure 1: The RMSE of distributed correlation estimators with $m = (1, 45, 90, 180)$, where SA, AC, and rAC stand for $\bar{\omega}_j$, $\widetilde{\omega}_j$, and reinforced $\widetilde{\omega}_j$ respectively.

with the centralized estimators ($m = 1$ case), the distributed estimators $\widetilde{\omega}_j$ and $\bar{\omega}_j$ are computationally more attractive, in particular when m is large. As expected, the reinforced aggregated estimators help to further improve the estimation accuracy of $\widetilde{\omega}_j$ at a higher computational cost.

4.1.2. Example 2

The promising performance of $\widetilde{\omega}_j$ encourages us to further check whether the associated screening procedure ACS also works well. To this end, we generate N independent copies of $\mathbf{X} = (X_1, \ldots, X_p)$ from a multivariate normal distribution with zero mean. The corresponding response Y is generated based on the following models.

$$
\begin{aligned}
&\text{(a)} \quad Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_8 X_8 + \varepsilon, \\
&\text{(b)} \quad Y = \beta_1 X_1 + \beta_2 X_4 + \beta_3 X_7 + \beta_4 X_{10} + \varepsilon, \\
&\text{(c)} \quad Y = \exp(\beta_1 X_1 + \beta_2 X_4 + \beta_3 X_7 + \beta_4 X_{10} + \varepsilon), \\
&\text{(d)} \quad Y = \beta_1 X_1 + \beta_2 X_4 + \exp(|\beta_3| X_7 + |\beta_4| X_{10}) + \varepsilon, \\
&\text{(e)} \quad Y = \beta_1 X_1 + \beta_2 X_4^2 + \beta_3 I(X_7 > 0) + \beta_4 |X_{10}| + \varepsilon, \\
&\text{(f)} \quad Y = 2\beta_1 X_1 X_2 + 2\beta_2 I(X_{12} > 0) + 3\beta_3 X_{22} + \varepsilon,
\end{aligned}
$$

Table 1: Mean computational time of distributed correlation estimators (in seconds) with fixed sample size $N = 2700$ and varied number of data segments; $m = 1$ case corresponds to the centralized estimates and $m > 1$ cases correspond to the distributed estimates.

Correlation	Estimator	$m = 1$	$m = 45$	$m = 90$	$m = 180$
Kendall τ	SA	$3.4 \cdot 10^{-1}$	$1.4 \cdot 10^{-4}$	$4.3 \cdot 10^{-5}$	$2.2 \cdot 10^{-5}$
	AC	$3.4 \cdot 10^{-1}$	$1.5 \cdot 10^{-4}$	$4.7 \cdot 10^{-5}$	$2.6 \cdot 10^{-5}$
	rAC	$--$	$3.7 \cdot 10^{-4}$	$1.2 \cdot 10^{-4}$	$5.5 \cdot 10^{-5}$
SIRS	SA	$1.7 \cdot 10^{-1}$	$1.1 \cdot 10^{-4}$	$5.0 \cdot 10^{-5}$	$4.0 \cdot 10^{-5}$
	AC	$1.7 \cdot 10^{-1}$	$7.5 \cdot 10^{-5}$	$2.5 \cdot 10^{-5}$	$1.4 \cdot 10^{-5}$
	rAC	$--$	$2.0 \cdot 10^{-4}$	$7.0 \cdot 10^{-5}$	$3.9 \cdot 10^{-5}$
DC	SA	$8.2 \cdot 10^{-1}$	$1.4 \cdot 10^{-4}$	$4.7 \cdot 10^{-5}$	$3.3 \cdot 10^{-5}$
	AC	$7.8 \cdot 10^{-1}$	$1.3 \cdot 10^{-4}$	$4.0 \cdot 10^{-5}$	$2.7 \cdot 10^{-5}$
	rAC	$--$	$3.6 \cdot 10^{-4}$	$1.2 \cdot 10^{-4}$	$9.3 \cdot 10^{-5}$

where $\varepsilon \sim N(0, 1)$ is a noise term. Models (a) and (b) are two linear cases with different model sparsity and covariance structures. Models (c) and (d) are transformation model and multiple-index model, which are adopted from Li et al. (2012a) and Zhu et al. (2011) respectively. Models (e) and (f) are addictive model and interactive model, both of which were discussed in Li et al. (2012b). In Model (a), $\text{cov}(\mathbf{X})$ is set to be an identity matrix, while in Models (b)-(f) we set $\text{cov}(X_j, X_r) = 0.5^{|j-r|}$ for $j, r \in \{1, \ldots, p\}$ such that the features have an autoregressive correlation. In Models (a)-(f), the values of model coefficients are generated by $(-1)^W(2 + |V|)$, where $W \sim \text{Bernoulli}(0.6)$ and $V \sim N(0, 1)$.

We apply the proposed ACS on these simulated data sets for feature screening. In each case, we split the data into m segments and assess the performance of ACS based on Pearson, Kendall τ, SIRS, and DC correlations as discussed in Section 2.3.1. For each correlation scenario, we set the corresponding screening threshold by

$$\gamma = \rho \cdot \min_{j \in \mathcal{M}} \hat{\omega}_j, \tag{8}$$

where $\hat{\omega}_j$ is the centralized estimator of that correlation and $\rho = 0.8, 0.6$ is a scale parameter. The choice of γ in (8) guarantees that all relevant features will be retained by the classic screening method based on $\hat{\omega}_j$; it purely serves for evaluating the proposed distributed method under the setup, in which the classic method is likely to work well. In practice, a proper γ is usually determined by users based on their research goals as well as the prior information about their data. We will test ACS in our next example with a data-driven choice of γ.

We evaluate the performance of ACS in terms of successful screening rate (SSR), screened model size (MS), positive selection rate (PSR), false discovery rate (FDR), Specifically, let $\hat{\mathcal{M}}(t)$ denote the index set of the features retained after screening based on the t-th repeti-

Table 2: Simulation results for Model (a) with $N = 2400$, $p = 10000$, $\|\mathcal{M}\|_0 = 8$, $m = (40, 60)$. The two values a, b in the same column correspond to $\rho = 0.8, 0.6$ cases.

m	Correlation	Method	SSR	MS	Std(MS)	PSR	FDR	Time^n	Time^N
40	Pearson	SAS	1.0, 1.0	8, 9	5, 531	1.0, 1.0	0.0, .11	0.012	0.176
		SVS	.97, 1.0	8, 13	6, 256	1.0, 1.0	0.0, .36	0.012	
		ACS	1.0, 1.0	8, 8	0, 0	1.0, 1.0	0.0, 0.0	0.012	
	Kendall τ	SAS	1.0, 1.0	8, 18	64, 935	1.0, 1.0	0.0, .56	0.678	1504
		SVS	.97, 1.0	8, 28	34, 507	1.0, 1.0	0.0, .71	0.678	
		ACS	1.0, 1.0	8, 8	0, 0	1.0, 1.0	0.0, 0.0	0.678	
		rACS	1.0, 1.0	8, 8	0, 0	1.0, 1.0	0.0, 0.0	2.030	
	SIRS	SAS	1.0, 1.0	10, 1042	1536, 3388	1.0, 1.0	.20, .99	0.013	1.672
		SVS	.99, 1.0	8, 34	781, 1540	1.0, 1.0	0.0, .76	0.013	
		ACS	.69, .93	45, 176	122, 229	1.0, 1.0	.83, .95	0.013	
		rACS	.86, .99	8, 10	13, 46	1.0, 1.0	0.0, .20	0.033	
	DC	SAS	1.0, 1.0	9996, 10000	2807, 715	1.0, 1.0	.99, .99	0.742	6625
		SVS	1.0, 1.0	9491, 10000	3712, 1076	1.0, 1.0	.99, .99	0.742	
		ACS	.96, 1.0	8, 8	3, 20	1.0, 1.0	0.0, 0.0	0.742	
		rACS	.99, 1.0	8, 8	0, 0	1.0, 1.0	0.0, 0.0	2.223	
60	Pearson	SAS	1.0, 1.0	8, 156	192, 1753	1.0, 1.0	0.0, .95	0.011	0.176
		SVS	1.0, 1.0	8, 80	60, 840	1.0, 1.0	0.0, .90	0.011	
		ACS	1.0, 1.0	8, 8	0, 0	1.0, 1.0	0.0, 0.0	0.011	
	Kendall τ	SAS	1.0, 1.0	8, 1293	675, 2772	1.0, 1.0	0.0, .99	0.329	1504
		SVS	.98, 1.0	8, 313	265, 1429	1.0, 1.0	0.0, .97	0.329	
		ACS	1.0, 1.0	8, 8	0, 0	1.0, 1.0	0.0, 0.0	0.329	
		rACS	1.0, 1.0	8, 8	0, 0	1.0, 1.0	0.0, 0.0	0.982	
	SIRS	SAS	1.0, 1.0	4096, 9916	3927, 2392	1.0, 1.0	.99, .99	0.010	1.672
		SVS	1.0, 1.0	85, 4036	2479, 3752	1.0, 1.0	.90, .99	0.010	
		ACS	.61, .85	124, 378	200, 327	1.0, 1.0	.94, .98	0.010	
		rACS	.88, .98	8, 22	36, 94	1.0, 1.0	0.0, .63	0.025	
	DC	SAS	1.0, 1.0	10000, 10000	158, 0	1.0, 1.0	.99, .99	0.479	6625
		SVS	1.0, 1.0	10000, 10000	785, 0	1.0, 1.0	.99, .99	0.479	
		ACS	.93, 1.0	8, 8	9, 37	1.0, 1.0	0.0, 0.0	0.479	
		rACS	.98, 1.0	8, 8	0, 2	1.0, 1.0	0.0, 0.0	1.434	

tion. The aforementioned four indices are calculated as follows.

$$\text{SSR} = \frac{1}{T}\sum_{t=1}^{T} I_{\{\mathcal{M}\subset\hat{\mathcal{M}}(t)\}}, \quad \text{MS} = \left\lfloor \|\hat{\mathcal{M}}(t)\|_0 \right\rfloor_{med},$$

$$\text{PSR} = \left\lfloor \frac{\|\mathcal{M}\cap\hat{\mathcal{M}}(t)\|_0}{\|\mathcal{M}\|_0} \right\rfloor_{med}, \quad \text{FDR} = \left\lfloor \frac{\|\hat{\mathcal{M}}(t)-\mathcal{M}\|_0}{\|\hat{\mathcal{M}}(t)\|_0} \right\rfloor_{med},$$

where $I_{\{\cdot\}}$ is an indicator function, $\lfloor\cdot\rfloor_{med}$ denotes the median of a series of values, and $\|\cdot\|_0$ denotes the number of elements in a set. For comparison, we report as well the screening outcomes of SAS, which is based on the simple averaging estimators (1). Also, we compare ACS with a subsampling-voting screening (SVS) method. In SVS, parallel screening is first

Table 3: Simulation results for Model (b) with $N = 1200$, $p = 1500$, $\|\mathcal{M}\|_0 = 4$, $m = (20, 40)$.

m	Correlation	Method	SSR	MS	Std(MS)	PSR	FDR	Time^n	Time^N
20	Pearson	SAS	1.0, 1.0	6, 9	19, 109	1.0, 1.0	.33, .53	0.001	0.022
		SVS	.96, 1.0	6, 9	16, 66	1.0, 1.0	.33, .53	0.001	
		ACS	1.0, 1.0	6, 9	2, 2	1.0, 1.0	.33, .53	0.001	
	Kendall τ	SAS	1.0, 1.0	6, 9	37, 132	1.0, 1.0	.33, .53	0.105	51.20
		SVS	.95, .98	6, 9	26, 85	1.0, 1.0	.33, .56	0.105	
		ACS	1.0, 1.0	6, 8	2, 2	1.0, 1.0	.33, .50	0.105	
	SIRS	SAS	1.0, 1.0	6, 8	203, 217	1.0, 1.0	.33, .50	0.001	0.111
		SVS	.97, .99	6, 7	149, 200	1.0, 1.0	.33, .43	0.001	
		ACS	.89, .98	6, 9	38, 50	1.0, 1.0	.33, .56	0.001	
	DC	SAS	1.0, 1.0	8, 10	310, 510	1.0, 1.0	.50, .60	0.122	105.3
		SVS	1.0, 1.0	7, 9	281, 419	1.0, 1.0	.43, .56	0.122	
		ACS	.98, .99	5, 7	6, 15	1.0, 1.0	.20, .43	0.122	
40	Pearson	SAS	1.0, 1.0	6, 9	140, 219	1.0, 1.0	.33, .56	0.001	0.022
		SVS	.97, .99	7, 9	89, 181	1.0, 1.0	.43, .56	0.001	
		ACS	1.0, 1.0	6, 9	2, 2	1.0, 1.0	.33, .53	0.001	
	Kendall τ	SAS	1.0, 1.0	7, 9	171, 263	1.0, 1.0	.43, .56	0.035	51.20
		SVS	0.99, 1.0	7, 10	123, 210	1.0, 1.0	.43, .60	0.035	
		ACS	1.0, 1.0	6, 9	2, 2	1.0, 1.0	.33, .53	0.035	
	SIRS	SAS	1.0, 1.0	8, 11	359, 509	1.0, 1.0	.50, .64	0.001	0.111
		SVS	1.0, 1.0	7, 9	248, 361	1.0, 1.0	.43, .56	0.001	
		ACS	.81, .92	8, 21	59, 75	1.0, 1.0	.56, .80	0.001	
	DC	SAS	1.0, 1.0	89, 1466	668, 646	1.0, 1.0	.95, .99	0.038	105.3
		SVS	1.0, 1.0	19, 1116	606, 667	1.0, 1.0	.78, .99	0.038	
		ACS	.93, .99	6, 7	19, 31	1.0, 1.0	.29, .43	0.038	

conducted on m random subsamples of $\mathcal{D}$, each of which is of size $n = N/m$; a feature is then retained if it is selected by more than 50% of the subsamples. To check the improving strategy in Section 2.3.2, we further run the reinforced ACS (rACS) with $R = 3$ for the data generated from Model (a). In our setup, the classic screening method based on $\hat{\omega}_j$ would have SSR = 1, PSR = 1 in all cases and is likely to have small MS and FDR as N is large. Yet, as to be revealed, its computational cost can be unaffordable in many cases. For the computational convenience, we choose to exclude the classic screening from the comparison and use directly the optimal values of the aforementioned indices as a benchmark.

We summarize the simulation results in Tables 2-4 based on $T = 100$ repetitions. For Models (c)-(f), we only exhibit the selected results due to the page limit. In the tables, Time^n and Time^N report the averaged computational time (in seconds) respectively for a distributed screening and the corresponding classic screening based on a few pilot runs. The two values in the same column correspond to the two setups of ρ in (8). Std(MS) reports the sample standard deviation of $\|\hat{\mathcal{M}}(t)\|_0$s, which measures the screening precision.

With the oracle choice of γ, we see that all methods perform well in terms of keeping relevant features; this is indicated by their high SSRs in most cases. Regarding the screening

Table 4: Simulation results for Models (c)-(f) with case-specific (N, p, m) setup listed in the table.

m	Correlation	Method	SSR	MS	Std(MS)	PSR	FDR	Time^n	Time^N
Model (c), $N = 2400$, $p = 10000$, $\|\mathcal{M}\|_0 = 4$									
40	Pearson	SAS	1.0, 1.0	10000,10000	0, 0	1.0, 1.0	.99, .99	0.007	0.166
		ACS	1.0, 1.0	1825, 3207	2333, 2350	1.0, 1.0	.99, .99	0.007	
	Kendall τ	SAS	1.0, 1.0	6, 9	2, 141	1.0, 1.0	.33, .56	0.666	1429.2
		ACS	1.0, 1.0	6, 8	2, 2	1.0, 1.0	.33, .50	0.666	
80	Pearson	SAS	1.0, 1.0	10000, 10000	0, 0	1.0, 1.0	.99, .99	0.005	0.166
		ACS	1.0, 1.0	1825, 3207	2333, 2350	1.0, 1.0	.99, .99	0.005	
	Kendall τ	SAS	1.0, 1.0	7, 10	343, 2439	1.0, 1.0	.43, .60	0.220	1429.2
		ACS	1.0, 1.0	6, 8	2, 2	1.0, 1.0	.33, .50	0.220	
Model (d), $N = 3600$, $p = 10000$, $\|\mathcal{M}\|_0 = 4$									
50	Pearson	SAS	1.0, 1.0	10000, 10000	0, 0	1.0, 1.0	.99, .99	0.009	0.259
		ACS	1.0, 1.0	7478, 8128	2473, 2049	1.0, 1.0	.99, .99	0.009	
	SIRS	SAS	1.0, 1.0	8, 10	270, 1377	1.0, 1.0	.50, .60	0.015	3.393
		ACS	.95, 1.0	7, 9	22, 66	1.0, 1.0	.43, .56	0.015	
100	Pearson	SAS	1.0, 1.0	10000, 10000	0, 0	1.0, 1.0	.99, .99	0.006	0.259
		ACS	1.0, 1.0	7478, 8128	2473, 2049	1.0, 1.0	.99, .99	0.006	
	SIRS	SAS	1.0, 1.0	11, 99	3167, 4192	1.0, 1.0	.64, .96	0.009	3.393
		ACS	.83, .92	8, 13	79, 177	1.0, 1.0	.50, .69	0.009	
Model (e), $N = 4800$, $p = 10000$, $\|\mathcal{M}\|_0 = 4$									
60	Pearson	SAS	1.0, 1.0	10000, 10000	0, 0	1.0, 1.0	.99, .99	0.010	0.334
		ACS	1.0, 1.0	1508, 2807	2596, 2578	1.0, 1.0	.99, .99	0.010	
	DC	SAS	1.0, 1.0	10000, 10000	1813, 1188	1.0, 1.0	.99, .99	1.147	26221
		ACS	.89, .95	6, 9	72, 162	1.0, 1.0	.33, .53	1.147	
120	Pearson	SAS	1.0, 1.0	10000, 10000	0, 0	1.0, 1.0	.99, .99	0.006	0.334
		ACS	1.0, 1.0	1508, 2807	2596, 2578	1.0, 1.0	.99, .99	0.006	
	DC	SAS	1.0, 1.0	10000, 10000	0, 0	1.0, 1.0	.99, .99	0.489	26221
		ACS	.86, .94	11, 73	175, 317	1.0, 1.0	.65, .95	0.489	
Model (f), $N = 10000$, $p = 10000$, $\|\mathcal{M}\|_0 = 4$									
100	Pearson	SAS	1.0, 1.0	10000, 10000	0, 0	1.0, 1.0	.99, .99	0.016	0.729
		ACS	1.0, 1.0	8600, 8950	3365, 2946	1.0, 1.0	.99, .99	0.016	
	DC	SAS	1.0, 1.0	10000, 10000	1809, 0	1.0, 1.0	.99, .99	1.923	213998
		ACS	.98, 1.0	8, 8	1, 2	1.0, 1.0	.50, .50	1.923	
250	Pearson	SAS	1.0, 1.0	10000, 10000	0, 0	1.0, 1.0	.99, .99	0.017	0.729
		ACS	1.0, 1.0	8600, 8950	3365, 2946	1.0, 1.0	.99, .99	0.017	
	DC	SAS	1.0, 1.0	10000, 10000	0, 0	1.0, 1.0	.99, .99	0.441	213998
		ACS	.85, 1.0	8, 10	9, 50	1.0, 1.0	.50, .60	0.441	

accuracy, SAS seems to be inferior, as it tends to keep too many irrelevant features after screening. This phenomenon is particularly severe for non-linear correlation measures SIRS and DC under the large-m-small-γ setup. As an extreme case, when DC is used in Models (e) and (f), SAS suggests keeping all the 10000 features; this completely fails in the mission of screening. The over-selection of SAS here is a direct result from the inaccuracy of the

corresponding simple averaging estimators $\bar{\omega}_j$s. When Pearson and Kendall τ correlations are used, this issue is less severe, as the corresponding $\bar{\omega}_j$s are less biased due to their nature. The proposed ACS, in comparison, is built upon the stable $\widetilde{\omega}_j$s, and thus achieves a reasonably high screening accuracy in most setups. For all the four correlation choices, it is able to screen most irrelevant features out, while keep relevant ones with a high probability. Such a performance is very promising.

Moreover, the SAS-based screening tends to have a high variability in most cases; this makes it less trustable in practice. In Models (a) and (b), SVS seems to help reducing the variability of SAS, but the improvement is less significant as compared with ACS or rACS. This is somewhat expected, as both SAS and SVS are based on the local estimators $\hat{\omega}_{l,j}$, the bias of which can be substantial when m is large.

We observe that, when SIRS is used in Model (a) with $m = 60$, none of SAS, SVS and ACS works very well, if SSR and Std(MS) are considered jointly. This might be due to the relatively low sensitivity of SIRS in detecting linear correlations when n is small. When $m = 40$, all methods get improved due to the increased accuracy in both $\bar{\omega}_j$ and $\widetilde{\omega}_j$. Apparently, the multiple data partition strategy in rACS helps a lot in this case, as indicated by its high SSR and low Std(MS) in both m setups.

We also observe that, with Pearson correlation, neither ACS nor SAS performs satisfactorily in Models (c)-(f). This is because relevant features in those models mainly have non-linear correlations with the response. The superior screening accuracy of ACS is observed when advanced correlation measures are used in those models.

Benefited from its distributed framework, the proposed ACS enables parallel computing and enjoys a great numerical advantage over the classic screening procedures (i.e., $m = 1$ case). As shown in Tables 2-4, the computational cost of ACS can be even less than 1% of the traditional cost with a large m setup, while it still maintains relatively high screening accuracy. This merit together with its broad compatibility makes ACS an attractive approach for screening with large-N-large-p data.

4.1.3. Example 3

We further test ACS when the screening threshold γ is chosen by a data-driven method. Similar to Zhu et al. (2011), after the training data are generated, we additionally generate N independent copies of a q-dimensional auxiliary variable $(X'_1, \ldots, X'_q)^T$ from a standard multivariate normal distribution. Note that X'_js are independent of the response Y and thus are all irrelevant. A reasonable screening threshold is therefore set to be

$$\gamma = \max_{j=1,\ldots,q} \widetilde{\omega}'_j,$$

where $\widetilde{\omega}'_j$ is the aggregated correlation estimator between Y and X'_j.

We repeat our simulation in Model (b) of Example 2 with the new choice of γ and report the results in Table 5, where $N = 2400$, $p = 5000$, $m = (80, 100)$, and $q = (1000, 500)$. It is seen that the performances of ACS and rACS with $R = 5$ are still sharp with the non-oracle γ. In fact, the significance of the proposed method is even better observed in this example, as SAS seems to be more sensitive to γ.

Table 5: Simulation results of Example 3: data are generated from Model (b) with $N = 2400$, $p = 5000$, $\|\mathcal{M}\|_0 = 4$; distributed screening is conducted based on a data-driven γ with $m = (80, 100)$. The two values a, b in the same column correspond to $q = (1000, 500)$ cases.

m	Correlation	Method	SSR	MS	Std(MS)	PSR	FDR	Time^n	Time^N
80	Pearson	SAS	1.0, 1.0	5000, 5000	0, 0	1.0, 1.0	.99, .99	0.003	0.086
		ACS	1.0, 1.0	15, 20	6, 12	1.0, 1.0	.73, .80	0.003	
	Kendall τ	SAS	1.0, 1.0	5000, 5000	0, 0	1.0, 1.0	.99, .99	0.112	738.0
		ACS	1.0, 1.0	15, 21	5, 10	1.0, 1.0	.73, .81	0.112	
		rACS	1.0, 1.0	14, 17	4, 7	1.0, 1.0	.71, .76	0.559	
	SIRS	SAS	1.0, 1.0	17, 74	368, 769	1.0, 1.0	.76, .95	0.005	0.850
		ACS	.84, .84	11, 13	6, 9	1.0, 1.0	.64, .71	0.005	
		rACS	.86, .88	7, 8	2, 2	1.0, 1.0	.43, .50	0.024	
	DC	SAS	1.0, 1.0	5000, 5000	0, 0	1.0, 1.0	.99, .99	0.146	3240.6
		ACS	.99, 1.0	12, 15	7, 12	1.0, 1.0	.67, .73	0.146	
		rACS	1.0, 1.0	9, 9	2, 2	1.0, 1.0	.56, .56	0.728	
100	Pearson	SAS	1.0, 1.0	5000, 5000	0, 0	1.0, 1.0	.99, .99	0.002	0.086
		ACS	1.0, 1.0	15, 19	5, 11	1.0, 1.0	.73, .79	0.002	
	Kendall τ	SAS	1.0, 1.0	5000, 5000	0, 0	1.0, 1.0	.99, .99	0.086	738.0
		ACS	1.0, 1.0	15, 18	5, 9	1.0, 1.0	.73, .78	0.086	
		rACS	1.0, 1.0	13, 15	4, 6	1.0, 1.0	.69, .73	0.428	
	SIRS	SAS	1.0, 1.0	448, 1404	1183, 1738	1.0, 1.0	.99, .99	0.004	0.850
		ACS	.80, .82	11, 15	6, 9	1.0, 1.0	.64, .75	0.004	
		rACS	.83, .85	7, 7	2, 2	1.0, 1.0	.43, .43	0.020	
	DC	SAS	1.0, 1.0	5000, 5000	0, 0	1.0, 1.0	.99, .99	0.128	3240.6
		ACS	.98, .98	12, 15	8, 13	1.0, 1.0	.67, .75	0.128	
		rACS	1.0, 1.0	9, 9	2, 2	1.0, 1.0	.56, .56	0.640	

4.1.4. Example 4

In the previous examples, we have observed the promising performance of ACS on a few parametric models. We now extend our numerical assessment on ACS to a model-free learning framework, where the true model is unknown or may not even exist.

To be specific, with a given data set $\mathcal{D} = \{(Y_i, \mathbf{X}_i)\}_{i=1}^N$, we consider a kernel ridge regression (KRR), where the goal is to find a predictive function $\hat{f}$ by minimizing

$$\hat{f} = \arg\min_f \left\{ \frac{1}{N} \sum_{i=1}^{N} (Y_i - f(\mathbf{X}_i))^2 + \lambda \|f\|_K^2 \right\}, \tag{9}$$

where f has the form

$$f(\mathbf{X}) = \sum_{j=1}^{N} \beta_j K(\mathbf{X}, \mathbf{X}_j), \tag{10}$$

$\|f\|_K^2 = \sum_{i,j=1}^N \beta_i \beta_j K(\mathbf{X}_i, \mathbf{X}_j)$ is the norm of f induced by a semi-positive definite kernel function K, and $\lambda > 0$ is a tuning parameter. Let $\mathbf{K} = \{K(\mathbf{X}_i, \mathbf{X}_j),\ i, j = 1, \ldots, N\}$ be the working kernel matrix associated with K. The coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)^T$ of $\hat{f}$ in

form of (10) is estimated by

$$\hat{\boldsymbol{\beta}} = (\mathbf{K} + N\lambda \mathbf{I}_N)^{-1}\mathbf{Y},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$ and $\mathbf{I}_N$ is a $N \times N$ identity matrix. With a new input $\mathbf{X}_{new}$, KRR predicts the corresponding response by $\hat{Y}_{new} = \hat{f}(\mathbf{X}_{new}) = \sum_{j=1}^{N} \hat{\beta}_j K(\mathbf{X}_{new}, \mathbf{X}_j)$.

When N is huge, it is often reasonable to assume that not all kernel atoms (features) $K_j = K(\cdot, \mathbf{X}_j)$ are relevant for prediction; this amounts to assume that some β_js in (10) are zero. When a K_j is irrelevant, one may also consider $\mathbf{X}_j$ to be less important in learning $\hat{f}$ and thus it can be eliminated from (9). This suggests that one may remove the jth column and the jth row from $\mathbf{K}$ to reduce the computational cost in KRR. Thus, the goal of feature screening here is to screen out most irrelevant kernel atoms K_js before carrying out KRR. In KRR, the number of features equals to the sample size N and $\mathbf{K}$ has a natural dependent structure in both rows and columns. It is thus of interest to see how well ACS could do for feature screening in this challenging setup.

To this end, we generate $\mathcal{D}$ from model (b) in Example 2 with $N = 6000$ and $p = 30$, from which $N = 4800$ entries are randomly selected as a training set and the remaining 1200 ones are treated as a testing set. We then generate the $N \times N$ working kernel matrix $\mathbf{K}$ based on the training set using a Gaussian kernel $K(\mathbf{X}_i, \mathbf{X}_j) = \exp(-\|\mathbf{X}_i - \mathbf{X}_j\|_2^2/100)$. We treat $\mathbf{K}$ as an input design matrix and randomly partition it by rows into $m = 10, 40, 120, 160$ sub-kernel matrices $\mathbf{K}_l$ for $l = 1, \ldots m$, each of which is of dimension $n \times N$ with $n = N/m$. Those sub-kernel matrices $\mathbf{K}_l$s together with the corresponding $n \times 1$ sub-responses $\mathbf{Y}_l$s are then treated as m data segments $\mathcal{D}_l = \{(\mathbf{Y}_l, \mathbf{K}_l)\}_{l=1}^m$, on which we apply ACS with SIRS to screen the irrelevant kernel atoms (i.e. screen irrelevant columns of $\mathbf{K}$). For comparison, we also run SAS and rACS with $R = 3$ for each choice of m. Since rACS shows its superior accuracy in our previous examples, we set the screening threshold γ such that 500 important features will be retained by rACS with $m = 10$; the same γ is then used for all screening methods in all m setups.

In this example, the true model in form of (10) is unknown. We thus evaluate the screening performance in terms of prediction. Specifically, after screening, we obtain a refined kernel matrix by removing the irrelevant columns and the corresponding rows from $\mathbf{K}$. We then carry out KKR to obtain a fitted $\hat{f}$ based on this refined kernel matrix with λ determined by a 10-fold cross validation. We assess the predictive power of $\hat{f}$ based on the testing set in terms of the prediction RMSE

$$\text{RMSE}(\hat{f}) = \left[\frac{1}{n_{test}} \sum_{i \in \mathcal{S}_{test}} (Y_i - \hat{f}(\mathbf{X}_i))^2 \right]^{1/2},$$

where $\mathcal{S}_{test}$ and n_{test} denote the observation index set and the sample size of the testing set respectively.

In Figure 2, we report the mean prediction RMSE as well as the averaged model size (AMS) based on 100 repetitions. We also report the performance of $\hat{f}$ based on a full KRR without screening. It is seen that all screening methods under consideration lead to a similar predictive accuracy comparable to the full KRR. When m is small, SAS and ACS tend to keep the same amount of relevant features. As m increases, SAS becomes more liberal by retaining more features after screening, while ACS remains restrictive. When $m = 160$, SAS suggests 1308 "relevant" features, which is about 2.6 times the number of features

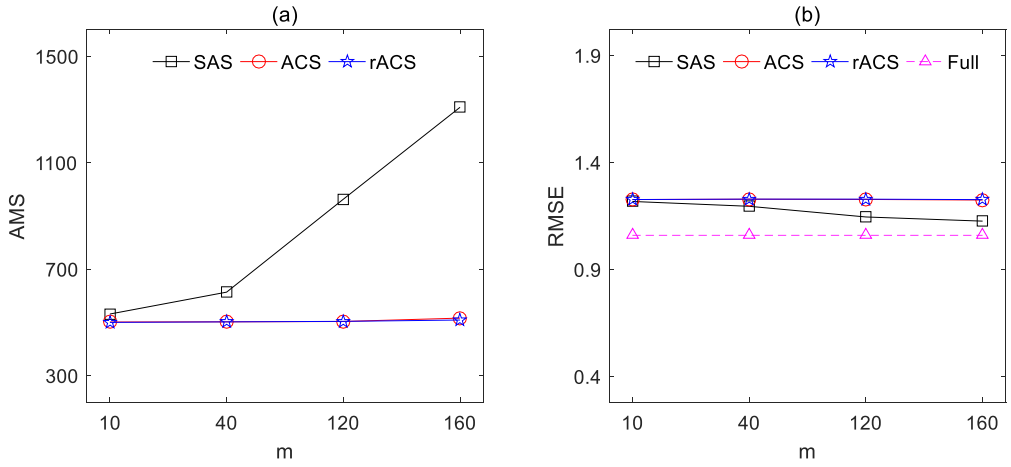

Figure 2: Simulation results for Example 4. (a) The averaged model size (AMS) selected by SAS, ACS, and rACS with different choices of m. (b) The predictive root-mean-squared error of KRR based on the features retained by SAS, ACS, and rACS. The dashed line corresponds to the full KRR without feature screening.

suggested by ACS. Yet, as indicated by their prediction RMSEs, including a large number of features in $\hat{f}$ does not help to significantly improve the predictive power. This implies that a large portion of the SAS-suggested features are actually redundant. In comparison, ACS is stable among all m setups and leads to a more accurate screening in general.

Admittedly, the current theoretical support for ACS does not directly apply to the non-i.i.d KRR learning. Yet, we observe that ACS still performs relatively well in this challenging case. This further indicates that ACS can be a reliable method in practice.

4.2. A real data analysis

We apply the proposed ACS to a real data set[1], which contains 81 covariates extracted from 21,263 superconductors along with the associated critical temperature (response). Readers may refer to Hamidieh (2018) for a detailed description of this data set. It is of interest to predict the unknown response given a set of new values of the covariates. It is likely that the covariates are linked to the response with a non-linear relationship. To avoid potential model mis-specification, we analyze this data set using the gaussian KRR as discussed in Example 4 of the previous subsection.

We remove data entries with missing values in $\mathbf{X}_i$ and get 20,877 available data entries, from which we randomly select $20,000$ entries as a training set and treat the remaining 877 ones as a testing set. This leads to a working kernel matrix $\mathbf{K}$ with $N = 20,000$ observations and $20,000$ kernel atoms $K(\cdot, \mathbf{X}_j)$. Apparently, it is numerically costly to conduct the full KRR on $\mathbf{K}$, which is likely to contain many redundant kernel atoms (features). It is therefore beneficial to conduct distributed feature screening before KRR. To this end, we randomly partition the training set into $m = 10, 100, 200, 500$ segments and run ACS, rACS, and SAS

1. The data set is available at http://archive.ics.uci.edu/ml/datasets/Superconductivty+Data.

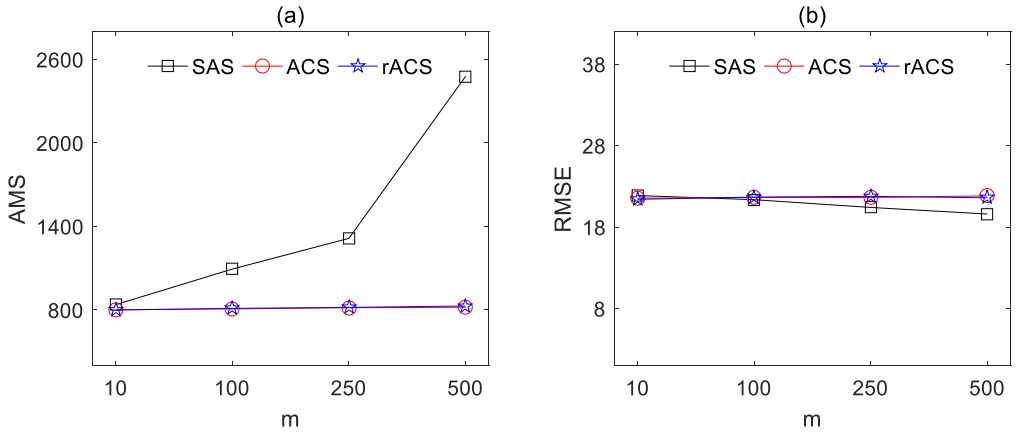

Figure 3: Analysis of superconductor data. (a) The averaged model size (AMS) selected by SAS, ACS, and rACS with different choices of m. (b) The predictive root-mean-squared error of KRR based on the features retained by SAS, ACS, and rACS.

under the same setup as in Example 4, except that the screening threshold γ here is set by the 800th largest $\widetilde{\omega}_j$ in rACS with $m = 10$.

The results are summarized in Figure 3 with 100 repetitions. The pattern in the plots is similar to Figure 2. The screening accuracy and high stability are again observed for the proposed method.

5. Concluding Remarks

Technological innovations have made a profound impact on knowledge discovery. Extracting useful features from massive amount of high dimensional data is essential in many modern scientific areas. In this paper, we proposed a distributed framework (ACS) for feature screening with large-N-large-p data sets. In the spirit of "divide-and-conquer", ACS enables distributed storage and paralleling computing, and thus enjoys a great numerical advantage over the classic screening methods. The key of success for ACS is that we express a correlation measure as a function of several component parameters and conduct distributive unbiased estimation for each of them. With the unbiased component estimates combined together, we then obtained an aggregated correlation estimate $\widetilde{\omega}_j$, which is accurate and insensitive to the number of data segments used in the analysis. This further leads to a computationally efficient and performance reliable screening procedure. Under mild conditions, we showed that $\widetilde{\omega}_j$ is as efficient as the classic centralized estimator, while it drastically reduces the computational cost. The corresponding screening procedure is compatible with a broad range of correlation measures and enjoys the desirable sure screening property.

It should be noted that our current discussion is based on the i.i.d assumption of $(Y_i, \mathbf{X}_i)$, which can be impractical when data segments are naturally stored at different locations. In such a scenario, it is likely that data segments are of different sizes and qualities. To make

the proposed ACS more adaptive, one may replace $\bar{U}_{j,h}$ in (4) by a weighted average, where the weight is proportional to the inverse-variance of the local component estimator $U^l_{j,h}$.

The proposed ACS focuses on model-free feature screening, where most screening measures have a natural expression in (2). It would be promising to extend the distributive idea to other model-based screening methods, where a screening measure may not have a closed-form expression. One possible way is to conduct distributed optimization for estimating a model-based screening measure. We leave this interesting topic for future research.

Acknowledgments

Xu's research was supported by NSERC grant RGPIN-2016-05024 and NSFC grant 116900 14. R. Li's research was supported by NSF grant DMS 1820702 and NIDA, NIH grant P50 DA039838. Xia's research was supported by NSFC grant 11771353. X. Li's research was supported by JSPSP grant 2019SJA2093. The content is solely the responsibility of the authors and does not necessarily represent the official views of the aforementioned funding agencies.

Appendix A. Expressions of $\hat{\omega}_{l,j}$ for the examples in Section 2.3.1

For comparison with the proposed ACS, we list the corresponding expressions of $\hat{\omega}_{l,j}$ used in SAS for the examples in Section 2.3.1.

1. **Pearson correlation**

$$\hat{\omega}_{l,j} = \left| \frac{\frac{1}{n}\sum_{i\in\mathcal{D}_l} X_{ij}Y_i - \frac{1}{n}\sum_{i\in\mathcal{D}_l} X_{ij} \cdot \frac{1}{n}\sum_{i\in\mathcal{D}_l} Y_i}{\sqrt{(\frac{1}{n}\sum_{i\in\mathcal{D}_l} X_{ij}^2 - [\frac{1}{n}\sum_{i\in\mathcal{D}_l} X_{ij}]^2)(\frac{1}{n}\sum_{i\in\mathcal{D}_l} Y_i^2 - [\frac{1}{n}\sum_{i\in\mathcal{D}_l} Y_i]^2)}} \right|.$$

2. **Kendall τ rank correlation**

$$\hat{\omega}_{l,j} = \left| \frac{1}{n(n-1)} \sum_{i_1,i_2\in\mathcal{D}_l} I(X_{i_1 j} < X_{i_2 j}) I(Y_{i_1} < Y_{i_2}) - 1/4 \right|.$$

3. **SIRS correlation**

$$\hat{\omega}_{l,j} = \frac{1}{n(n-1)(n-2)} \sum_{i_1\in\mathcal{D}_l} \left\{ \sum_{i_2\in\mathcal{D}_l} X_{i_2 j} I(Y_{i_2} < Y_{i_1}) \right\}^2,$$

   where $X_{ij} \in \mathcal{D}_l$ are standardized for each feature j.

4. **Distance correlation**

$$\hat{\omega}_{l,j} = \frac{\hat{\theta}_{l,j,1} + \hat{\theta}_{l,j,2} \cdot \hat{\theta}_{l,j,3} - 2\hat{\theta}_{l,j,4}}{\sqrt{(\hat{\theta}_{l,j,5} + \hat{\theta}^2_{l,j,2} - 2\hat{\theta}_{l,j,6})(\hat{\theta}_{l,j,7} + \hat{\theta}^2_{l,j,3} - 2\hat{\theta}_{l,j,8})}},$$

where

$$\hat{\theta}_{l,j,1} = \frac{1}{n^2} \sum_{i_1,i_2 \in \mathcal{D}_l} |Y_{i_1} - Y_{i_2}| \cdot |X_{i_1 j} - X_{i_2 j}|, \tag{11}$$

$$\hat{\theta}_{l,j,4} = \frac{1}{n^3} \sum_{i_1,i_2,i_3 \in \mathcal{D}_l} |Y_{i_1} - Y_{i_3}| \cdot |X_{i_2 j} - X_{i_3 j}|. \tag{12}$$

The expression of $\hat{\theta}_{l,j,h}$ for $h = 2, 3, 5, 7$ is similar to (11); the expression of $\hat{\theta}_{l,j,h}$ for $h = 6, 8$ is similar to (12).

In the above examples, $\hat{\omega}_{l,j}$s are biased with the scale of bias heavily depending on the local sample size $n = N/m$. When m is small, the simple average estimator $\bar{\omega}_j = \frac{1}{m}\sum_{l=1}^{m} \hat{\omega}_{l,j}$ can be severely biased, as the bias can not be eliminated by simple averaging. In comparison, the aggregated correlation estimator $\widetilde{\omega}_j$ conducts distributed unbiased estimation for the components within ω_j and thus leads to a more accurate estimation that is insensitive to m.

Appendix B. Proof of Proposition 1

Let $\Psi_{j,h,q} = E_q[\hat{\theta}_{j,h}(Z_{i_1 j}, \ldots, Z_{i_q j}, Z_{i_{q+1} j}, \ldots, Z_{i_{k_h} j})]$ be the expectation of $\hat{\theta}_{j,h}$ with respect to $(Z_{i_{q+1} j}, \ldots, Z_{i_{k_h} j})$ for $0 \le q \le k_h - 1$ and $\zeta_{j,h,q} = \mathrm{Var}(\Psi_{j,h,q})$.

By equation (5.13) of Hoeffding (1992), we have

$$\mathrm{Var}(U_{j,h}^l) = \binom{n}{k_h}^{-1} \sum_{q=1}^{k_h} \binom{k_h}{q} \binom{n-k_h}{k_h - q} \zeta_{j,h,q}.$$

The variance of $\bar{U}_{j,h}$ is therefore given by

$$\mathrm{Var}(\bar{U}_{j,h}) = \frac{1}{m}\mathrm{Var}(U_{j,h}^l) = \frac{1}{m}\binom{n}{k_h}^{-1} \sum_{q=1}^{k_h} \binom{k_h}{q} \binom{n-k_h}{k_h - q} \zeta_{j,h,q}. \tag{13}$$

By Theorem 5.1 of Hoeffding (1992), we have

$$\frac{\zeta_{j,h,q}}{q} \le \frac{\zeta_{j,h,q+1}}{q+1}$$

for $q = 1, ..., k_h - 1$. This together with Condition C1 implies that $\zeta_{j,h,1} \le ... \le \zeta_{j,h,k_h} = \mathrm{Var}(\hat{\theta}_{j,h}) \le E(\hat{\theta}_{j,h}^2) < 2D_0/\kappa_0^2$ for all j and h. Since k_h is a constant, the qth summand in (13) can be expressed by

$$\frac{1}{m}\binom{n}{k_h}^{-1} \binom{k_h}{q} \binom{n-k_h}{k_h - q} \zeta_{j,h,q} = O(\frac{1}{mn^q}) = O(\frac{m^{q-1}}{N^q}), \text{ for } q = 1, ..., k_h.$$

Expression (7) is hence implied.

To show $\mathrm{Var}(\bar{U}_{j,h})$ is increasing in m, let $\bar{U}_{j,h}[m, n]$ denote $\bar{U}_{j,h}$ based on m data segments, each of which is with size n. By Theorem 5.2 of Hoeffding (1992), we have

$N \cdot \mathrm{Var}(\bar{U}_{j,h}[1,N])$ is a decreasing function of N. Suppose $N = n_1 m_1 = n_2 m_2$ with $n_1 > n_2$. Then, we have

$$\begin{aligned} n_1 \mathrm{Var}(\bar{U}_{j,h}[1,n_1]) &\leq n_2 \mathrm{Var}(\bar{U}_{j,h}[1,n_2]), \\ \frac{n_1}{N}\mathrm{Var}(\bar{U}_{j,h}[1,n_1]) &\leq \frac{n_2}{N}\mathrm{Var}(\bar{U}_{j,h}[1,n_2]), \\ \frac{1}{m_1}\mathrm{Var}(\bar{U}_{j,h}[1,n_1]) &\leq \frac{1}{m_2}\mathrm{Var}(\bar{U}_{j,h}[1,n_2]), \\ \mathrm{Var}(\bar{U}_{j,h}[m_1,n_1]) &\leq \mathrm{Var}(\bar{U}_{j,h}[m_2,n_2]). \end{aligned}$$

The incremental property is therefore proved. ■

Appendix C. Proof of Theorem 2

To prove Theorem 2, we first derive a probability convergence bound of the component estimator $\bar{U}_{j,h}$ in the following Lemma.

Lemma 5 *Suppose Condition C1 is satisfied and $\varepsilon \in (0, \delta_0]$ with an arbitrarily large $\delta_0 > 0$. There exists a sufficiently small $c_0 > 0$ such that*

$$P(|\bar{U}_{j,h} - \theta_{j,h})| \geq \varepsilon) \leq 2(1 - c_0\varepsilon^2/2)^{m\lfloor n/k_h \rfloor},$$

for $j = 1, \ldots, p$ and $h = 1, \ldots, s$, where $\lfloor n/k_h \rfloor$ denotes the largest integer no larger than n/k_h.

Proof of Lemma 5. Let $\hat{\theta}_{j,h}$ be a basis unbiased estimator of $\theta_{j,h}$ with degree k_h. By Markov's inequality, we have

$$\begin{aligned} P(\bar{U}_{j,h} - \theta_{j,h} \geq \varepsilon) &= P(\exp\{\nu(\bar{U}_{j,h} - \theta_{j,h})\} \geq \exp\{\nu\varepsilon\}) \\ &\leq \exp\{-\nu\varepsilon\}\exp\{-\nu\theta_{j,h}\}E[\exp\{\nu\bar{U}_{j,h}\}], \end{aligned} \tag{14}$$

for any $\varepsilon > 0$ and $0 < \nu \leq \kappa_0 m r_h$ with $r_h = \lfloor n/k_h \rfloor$.

Let $\mathcal{S}_l = \{l_1, ..., l_n\}$ denote the index set of $\{Y, \mathbf{X}\}$ copies based on $\mathcal{D}_l$, on which we can construct r_h independent $\hat{\theta}_{j,h}$s. We define an averaged estimator based on those $\hat{\theta}_{j,h}$s by

$$V_{j,h}(Z_{l_1 j}, ..., Z_{l_n j}) = \frac{1}{r_h}\sum_{u=1}^{r_h} \hat{\theta}_{j,h}(Z_{l_{(u-1)k_h+1} j}, ..., Z_{l_{uk_h} j}).$$

Then, the local U-statistic in (3) can be expressed by

$$U^l_{j,h} = \frac{1}{n!}\sum_{\{i_1,...,i_n\}\in\Omega} V_{j,h}(Z_{l_{i_1} j}, ..., Z_{l_{i_n} j}),$$

where $\Omega = \{1, ..., n\}$ and the summation is over all $\{Z_{l_{i_1} j}, ..., Z_{l_{i_n} j}\}$ permutations from $\mathcal{D}_l$. Consequently,

$$\bar{U}_{j,h} = \frac{1}{m}\sum_{l=1}^{m} U^l_{j,h} = \frac{1}{n!}\sum_{\{i_1,...,i_n\}\in\Omega} \frac{1}{m}\sum_{l=1}^{m} V_{j,h}(Z_{l_{i_1} j}, ..., Z_{l_{i_n} j}).$$

Since exponential function is convex, Jensen's inequality implies that

$$\begin{aligned} E[\exp\{\nu \bar{U}_{j,h}\}] &= E\left[\exp\left\{\frac{\nu}{n!}\sum_{\{i_1,...,i_n\}\in\Omega}\left(\frac{1}{m}\sum_{l=1}^{m} V_{j,h}(Z_{l_{i_1}j},...,Z_{l_{i_n}j})\right)\right\}\right] \\ &\leq \frac{1}{n!}\sum_{\{i_1,...,i_n\}\in\Omega} E\left[\exp\left\{\frac{\nu}{m}\sum_{l=1}^{m} V_{j,h}(Z_{l_{i_1}j},...,Z_{l_{i_n}j})\right\}\right] \\ &= \psi_{j,h}^{mr_h}(\kappa), \end{aligned} \tag{15}$$

where $\kappa = \nu/(mr_h)$ and $\psi_{j,h}(\kappa) = E[\exp\{\kappa\hat{\theta}_{j,h}\}]$.

Combining (14) and (15), we have

$$P(\bar{U}_{j,h} - \theta_{j,h} \geq \varepsilon) \leq [\exp\{-\kappa\varepsilon\}\exp\{-\kappa\theta_{j,h}\}\psi_{j,h}(\kappa)]^{mr_h}. \tag{16}$$

Let V be a generic variable. By Taylor expansion, we have $\exp\{\kappa V\} = 1+\kappa V+\kappa^2 V'/2$, where $0 < V' < V^2\exp\{\kappa_1 V\}$ for some $\kappa_1 \in (0,\kappa)$. Thus, factor $\exp\{-\kappa\theta_{j,h}\}\psi_{j,h}(\kappa)$ in (16) can be bounded by

$$\begin{aligned} \exp\{-\kappa\theta_{j,h}\}\psi_{j,h}(\kappa) &= E[\exp\{\kappa(\hat{\theta}_{j,h}-\theta_{j,h})\}] \\ &= E\left[1+\kappa(\hat{\theta}_{j,h}-\theta_{j,h}) + \kappa^2\exp\{\kappa_1(\hat{\theta}_{j,h}-\theta_{j,h})\}(\hat{\theta}_{j,h}-\theta_{j,h})^2/2\right] \\ &= 1+\kappa^2 E\left[(\hat{\theta}_{j,h}-\theta_{j,h})^2\exp\{\kappa_1(\hat{\theta}_{j,h}-\theta_{j,h})\}\right]/2 \\ &\leq 1+\kappa^2[E\hat{\theta}_{j,h}^4 \cdot E\exp\{2\kappa_1(\hat{\theta}_{j,h}-\theta_{j,h})\}]^{1/2}/2, \end{aligned} \tag{17}$$

where (17) is implied by Hölder's inequality.

By Condition C1 and the compactness of Θ, we know (17) can be bounded by $1+D_1\kappa^2$ with some $D_1 > 0$ for all $j = 1,...,p, h = 1,...,s$. Also, when $\kappa\varepsilon < 1$, we have $\exp(-\kappa\varepsilon) \leq 1-\varepsilon\kappa+D_2\varepsilon^2\kappa^2$ with some $D_2 > 0$. Thus, we have the base term in (16) bounded by

$$\begin{aligned} \exp\{-\kappa\varepsilon\}\exp\{-\kappa\theta_{j,h}\}\psi_{j,h}(\kappa) &\leq (1+D_1\kappa^2)(1-\varepsilon\kappa+D_2\varepsilon^2\kappa^2) \\ &= 1-\varepsilon\kappa+D_2\kappa^2\varepsilon^2+D_1\kappa^2-D_1\kappa^3\varepsilon+D_1D_2\kappa^4\varepsilon^2 \\ &\leq 1-\varepsilon\kappa+D_2\kappa^2\varepsilon^2+D_1\kappa^2+D_1D_2\kappa^4\varepsilon^2 \\ &= 1-\varepsilon\kappa+E_1, \end{aligned}$$

where $E_1 = D_2\kappa^2\varepsilon^2+D_1\kappa^2+D_1D_2\kappa^4\varepsilon^2$. By setting $\kappa = c_0\varepsilon$, we have

$$\begin{aligned} \frac{E_1}{\kappa\varepsilon} &= D_2c_0\varepsilon^2+D_1c_0+D_1D_2c_0^3\varepsilon^4 \\ &\leq D_2c_0\delta_0^2+D_1c_0+D_1D_2c_0^3\delta_0^4. \end{aligned} \tag{18}$$

Note that, when $c_0 > 0$ is small enough, we have $\kappa \in (0,\kappa_0)$, $\kappa\varepsilon < 1$, and (18) is bounded by $1/2$. Thus, the base term in (16) is further bounded by

$$\exp\{-\kappa\varepsilon\}\exp\{-\kappa\theta_{j,h}\}\psi_{j,h}(\kappa) \leq 1-\varepsilon\kappa/2. \tag{19}$$

Combining (16) and (19), we have

$$P(\bar{U}_{j,h} - \theta_{j,h} \geq \varepsilon) \leq (1 - c_0\varepsilon^2/2)^{mr_h}.$$

Similarly, we can show that $P(\bar{U}_{j,h} - \theta_{j,h} \leq -\varepsilon) \leq (1 - c_0\varepsilon^2/2)^{mr_h}$. Therefore, we obtain

$$P(|\bar{U}_{j,h} - \theta_{j,h}| \geq \varepsilon) \leq 2(1 - c_0\varepsilon^2/2)^{m\lfloor n/k_h \rfloor}$$

under the conditions specified in the proposition. The proof is complete. ■

With Lemma 5, we prove Theorem 2 based on the following fact.

Lemma 6 *Suppose that $\theta_h, h = 1, ..., s$ are bounded, that is, there exists a positive constant $a > 0$ such that $|\theta_h| < a$. Let $\widetilde{\theta}_h$ be an estimator of θ_h. Suppose for any $\varepsilon \in (0, c]$, there exists a constant $c_1 > 0$ such that, for any $h \in \{1, ..., s\}$,*

$$P(|\widetilde{\theta}_h - \theta_h| \geq \varepsilon) \leq c_1(1 - \varepsilon^2/c_1)^{m\lfloor n/k \rfloor}, \tag{20}$$

where k is a positive integer. Then, there exists a positive constant c' such that

$$\begin{aligned}
P\left(\left||\widetilde{\theta}_h| - |\theta_h|\right| \geq \varepsilon\right) &\leq c'(1 - \varepsilon^2/c')^{m\lfloor n/k \rfloor}, &(21)\\
P(|(\widetilde{\theta}_{h_1} + \widetilde{\theta}_{h_2}) - (\theta_{h_1} + \theta_{h_2})| \geq \varepsilon) &\leq c'(1 - \varepsilon^2/c')^{m\lfloor n/k \rfloor}, &(22)\\
P(|(\widetilde{\theta}_{h_1} - \widetilde{\theta}_{h_2}) - (\theta_{h_1} - \theta_{h_2})| \geq \varepsilon) &\leq c'(1 - \varepsilon^2/c')^{m\lfloor n/k \rfloor}, &(23)\\
P(|\widetilde{\theta}_{h_1}\widetilde{\theta}_{h_2} - \theta_{h_1}\theta_{h_2}| \geq \varepsilon) &\leq c'(1 - \varepsilon^2/c')^{m\lfloor n/k \rfloor}, &(24)\\
P(|\widetilde{\theta}_h^2 - \theta_h^2| \geq \varepsilon) &\leq c'(1 - \varepsilon^2/c')^{m\lfloor n/k \rfloor}. &(25)
\end{aligned}$$

Moreover, suppose there exists a constant $b > 0$ such that $|\theta_{h_2}| > b$. Then, we have

$$P(|\widetilde{\theta}_{h_1}/\widetilde{\theta}_{h_2} - \theta_{h_1}/\theta_{h_2}| \geq \varepsilon) \leq c'(1 - \varepsilon^2/c')^{m\lfloor n/k \rfloor}. \tag{26}$$

If we further assume $\theta_h > 0$, then

$$P\left(\left|\sqrt{\widetilde{\theta}_h} - \sqrt{\theta_h}\right| \geq \varepsilon\right) \leq c'(1 - \varepsilon^2/c')^{m\lfloor n/k \rfloor}. \tag{27}$$

Proof of Lemma 6. We prove the lemma by justifying (21)-(27) sequentially.

The proof of (21) is straightforward. By (20), for any $\varepsilon \in (0, c]$, we have

$$\begin{aligned}
P\left(\left||\widetilde{\theta}_h| - |\theta_h|\right| \geq \varepsilon\right) &\leq P\left(\left|\widetilde{\theta}_h - \theta_h\right| \geq \varepsilon\right)\\
&\leq c_1(1 - \varepsilon^2/c_1)^{m\lfloor n/k \rfloor}.
\end{aligned}$$

We now work on (22). For any $\varepsilon \in (0, c]$, we have

$$\begin{aligned}
& P(|(\widetilde{\theta}_{h_1} + \widetilde{\theta}_{h_2}) - (\theta_{h_1} + \theta_{h_2})| \geq \varepsilon)\\
\leq\ & P(|\widetilde{\theta}_{h_1} - \theta_{h_1}| \geq \varepsilon/2) + P(|\widetilde{\theta}_{h_2} - \theta_{h_2}| \geq \varepsilon/2)\\
\leq\ & 2c_1(1 - \varepsilon^2/(4c_1))^{m\lfloor n/k \rfloor} \leq c_2(1 - \varepsilon^2/c_2)^{m\lfloor n/k \rfloor},
\end{aligned}$$

where $c_2 = 4c_1$. Similarly, we can also show (23).

To show (24), we first prove that $\widetilde{\theta}_h$s are bounded in probability. Specifically, since $|\theta_h| \le a$, we have, for any $\varepsilon \in (0, c]$,

$$\begin{aligned} P\left(|\widetilde{\theta}_h| \ge a + \varepsilon\right) &\le P\left(|\widetilde{\theta}_h - \theta_h| + |\theta_h| \ge a + \varepsilon\right) \\ &\le P\left(|\widetilde{\theta}_h - \theta_h| \ge \varepsilon\right) \\ &\le c_1(1 - \varepsilon^2/c_1)^{m\lfloor n/k \rfloor}. \end{aligned} \tag{28}$$

Therefore,

$$\begin{aligned} & P(|\widetilde{\theta}_{h_1}\widetilde{\theta}_{h_2} - \theta_{h_1}\theta_{h_2}| \ge \varepsilon) \\ \le\ & P(|\widetilde{\theta}_{h_1}\widetilde{\theta}_{h_2} - \widetilde{\theta}_{h_1}\theta_{h_2} + \widetilde{\theta}_{h_1}\theta_{h_2} - \theta_{h_1}\theta_{h_2}| \ge \varepsilon) \\ \le\ & P(|\widetilde{\theta}_{h_1}| \cdot |\widetilde{\theta}_{h_2} - \theta_{h_2}| + |\theta_{h_2}| \cdot |\widetilde{\theta}_{h_1} - \theta_{h_1}| \ge \varepsilon) \\ \le\ & P(|\widetilde{\theta}_{h_1}| \cdot |\widetilde{\theta}_{h_2} - \theta_{h_2}| \ge \varepsilon/2) + P(|\theta_{h_2}| \cdot |\widetilde{\theta}_{h_1} - \theta_{h_1}| \ge \varepsilon/2). \end{aligned} \tag{29}$$

By (20) and (28), the first term of (29) can be bounded by

$$\begin{aligned} & P(|\widetilde{\theta}_{h_1}| \cdot |\widetilde{\theta}_{h_2} - \theta_{h_2}| \ge \varepsilon/2) \\ =\ & P(|\widetilde{\theta}_{h_1}| \cdot |\widetilde{\theta}_{h_2} - \theta_{h_2}| \ge \varepsilon/2, |\widetilde{\theta}_{h_1}| \ge a + \varepsilon) \\ & + P(|\widetilde{\theta}_{h_1}| \cdot |\widetilde{\theta}_{h_2} - \theta_{h_2}| \ge \varepsilon/2, |\widetilde{\theta}_{h_1}| < a + \varepsilon) \\ \le\ & P(|\widetilde{\theta}_{h_1}| \ge a + \varepsilon) + P((a + \varepsilon) \cdot |\widetilde{\theta}_{h_2} - \theta_{h_2}| \ge \varepsilon/2) \\ \le\ & c_1(1 - \varepsilon^2/c_1)^{m\lfloor n/k \rfloor} + c_1(1 - \varepsilon^2/c_3)^{m\lfloor n/k \rfloor}, \end{aligned}$$

where $c_3 = \max\{4(a + c)^2 c_1, c_1\}$. The second term of (29) can be bounded by

$$\begin{aligned} P(|\theta_{h_2}| \cdot |\widetilde{\theta}_{h_1} - \theta_{h_1}| \ge \varepsilon/2) &\le P(|\widetilde{\theta}_{h_1} - \theta_{h_1}| \ge \varepsilon/(2a)) \\ &\le c_1(1 - \varepsilon^2/c_4)^{m\lfloor n/k \rfloor} \end{aligned}$$

with $c_4 = \max\{4a^2 c_1, c_1\}$. Then, by setting $c_5 = \max\{3c_1, c_3\}$, we have

$$P(|\widetilde{\theta}_{h_1}\widetilde{\theta}_{h_2} - \theta_{h_1}\theta_{h_2}| \ge \varepsilon) \le 3c_1(1 - \varepsilon^2/c_3)^{m\lfloor n/k \rfloor} \le c_5(1 - \varepsilon^2/c_5)^{m\lfloor n/k \rfloor},$$

which proves (24). By setting $\widetilde{\theta}_{h_2} = \widetilde{\theta}_{h_1} = \widetilde{\theta}_h$ in (24), we immediately have result (25).

To prove (26), let us first show that $\widetilde{\theta}_{h_2}$ is bounded away from 0 in probability. Since $|\theta_{h_2}| > b > 0$, there exists a constant $\delta_1 \in (0, c)$ such that for some $b' = b - \delta_1 > 0$,

$$\begin{aligned} P(|\widetilde{\theta}_{h_2}| \le b') &\le P(|\theta_{h_2}| - |\widetilde{\theta}_{h_2} - \theta_{h_2}| \le b - \delta_1) \\ &\le P(|\widetilde{\theta}_{h_2} - \theta_{h_2}| \ge \delta_1) \\ &\le c_1(1 - \delta_1^2/c_1)^{m\lfloor n/k \rfloor}. \end{aligned}$$

Let $c_6 = c_1 c^2/\delta_1^2$. Then, for $\varepsilon \in (0, c)$, we have

$$P(|\widetilde{\theta}_{h_2}| \le b') \le c_1(1 - \varepsilon^2/c_6)^{m\lfloor n/k \rfloor}. \tag{30}$$

Based on (30), we have

$$\begin{aligned} & P(|\widetilde{\theta}_{h_1}/\widetilde{\theta}_{h_2} - \theta_{h_1}/\theta_{h_2}| \geq \varepsilon) \\ = \; & P(|\widetilde{\theta}_{h_1}/\widetilde{\theta}_{h_2} - \theta_{h_1}/\theta_{h_2}| \geq \varepsilon, |\widetilde{\theta}_{h_2}| \leq b') + P(|\widetilde{\theta}_{h_1}/\widetilde{\theta}_{h_2} - \theta_{h_1}/\theta_{h_2}| \geq \varepsilon, |\widetilde{\theta}_{h_2}| > b') \\ \leq \; & P(|\widetilde{\theta}_{h_2}| \leq b') + P(|\widetilde{\theta}_{h_1}/\widetilde{\theta}_{h_2} - \theta_{h_1}/\theta_{h_2}| \geq \varepsilon, |\widetilde{\theta}_{h_2}| > b') \\ \leq \; & c_1(1-\varepsilon^2/c_6)^{m\lfloor n/k \rfloor} + P(|\widetilde{\theta}_{h_1}/\widetilde{\theta}_{h_2} - \theta_{h_1}/\theta_{h_2}| \geq \varepsilon, |\widetilde{\theta}_{h_2}| > b'). \end{aligned} \tag{31}$$

In (31), the second term can be bounded by

$$\begin{aligned} & P(|\widetilde{\theta}_{h_1}/\widetilde{\theta}_{h_2} - \theta_{h_1}/\theta_{h_2}| \geq \varepsilon, |\widetilde{\theta}_{h_2}| > b') \\ \leq \; & P(|\widetilde{\theta}_{h_1}/\widetilde{\theta}_{h_2} - \theta_{h_1}/\widetilde{\theta}_{h_2}| + |\theta_{h_1}/\widetilde{\theta}_{h_2} - \theta_{h_1}/\theta_{h_2}| \geq \varepsilon, |\widetilde{\theta}_{h_2}| > b') \\ \leq \; & P\left(\frac{1}{b'}|\widetilde{\theta}_{h_1} - \theta_{h_1}| \geq \varepsilon/2\right) + P\left(\frac{|\theta_{h_1}|}{|\widetilde{\theta}_{h_2}| \cdot |\theta_{h_2}|}|\widetilde{\theta}_{h_2} - \theta_{h_2}| \geq \varepsilon/2\right) \\ \leq \; & c_1(1-\varepsilon^2/c_7)^{m\lfloor n/k \rfloor} + c_1(1-\varepsilon^2/c_8)^{m\lfloor n/k \rfloor}, \end{aligned} \tag{32}$$

where $c_7 = \max\{4c_1/(b')^2, c_1\}$ and $c_8 = \max\{4a^2c_1/(b'b)^2, c_1\}$. Let $c_9 = \max\{3c_1, c_6, c_7, c_8\}$, then we have

$$P(|\widetilde{\theta}_{h_1}/\widetilde{\theta}_{h_2} - \theta_{h_1}/\theta_{h_2}| \geq \varepsilon) \leq c_9(1-\varepsilon^2/c_9)^{m\lfloor n/k \rfloor}.$$

Let us work on (27). Since $\theta_h > 0$, there exists a $\tilde{b} > 0$ such that $\theta_h > \tilde{b}$. Similar to (30)-(32), there exist two positive constants $\tilde{b}'$ and c_{10} such that

$$\begin{aligned} & P\left(\left|\sqrt{\widetilde{\theta}_h} - \sqrt{\theta_h}\right| \geq \varepsilon\right) \\ \leq \; & P(|\widetilde{\theta}_h| \leq \tilde{b}') + P\left(\frac{\widetilde{\theta}_h - \theta_h}{\sqrt{\widetilde{\theta}_h} + \sqrt{\theta_h}} \geq \varepsilon, |\widetilde{\theta}_h| > \tilde{b}'\right) \\ \leq \; & c_1(1-\varepsilon^2/c_{10})^{m\lfloor n/k \rfloor} + P\left(|\widetilde{\theta}_h - \theta_h| \geq (\sqrt{\tilde{b}'} + \sqrt{\tilde{b}})\varepsilon\right) \\ \leq \; & c_1(1-\varepsilon^2/c_{10})^{m\lfloor n/k \rfloor} + c_1(1-\varepsilon^2/c_{11})^{m\lfloor n/k \rfloor}, \end{aligned}$$

where $c_{11} = \max\{c_1/(\sqrt{\tilde{b}'} + \sqrt{\tilde{b}})^2, c_1\}$. By setting $c_{12} = \max\{2c_1, c_{10}, c_{11}\}$, we obtain that

$$P\left(\left|\sqrt{\widetilde{\theta}_h} - \sqrt{\theta_h}\right| \geq \varepsilon\right) \leq c_{12}(1-\varepsilon^2/c_{12})^{m\lfloor n/k \rfloor}.$$

Result (27) is therefore proved.

Combining the results in (21)-(27), we prove Lemma 5 by setting $c' = \max\{c_1, c_2, c_5, c_9, c_{12}\}$. ■

Proof of Theorem 2. By Lemma 5, for any $\varepsilon \in (0, \delta_0]$, there exists a $c_0 > 0$ such that

$$P(|\bar{U}_{j,h} - \theta_{j,h}| \geq \varepsilon) \leq 2(1-c_0\varepsilon^2)^{m\lfloor n/k_h \rfloor} \leq 2(1-c_0\varepsilon^2)^{m\lfloor n/k \rfloor},$$

where $k = \max\{k_h, h = 1, \ldots, s\} \leq n$. Since δ_0 can be arbitrarily large, the inequality holds with $\varepsilon = cN^{-\tau} \in (0, c]$ for some $0 < \tau < 1/2$. Thus, we have

$$\begin{aligned} P(|\bar{U}_{j,h} - \theta_{j,h}| \geq cN^{-\tau}) &\leq 2(1 - c_{13}N^{-2\tau}/2)^{m\lfloor n/k \rfloor} \\ &\leq c_{14}(1 - N^{-2\tau}/c_{14})^{m\lfloor n/k \rfloor}, \quad h = 1, ..., s, \end{aligned} \tag{33}$$

where $c_{14} = \max\{2, 2/c_{13}\}$. This implies that the results of Lemma 5 are applicable by setting $\widetilde{\theta}_h = \bar{U}_{j,h}$.

By Condition C2, we require that $\widetilde{\omega}_j = g(\bar{U}_{j,1}, \ldots, \bar{U}_{j,s})$ is constructed by a finite number of simple numerical operations, which serve as basic building blocks of $g(\cdot)$. For each building block, Lemma 6 can be used immediately to establish the convergence bound for the corresponding $\widetilde{\omega}_j$. With finite combination of those building blocks, (33) further implies that

$$P(|\widetilde{\omega}_j - \omega_j| \geq cN^{-\tau}) \leq \eta(1 - N^{-2\tau}/\eta)^{m\lfloor n/k \rfloor} \tag{34}$$

for some generic positive constant η.

Consequently, we have

$$\begin{aligned} P\left(\max_{1 \leq j \leq p} |\widetilde{\omega}_j - \omega_j| \geq cN^{-\tau}\right) &\leq \sum_{j=1}^{p} P\left(|\widetilde{\omega}_j - \omega_j| \geq cN^{-\tau}\right) \\ &\leq \eta p(1 - N^{-2\tau}/\eta)^{m\lfloor n/k \rfloor}. \end{aligned}$$

The theorem is proved. ■

Appendix D. Proof of Theorem 3

When $\omega_j = g(\boldsymbol{\theta}_j)$ is Lipschitz continuous in $\boldsymbol{\theta}_j$, we have

$$\begin{aligned} \mathrm{MSE}(\widetilde{\omega}_j) &= \mathrm{MSE}(g(\bar{U}_{j,1}, ..., \bar{U}_{j,s})) \\ &= E(g(\bar{U}_{j,1}, ..., \bar{U}_{j,s}) - g(\boldsymbol{\theta}_j))^2 \\ &\leq E\left(L^2 \sum_{h=1}^{s} (\bar{U}_{j,h} - \theta_{j,h})^2\right) \\ &= L^2 \sum_{h=1}^{s} E(\bar{U}_{j,h} - \theta_{j,h})^2 \\ &\leq sL^2 \max_{j,h}\{\mathrm{Var}(\bar{U}_{j,h})\}, \end{aligned}$$

where L is defined in C2′. Thus, the theorem is implied directly by Proposition 1. ■

Appendix E. Proof of Theorem 4

Note that $\gamma = cN^{-\tau}$. If $\mathcal{M} \nsubseteq \widetilde{\mathcal{M}}$, there must exist some $j \in \mathcal{M}$ such that $\widetilde{\omega}_j < cN^{-\tau}$. Also, by Condition C3, we assume $\min_{j \in \mathcal{M}} \omega_j \geq 2cN^{-\tau}$. Thus, $\mathcal{M} \nsubseteq \widetilde{\mathcal{M}}$ implies $|\widetilde{\omega}_j - \omega_j| > cN^{-\tau}$

for some $j \in \mathcal{M}$. Therefore, by (34), we have

$$\begin{aligned} P\{\mathcal{M} \subseteq \widetilde{\mathcal{M}}\} &\geq P(\max_{j\in\mathcal{M}} |\widetilde{\omega}_j - \omega_j| \leq cN^{-\tau}) \\ &\geq 1 - P(\max_{j\in\mathcal{M}} |\widetilde{\omega}_j - \omega_j| > cN^{-\tau}) \\ &\geq 1 - d \cdot P(|\widetilde{\omega}_j - \omega_j| > cN^{-\tau}) \\ &\geq 1 - d\eta(1 - N^{-2\tau}/\eta)^{m\lfloor n/k \rfloor}, \end{aligned}$$

where d is the cardinality of $\mathcal{M}$. The theorem is proved.

To see that the above probability bound goes to 1 as $N \to \infty$, it suffice to show that

$$d(1 - N^{-2\tau}/\eta)^{m\lfloor n/k \rfloor} = o(1).$$

Note that

$$(1 - N^{-2\tau}/\eta)^{m\lfloor n/k \rfloor} = \left\{(1 - 1/(\eta N^{2\tau}))^{\eta N^{2\tau}}\right\}^{m\lfloor n/k \rfloor/(\eta N^{2\tau})}.$$

Since

$$\lim_{N\to\infty} (1 - 1/(\eta N^{2\tau}))^{\eta N^{2\tau}} = 1/e,$$

there exists a positive integer N_1 such that whenever $N > N_1$, we have $(1-1/(\eta N^{2\tau}))^{\eta N^{2\tau}} < 2/e < 1$.

Also note that $0 < \tau < 1/2$ and $m\lfloor n/k \rfloor$ is in the same order of N; there exists a positive integer N_2 and a constant $\upsilon \in (0, 1 - 2\tau)$ such that whenever $N > N_2$, we have $m\lfloor n/k \rfloor/(\eta N^{2\tau}) > N^{\upsilon}$.

Therefore, when $N > \max\{N_1, N_2\}$, we have

$$\left\{(1 - 1/(\eta N^{2\tau}))^{\eta N^{2\tau}}\right\}^{m\lfloor n/k \rfloor/\eta N^{2\tau}} < (2/e)^{N^{\upsilon}}.$$

It follows that, when $d = O(N)$ and N is large enough,

$$d(1 - N^{-2\tau}/\eta)^{m\lfloor n/k \rfloor} \leq d(2/e)^{N^{\upsilon}} = o(1).$$

The probability bound in Theorem 4 thus goes to 1 as $N \to \infty$. ■

References

Moulinath Banerjee, Cecile Durot, Bodhisattva Sen, et al. Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *The Annals of Statistics*, 47(2): 720–757, 2019.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3): 1352–1382, 2018.

Xueying Chen and Min Ge Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24(4):1655–1684, 2014.

Yixin Chen, Guozhu Dong, Jiawei Han, Jian Pei, Benjamin W Wah, and Jianyong Wang. Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering*, 18(12):1585–1599, 2006.

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

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

Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of machine learning research*, 10(Sep):2013–2038, 2009.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in Statistics*, pages 308–334. Springer, 1992.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Gaorong Li, Heng Peng, Jun Zhang, Lixing Zhu, et al. Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877, 2012a.

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

N Lin and R Xi. Fast surrogates of u-statistics. *Computational Statistics & Data Analysis*, 54(1):16–24, 2010.

Nan Lin and Ruibin Xi. Aggregated estimating equation estimation. *Statistics and Its Interface*, 4(1):73–83, 2011.

Chengchun Shi, Wenbin Lu, and Rui Song. A massive data framework for m-estimators with cubic-rate. *Journal of the American Statistical Association*, 113(524):1698–1709, 2018.

Yuanshan Wu and Guosheng Yin. Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika*, 102(1):65–76, 2015.

Chen Xu, Yongquan Zhang, Runze Li, and Xindong Wu. On the feasibility of distributed kernel regression for big data. *IEEE Transactions on Knowledge and Data Engineering*, 28(11):3041–3052, 2016.

Yuchen Zhang, John C Duchi, and Martin Wainwright. Comunication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(1):3321–3363, 2012.

Tingyou Zhou, Liping Zhu, Runze Li, and Chen Xu. Model-free forward regression via cumulative divergence. *Journal of the American Statistical Association*, page in press, 2019.

Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496): 1464–1475, 2011.