Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Feature Screening for Massive Data Analysis by Subsampling

Xuening Zhu, Rui Pan, Shuyuan Wu & Hansheng Wang

To cite this article: Xuening Zhu, Rui Pan, Shuyuan Wu & Hansheng Wang (2022) Feature Screening for Massive Data Analysis by Subsampling, Journal of Business & Economic Statistics, 40:4, 1892-1903, DOI: 10.1080/07350015.2021.1990771

To link to this article: https://doi.org/10.1080/07350015.2021.1990771

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 40, NO. 4, 1892–1903
https://doi.org/10.1080/07350015.2021.1990771

Taylor & Francis
Taylor & Francis Group

Feature Screening for Massive Data Analysis by Subsampling

Xuening Zhu[a], Rui Pan[b], Shuyuan Wu[c], and Hansheng Wang[c]

[a]School of Data Science, Fudan University, Shanghai, China; [b]School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China; [c]Guanghua School of Management, Peking University, Beijing, China

ABSTRACT
Modern statistical analysis often encounters massive datasets with ultrahigh-dimensional features. In this work, we develop a subsampling approach for feature screening with massive datasets. The approach is implemented by repeated subsampling of massive data and can be used for analyzing tasks with memory constraints. To conduct the procedure, we first calculate an *R*-squared screening measure (and related sample moments) based on subsamples. Second, we consider three methods to combine the local statistics. In addition to the simple average method, we design a jackknife debiased screening measure and an aggregated moment screening measure. Both approaches reduce the bias of the subsampling screening measure and therefore increase the accuracy of the feature screening. Last, we consider a novel sequential sampling method, that is more computationally efficient than the traditional random sampling method. The theoretical properties of the three screening measures under both sampling schemes are rigorously discussed. Finally, we illustrate the usefulness of the proposed method with an airline dataset containing 32.7 million records.

ARTICLE HISTORY
Received February 2021
Accepted September 2021

KEYWORDS
Feature screening; Jackknife debiased estimator; Massive dataset; Subsampling

1. Introduction

Modern statistical analysis often faces massive datasets with a huge number of features. Massive data size and the high-dimensional features impose challenges with respect to data storage, communication and statistical analysis (Fan et al. 2020). On the one hand, the large data size makes data storage and computing difficult. As a result, subsampling methods (Ma, Mahoney, and Yu 2015; Wang, Zhu, and Ma 2018; Wang, Yang, and Stufken 2019; Yu et al. 2020) and distributed statistical tools (Shamir, Srebro, and Zhang 2014; Chang, Lin, and Wang 2017; Jordan, Lee, and Yang 2019; Fan et al. 2019) are proposed, and the corresponding theoretical properties are extensively studied. On the other hand, thousands of features may be obtained from data in various disciplines (Fan and Lv 2008), which makes the problem of high-dimensional analysis important. To make statistical analysis (e.g., regression analysis) feasible, feature screening methods are proposed and widely used (Wang 2009; Li, Zhong, and Zhu 2012b; Zhou, Zhu, and Li 2020; Li et al. 2020).

In the past decade, various subsampling methods have been proposed due to the emergence of massive data. On the one hand, subsampling makes the computation feasible for massive datasets. On the other hand, it solves the problem of automatic statistical inference, for example, estimating the standard error of sample correlation. Kleiner et al. (2014) designed a bag of little bootstraps (BLB) method to calculate precision measures of interested estimations, which is more computationally efficient than traditional bootstrap method on massive datasets. Sengupta, Volgushev, and Shao (2016) revised the BLB method by reducing resampling times, which further reduces the computational cost. Recently, Pan et al. (2020) proposed to conduct subsampling on the hard drive and developed an algorithm called sequential addressing subsampling (SAS). In addition to bootstrap-based subsampling methods, research has focused on deriving optimal subsampling schemes for complete dataset. A two-step procedure was developed by Wang, Zhu, and Ma (2018) to minimize the asymptotic mean squared error. Furthermore, an information-based optimal subdata selection method was designed by Wang (2019) to preserve most of the information contained in the full data, and Yu et al. (2020) developed an optimal distributed subsampling algorithm, that is able to obtain quasi-likelihood estimators for massive datasets.

Although the existing subsampling methods can realize automatic statistical inference, they are typically designed for data with fixed dimensions. When the features are of high dimension, a natural strategy is to screen out irrelevant features to facilitate further analysis. Specifically, a correlation measure is calculated between the response and each feature to implement feature screening: features with weak correlation are treated as irrelevant and removed. The literature on feature screening is rich. Fan and Lv (2008) proposed a sure independence screening (SIS) procedure using Pearson correlation, that, with probability tending to 1, could ensure that the selected feature set covered the important features. Wang (2009) revised the SIS procedure using a forward regression approach. Li, Zhong, and Zhu (2012b) developed a model-free sure independence screening procedure with the distance correlation (DC-SIS). See Li et al. (2012a), Wang (2012), Wu and Yin (2015), Barut, Fan, and Verhasselt (2016), Pan, Wang, and Li (2016), and Zhou, Zhu, and Li (2020) for recent developments in this regard. Moreover, in a

CONTACT Shuyuan Wu ✉ *shuyuan.w@pku.edu.cn* Guanghua School of Management, Peking University, Beijing, China.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/UBES*.

recent work of Li et al. (2020), the authors studied a distributed feature screening approach via componentwise debiasing. The method is applicable to massive datasets and relies on a distributed architecture.

In this work, we focus on subsampling-based feature screening for massive datasets with ultrahigh-dimensional features. We first obtain a measure based on the R-squared for each subsample as a subsample measure; then, we develop two kinds of screening procedures based on the subsample measures.

Debiased average screening. The first procedure is called jackknife debiased simple average screening, that is, debiased average screening (DAS) for short. Specifically, for each feature, we average the debiased R-squared measure obtained for each subsample associated with the feature. The debiased procedure reduces the bias of the subsample measure, and allows for smaller subsample sizes while maintaining the same screening efficiency. The procedure is easy to implement based on subsamples and is sufficiently flexible to extend to various screening measures.

Aggregated moment screening. The second procedure is called the aggregated moment screening (AMS). In contrast to simple average screening, we first take the average of several moment estimators over all the subsamples respectively, and then combine the averaged moment estimators to form an aggregated moment measure. The idea is similar to the componentwise debiasing method in distributed feature screening (Li et al. 2020). The AMS procedure requires the screening measure to take the specific forms of several simple moments. As a result, it lacks the flexibility compared to the DAS procedure.

In this work, we develop two screening procedures based on subsamples for massive datasets with ultrahigh dimensions. The main contributions of this work are as follows. First, the sampling is conducted directly on the hard drive, which greatly reduces the time cost and enables subsampling under a memory constraint. This is particularly useful for massive datasets. Second, the proposed screening measure can handle qualitative features with a diverging number of levels, which have not been sufficiently investigated in literature but are frequently encountered in practice. Third, we derive the theoretical properties of uniform convergence and screening consistency under various subsampling schemes.

The rest of this article is organized as follows. Section 2 develops the model setting and the subsampling methods. Section 3 presents the theoretical properties of the proposed approaches. Numerical studies are presented in Section 4, including the simulation experiments and real dataset analysis. Finally, the article concludes with a brief discussion in Section 5.

2. Feature Screening for Massive Data

2.1. Model and Notations

Suppose there are N observations, which are indexed as $i = 1, \ldots, N$. For the ith observation, we record $Y_i \in \mathbb{R}^1$ as a continuous response and $\mathcal{X}_i \in \mathbb{R}^p$ as the associated covariate vector. We consider the case that p is of ultrahigh dimension. Specifically, we divide $\mathcal{X}_i$ by variable types, that is, $\mathcal{X}_i = (X_i^\top, Z_i^\top)^\top$, where $X_i = (X_{i1}, \ldots, X_{ip_1})^\top \in \mathbb{R}^{p_1}$ represents quantitative variables and $Z_i = (Z_{i1}, \ldots, Z_{ip_2})^\top \in \mathbb{R}^{p_2}$ represents qualitative variables. Immediately, we have $p = p_1 + p_2$. Suppose the jth qualitative variable Z_{ij} takes l_j levels. Then, we transform it to $l_j - 1$ dummy variables as $\mathcal{Z}_{ij} = (\mathcal{Z}_{ij1}, \ldots, \mathcal{Z}_{ij(l_j-1)})^\top \in \mathbb{R}^{l_j-1}$ for $1 \le j \le p_2$, where $\mathcal{Z}_{ijl_j}$ is taken as the baseline without loss of generality. To model the continuous type response Y_i, we consider a linear regression model,

$$Y_i = X_i^\top \beta + \sum_{j=1}^{p_2} \mathcal{Z}_{ij}^\top \gamma_j + \varepsilon_i, \tag{2.1}$$

where $\beta \in \mathbb{R}^{p_1}$ and $\gamma_j \in \mathbb{R}^{l_j-1}$ are the associated regression parameters. Here, ε_i is the independent noise term with $\mathrm{var}(\varepsilon_i) = \sigma^2$. In addition, assume ε_i and $\mathcal{X}_i$ are independent of each other.

Let $\mathbb{X} = (X_1, \ldots, X_N)^\top \in \mathbb{R}^{N\times p_1}$, $\mathbb{Z}_j = (\mathcal{Z}_{1j}, \ldots, \mathcal{Z}_{Nj})^\top \in \mathbb{R}^{N\times(l_j-1)}$, and $\mathbb{Y} = (Y_1, \ldots, Y_N)^\top \in \mathbb{R}^N$. Denote the jth column vector of $\mathbb{X}$ as $\mathbb{X}_j \in \mathbb{R}^N$. Under the high-dimensional setting, sparsity is typically assumed; that is, only a set of important features have significant effects on the response. Define $\mathcal{M}_T^\beta = \{1 \le j \le p_1 : \beta_j \ne 0\}$ as the true model for the quantitative variables. For the qualitative variables, we announce $\mathbb{Z}_j$ to be important if there exists at least one level $1 \le k \le l_j - 1$ such that $\gamma_{jk} \ne 0$. Equivalently, the true model for qualitative variables is defined as $\mathcal{M}_T^\gamma = \{1 \le j \le p_2 :$ there exists $1 \le k \le l_j - 1$ such that $\gamma_{jk} \ne 0\}$.

2.2. Sequential Addressing Subsampling

Under the setting of massive data, the sample size N is extremely large. Hence, performing statistical analysis using the whole data is time consuming and even infeasible under memory constraints. As an alternative, one could conduct repeated subsampling and perform statistical analysis based on the subsamples. A straightforward procedure is to repeatedly perform subsampling with replacement from the original dataset. This scheme is then referred to as random addressing sampling (RAS), which requires to sampling of n observations by addressing the pointers n times on the hard drive. Sampling directly from the hard drive is particularly useful when the whole data cannot be read into memory at once. However, the RAS procedure is time consuming for large-scale datasets.

To address this issue, Pan et al. (2020) proposed a sequential addressing subsampling (SAS) approach, that first shuffles the dataset by randomly permuting all the data points. The shuffling procedure is conducted only once. Then, a random data point is addressed on the hard drive, and the subsequent $n - 1$ data points are read into memory at once. This process yields one sequential subsample. Note that the SAS method requires only a single addressing on the hard drive to obtain a sequential subsample, and this procedure greatly reduces the sampling cost.

Remark 1. Compared to that of the conventional sampling approach, the sampling space of the SAS method is smaller. Specifically, for the SAS method, only $N - n + 1$ subsamples can be generated given the shuffling data. In contrast, for the conventional sampling method, the subsample has N^n possible combinations, which is much larger. As a result, we should carefully select the subsample size n and number of repeated subsamples to make the procedure work.

Remark 2. Thanks to an anonymous referee and we find that the idea of the sequential addressing method can be also used in the divide-and-conquer type method. Specifically, after data shuffling, we split the whole dataset into M nonoverlapped manageable segments with segment size $n = N/M$. As a result, relevant statistics can be computed based on each segment. We refer to this method as divide-and-conquer (DC) method. The method shares the same spirit with the SAS method but without subsampling procedure. Compared to the RAS- and SAS-type sampling methods, the DC method has much smaller number of segments M especially when n is larger. As we will discuss in the theoretical analysis, this can lead to inferior statistical inference performance.

2.3. Feature Screening Measures Via Subsampling

A straightforward way to conduct feature screening via subsampling is to calculate the screening measures using each subsampled data and then combine them by simple average. We refer to the screening measure derived by this procedure as the simple average screening measure. Although the simple average screening measure is easy to obtain, it produces a nonnegligible amount of bias. To rectify the bias issue, we propose two screening measures, both measures designed based on the R-squared value.

The first screening measure, called the debiased average screening (DAS) measure, is simply a simple average of the subsample measure with an additional jackknife bias reduction step. The second procedure is called the aggregated moment screening (AMS) measure, which first calculates sample moments based on subsamples and then reassembles them together to form the final result. The basic ideas of the above two screening measures are illustrated in Figure 1.

To illustrate the details of the screening measures, we first describe the R-squared screening measure based on one subsample. Suppose that we conduct the sampling procedure B times. Furthermore, denote the predictors and response associated with the kth subsample as

$$\begin{aligned}
\mathbb{X}_{(k)} &= (\widetilde{X}_{(k)1}, \widetilde{X}_{(k)2}, \cdots, \widetilde{X}_{(k)n})^\top \in \mathbb{R}^{n\times p_1},\\
\mathbb{Z}_{(k)j} &= (\widetilde{\mathcal{Z}}_{(k)1j}, \widetilde{\mathcal{Z}}_{(k)2j}, \cdots, \widetilde{\mathcal{Z}}_{(k)nj})^\top \in \mathbb{R}^{n\times(l_j-1)},\\
\mathbb{Y}_{(k)} &= (\widetilde{Y}_{(k)1}, \widetilde{Y}_{(k)1}, \ldots, \widetilde{Y}_{(k)n})^\top \in \mathbb{R}^{n}.
\end{aligned}$$

For convenience we can scale $\mathbb{X}_{(k)}$, $\mathbb{Z}_{(k)j}$ and $\mathbb{Y}_{(k)}$ to satisfy $\sum_i \widetilde{X}_{(k)i} = 0$, $\sum_i Z_{(k)ij} = 0$ and $\sum_i \widetilde{Y}_{(k)i} = 0$. Let the set of quantitative variables of interest be collected by $\mathcal{M} = \{j_1, \ldots, j_m\}$. For the kth subsample, let $\mathbb{X}^{\mathcal{M}}_{(k)} = (\mathbb{X}_{(k)j} : j \in \mathcal{M}) \in \mathbb{R}^{n\times|\mathcal{M}|}$. Correspondingly, define $H^{\mathcal{M}}_{(k)} = \mathbb{X}^{\mathcal{M}}_{(k)}\{(\mathbb{X}^{\mathcal{M}}_{(k)})^\top \mathbb{X}^{\mathcal{M}}_{(k)}\}^{-1}(\mathbb{X}^{\mathcal{M}}_{(k)})^\top$. Using $\mathbb{X}^{\mathcal{M}}_{(k)}$ as the covariates and $\mathbb{Y}_{(k)}$ as the response, the R-squared for the kth subsample can be expressed as

$$\begin{aligned}
R^2(\mathbb{X}^{\mathcal{M}}_{(k)}) &= \frac{\left(\mathbb{X}^{\mathcal{M}\top}_{(k)}\mathbb{Y}_{(k)}\right)^\top \left(\mathbb{X}^{\mathcal{M}\top}_{(k)}\mathbb{X}^{\mathcal{M}}_{(k)}\right)^{-1} \left(\mathbb{X}^{\mathcal{M}\top}_{(k)}\mathbb{Y}_{(k)}\right)}{||\mathbb{Y}_{(k)} - \overline{\mathbb{Y}}_{(k)}||^2}\\
&= \frac{\mathbb{Y}^\top_{(k)} H^{\mathcal{M}}_{(k)} \mathbb{Y}_{(k)}}{||\mathbb{Y}_{(k)} - \overline{\mathbb{Y}}_{(k)}||^2}, \qquad (2.2)
\end{aligned}$$

where $\overline{\mathbb{Y}}_{(k)} = n^{-1}\sum_{i=1}^{n} \widetilde{Y}_{(k)i}$. The R-squared screening measure is suitable for both quantitative variables and qualitative variables with multiple levels. Since the denominator (i.e., $||\mathbb{Y}_{(k)} - \overline{\mathbb{Y}}_{(k)}||^2$) in (2.2) does not vary with different covariates $\mathbb{X}_{(k)j}$, hence it does not change the rank of the jth covariate in the final screening result (Fan, Feng, and Song 2011). To make it more computationally efficient, we could omit $||\mathbb{Y}_{(k)} - \overline{\mathbb{Y}}_{(k)}||^2$ in (2.2) in the practical implementation.

Remark 3. The screening measure can be seen as an extension of single-variable case. Typically, for a quantitative variable $\mathbb{X}_j$,

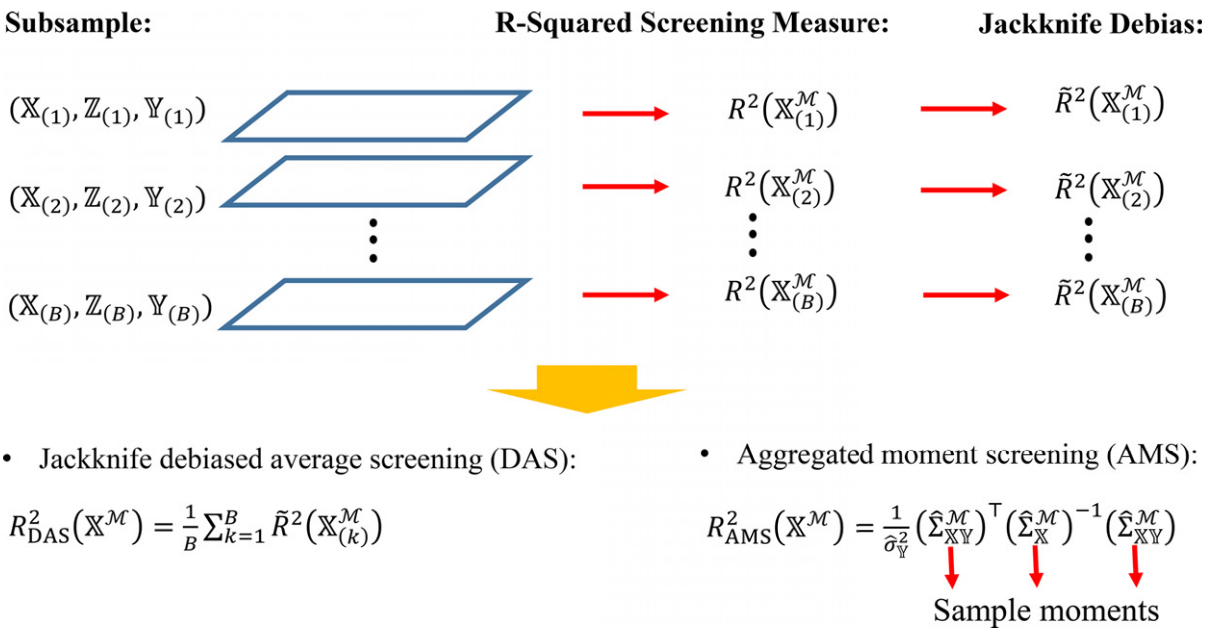
Subsample:
R-Squared Screening Measure:
Jackknife Debias:
$(\mathbb{X}_{(1)}, \mathbb{Z}_{(1)}, \mathbb{Y}_{(1)})$
$(\mathbb{X}_{(2)}, \mathbb{Z}_{(2)}, \mathbb{Y}_{(2)})$
$(\mathbb{X}_{(B)}, \mathbb{Z}_{(B)}, \mathbb{Y}_{(B)})$
$R^2(\mathbb{X}^{\mathcal{M}}_{(1)})$
$R^2(\mathbb{X}^{\mathcal{M}}_{(2)})$
$R^2(\mathbb{X}^{\mathcal{M}}_{(B)})$
$\tilde{R}^2(\mathbb{X}^{\mathcal{M}}_{(1)})$
$\tilde{R}^2(\mathbb{X}^{\mathcal{M}}_{(2)})$
$\tilde{R}^2(\mathbb{X}^{\mathcal{M}}_{(B)})$
Jackknife debiased average screening (DAS):
$R^2_{\mathrm{DAS}}(\mathbb{X}^{\mathcal{M}}) = \frac{1}{B}\sum_{k=1}^{B}\tilde{R}^2(\mathbb{X}^{\mathcal{M}}_{(k)})$
Aggregated moment screening (AMS):
$R^2_{\mathrm{AMS}}(\mathbb{X}^{\mathcal{M}}) = \frac{1}{\hat{\sigma}^2_{\mathbb{Y}}}(\hat{\Sigma}^{\mathcal{M}}_{\mathbb{XY}})^\top(\hat{\Sigma}^{\mathcal{M}}_{\mathbb{X}})^{-1}(\hat{\Sigma}^{\mathcal{M}}_{\mathbb{XY}})$
Sample moments

Figure 1. Basic ideas of the DAS and AMS measures. Specifically, $\widetilde{R}^2(\mathbb{X}^{\mathcal{M}}_{(k)}) = R^2(\mathbb{X}^{\mathcal{M}}_{(k)}) - \widehat{\Delta}_{(k)}$ is given in (2.5), and the AMS measure is given in (2.6).

we have $\mathcal{M} = \{j\}$. In this case the regression R-squared is equivalent to the square of the correlation between $\mathbb{Y}_{(k)}$ and $\mathbb{X}_{(k)j}$. As a result, the screening measure is the same as the SIS proposed by Fan and Lv (2008). For a qualitative variable with l levels, the signal strength of each level might be too weak to be detected by the SIS approach. As a result, it is more reasonable to evaluate such variables as a whole.

Remark 4. The proposed framework has a great potential for extensions. In particular, a practitioner might consider using different screening measures for different groups of variables. For instance, one may consider using R^2-based measure for quantitative variables and distance correlation (Li, Zhong, and Zhu 2012b) for qualitative variables. In this case, we could flexibly adjust the definition of Equation (2.2) to screen different groups of variables. The R-squared screening measure is also easy to extend to conditional sure independence screening (Fan and Lv 2008), pairwise interactions screening (Fan et al. 2016; Zhou et al. 2019) and many others according to the application scenarios. Last, the framework is easy to extend to generalized linear models. In this regards, we can revise the R-squared screening measure to adapt to generalized data types. Specifically, the likelihood based screening measure (Fan et al. 2010), distance correlation (Li, Zhong, and Zhu 2012b), model-free independence screening method (Zhou, Zhu, and Li 2020) can be used for this scenario.

Given the screening measure, feature screening is conducted as follows. For a quantitative variable $\mathbb{X}_j$, let $\mathcal{M} = \{j\}$ and denote $R^2_{\mathbb{X}_j} = R^2(\mathbb{X}^{\mathcal{M}}_{(k)})$ if only the kth subsample is employed. Then, for a given constant c_β, we can estimate $\mathcal{M}^\beta_T$ by $\widehat{\mathcal{M}}^\beta = \{1 \le j \le p_1 : R^2_{\mathbb{X}_j} \ge c_\beta\}$. Similarly, for a quantitative variable $\mathbb{Z}_j$, we use the notation $R^2_{\mathbb{Z}_j}$ to denote the marginal R-squared and estimate $\mathcal{M}^\gamma_T$ by $\widehat{\mathcal{M}}^\gamma = \{1 \le j \le p_2 : R^2_{\mathbb{Z}_j} \ge c_\gamma\}$, where c_γ is a prespecified constant. In the following, we discuss the proposed two screening measures in detail.

2.3.1. Debiased Average Screening

Based on the R-squared screening measure, we define a simple average screening (AVS) measure by taking the simple average over all subsamples,

$$R^2_{\text{AVS}}\left(\mathbb{X}^{\mathcal{M}}\right) = \frac{1}{B}\sum_{k=1}^{B} R^2\left(\mathbb{X}^{\mathcal{M}}_{(k)}\right). \tag{2.3}$$

This idea is widely used in literature (Kleiner et al. 2014; Sengupta, Volgushev, and Shao 2016) and is sufficiently flexible to extend to other screening measures. However, as we will show in the theoretical analysis, this approach produces a bias of order $O(n^{-1})$. The bias is ignorable as long as the subsample size n is large or the signal strength of important features is strong. However, increasing the subsample size requires addressing the hard disk a greater number of times, which is time consuming. To alleviate the bias issue while maintaining a compact subsampling size, we revise the simple average screening measure to reduce the bias.

We first introduce the jackknife debiased average screening measure. Denote $\mathbb{X}_{(k),-i} = (\widetilde{\mathbb{X}}_{(k)j} : j \neq i)^\top$ and $\mathbb{Y}_{(k),-i} = (\widetilde{Y}_{(k)j} : j \neq i)^\top$ as the kth subsample eliminating the ith subject. Correspondingly, define $\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}(k),-i} = (n-1)^{-1}(\mathbb{X}^{\mathcal{M}}_{(k),-i})^\top \mathbb{X}^{\mathcal{M}}_{(k),-i}$ $\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}\mathbb{Y}(k),-i} = (n-1)^{-1}(\mathbb{X}^{\mathcal{M}}_{(k),-i})^\top \mathbb{Y}_{(k),-i}$, and $\widehat{\sigma}^2_{y(k),-i} = (n-1)^{-1}||\mathbb{Y}_{(k),-i} - \overline{\mathbb{Y}}_{(k),-i}||^2$. Then the leave-one-out R-squared estimator is given as follows:

$$R^2\left(\mathbb{X}^{\mathcal{M}}_{(k),-i}\right) = \widehat{\sigma}^{-2}_{y(k),-i}\left(\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}\mathbb{Y}(k),-i}\right)^\top \left(\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}(k),-i}\right)^{-1} \left(\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}\mathbb{Y}(k),-i}\right). \tag{2.4}$$

The bias is estimated as $\widehat{\Delta}_{(k)} = n^{-1}(n-1)\sum_i R^2(\mathbb{X}^{\mathcal{M}}_{(k),-i}) - (n-1)R^2(\mathbb{X}^{\mathcal{M}}_{(k)})$. This leads to the jackknife debiased simple average screening (DAS) measure, defined as follows:

$$R^2_{\text{DAS}}\left(\mathbb{X}^{\mathcal{M}}\right) = \frac{1}{B}\sum_{k=1}^{B}\left\{R^2\left(\mathbb{X}^{\mathcal{M}}_{(k)}\right) - \widehat{\Delta}_{(k)}\right\}. \tag{2.5}$$

The bias of the above DAS measure is reduced substantially from $O(n^{-1})$ to $O(n^{-2})$. As a result, this approach allows for a smaller subsampling size while guaranteeing high accuracy.

2.3.2. Aggregated Moment Screening

The DAS measure is easy to extend to other screening measures using the same jackknife bias reduction procedure; hence, it is flexible. Beyond this approach, we note that the R-squared screening measure in Equation (2.2) is a nonlinear transformation of several simple moments. This motivates us to propose an AMS measure. Specifically, note that the R-squared (2.2) constitutes of three components, that is, $\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}(k)} = n^{-1}(\mathbb{X}^{\mathcal{M}}_{(k)})^\top \mathbb{X}^{\mathcal{M}}_{(k)}$, $\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}\mathbb{Y}(k)} = n^{-1}(\mathbb{X}^{\mathcal{M}}_{(k)})^\top \mathbb{Y}_{(k)}$, and $\widehat{\sigma}^2_{y(k)} = n^{-1}||\mathbb{Y}_{(k)} - \overline{\mathbb{Y}}_{(k)}||^2$. Each component is a simple moment estimator. Based on this observation, the AMS measure is designed as follows. First, we calculate the moment estimators of the above three components as the simple average over all the subsamples, $\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}} = B^{-1}\sum_{k=1}^{B}\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}(k)}$, $\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}\mathbb{Y}} = B^{-1}\sum_{k=1}^{B}\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}\mathbb{Y}(k)}$, and $\widehat{\sigma}^2_{\mathbb{Y}} = B^{-1}\sum_{k=1}^{B}\widehat{\sigma}^2_{y(k)}$. Then we define the AMS measure as

$$R^2_{\text{AMS}}\left(\mathbb{X}^{\mathcal{M}}\right) = \frac{1}{\widehat{\sigma}^2_{\mathbb{Y}}}\left(\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}\mathbb{Y}}\right)^\top \left(\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}}\right)^{-1}\left(\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{X}\mathbb{Y}}\right). \tag{2.6}$$

Similarly, for a set of qualitative variables collected in $\mathcal{M}$, one could define $\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{Z}}$, $\widehat{\Sigma}^{\mathcal{M}}_{\mathbb{Z}\mathbb{Y}}$, and $R^2(\mathbb{Z}^{\mathcal{M}})$ in the same way. The AMS measure in Equation (2.6) enjoys small bias and can be treated as a componentwise debiasing method (Li et al. 2020). When high correlation level is presented among the predictors of ultrahigh dimensional linear regression models, we recommend to use an iterative screening strategy (Fan and Lv 2008; Cho and Fryzlewicz 2012). The basic idea is to iteratively apply feature screening and post-variable-selection procedures to enhance the methodology power. We discuss this extension in details in Appendix A.1. In the following section, we further study the theoretical properties of the screening measures.

3. Theoretical Property

3.1. Convergence Properties Under RAS

To motivate the theoretical discussion, we first investigate the properties under the basic random sampling (RAS) scheme. Specifically, the subsampling procedure is conducted for B

rounds. In each round, we sample n observations from the data with replacement. As a result, the sampling probability for each data point is $1/N$. By slightly abusing the notation, we use $\mathbb{X}_{(k)}^{\mathcal{M}}$ and $\mathbb{Y}_{(k)}$ to denote the data of the kth subsampling procedure. Specifically, we denote $R^2_{\text{AMS}}(\mathbb{X}_j)$, $R^2_{\text{AVS}}(\mathbb{X}_j)$, $R^2_{\text{DAS}}(\mathbb{X}_j)$ as $R^2_{\mathbb{X}_j,\text{AMS}}$, $R^2_{\mathbb{X}_j,\text{AVS}}$, $R^2_{\mathbb{X}_j,\text{DAS}}$ for convenience if a single covariate $\mathbb{X}_j$ is considered. Without loss of generality, we let $N \geq nB$ in our following analysis.

Under the RAS setting, calculating the AMS measure is equivalent to conducting one-time subsampling with subsample size nB. Hence, it is straightforward to derive the theoretical properties of the AMS measure first. To this end, assume the following conditions.

Assumption 1 (Sub-Gaussian distribution). Assume the covariates X_{ij}, $\mathcal{Z}_{ijl}$ and ε_i independently follow sub-Gaussian distributions, that is, $E\{\exp(tX_{ij})\} \leq \exp(\sigma_x^2 t^2/2)$, $E\{\exp(t\mathcal{Z}_{ijl})\} \leq \exp(\sigma_z^2 t^2/2)$, and $E\{\exp(t\varepsilon_i)\} \leq \exp(\sigma_\varepsilon^2 t^2/2)$ for any $t > 0$. In addition, let $E(\varepsilon_i) = 0$, $E(X_{ij}) = 0$, and $\text{var}(Y_i) = \sigma_y^2 < \infty$.

Assumption 2 (Dimensionality for the AMS measure under RAS).

(a) (Quantitative covariates). Assume $\log p_1 \ll \min\{nBN^{-2\nu}, n^{1/2}BN^{-\nu}\}$ with $\nu \in [0, 1/2)$. In addition, assume $\log p_1 + \log B \ll n^{1/2}$.

(b) (Qualitative covariates). Let $\pi_{jl} = P(Z_j = l)$ and $\pi_{\min} = \min_{j,l} \pi_{jl}$. Assume $\log p_2 + \max_j \log l_j \ll \min\{nBN^{-2\nu} l_j^{-2}, n^{1/2}BN^{-\nu} l_j^{-1}, nB\pi_{\min}\}$, where $\nu \in [0, 1/2)$. In addition, assume $\log p_1 + \log B \ll n^{1/2}$.

The first condition assumes a sub-Gaussian distribution for all covariates and noise terms. Compared to the normality assumption on covariates (Fan and Lv 2008; Wang 2009; Wang, Kim, and Li 2013) in the feature screening literature, the sub-Gaussian assumption is more flexible. Assumption 2 is concerned with the dimensionality of the quantitative and qualitative covariates. The feature dimension is allowed to grow exponentially with the subsample size n and the number of subsampling times B. Specifically, with respect to the qualitative covariates, the number of levels (l_j) and the smallest ratio of all levels ($\pi_{\min}$) are also critical factors, which implies that the number of levels (i.e., l_j) should not be too large and that the features cannot be too sparse. Under the RAS, we illustrate the convergence property for the AMS measure.

Theorem 1. Assume Assumptions 1 and 2. (a) For a quantitative variable $\mathbb{X}_j$, we have $\max_j \left|R^2_{\mathbb{X}_j,\text{AMS}} - \mathcal{R}^2_{\mathbb{X}_j}\right| = O_p(N^{-\nu})$, where $\mathcal{R}^2_{\mathbb{X}_j} = \text{cor}(\mathbb{X}_j, \mathbb{Y})^2$. (b) For a quantitative variable $\mathbb{Z}_j$ with l_j levels, we have $\max_j \left|R^2_{\mathbb{Z}_j,\text{AMS}} - \mathcal{R}^2_{\mathbb{Z}_j}\right| = O_p(N^{-\nu})$, where $\mathcal{R}^2_{\mathbb{Z}_j} = \sigma_y^{-2} \sum_{l=1}^{l_j-1} \pi_{jl}^{-1} \sigma^2_{zy,jl}$ with $\sigma_y^2 = \text{var}(Y_i)$, $\pi_{jl} = P(\mathcal{Z}_{ijl} = 1)$, and $\sigma_{zy,jl} = E(\mathcal{Z}_{ijl} Y_i)$.

The proof of Theorem 1 is given in Appendix C.1. The results suggest that the resampling based R-squared values of all covariates converge uniformly to their population values.

Next, we investigate the convergence property of the AVS measure under RAS. Although the implementation of the AVS measure is simple, it has the disadvantage of producing a nonnegligible bias of $O(n^{-1})$. To make this point clear, we first study the theoretical properties of $R^2_{\mathbb{X}_j(k),\text{AVS}}$ for a single subsampling round k in the following Lemma.

Lemma 1. Assume Assumption 1. Then, we have $R^2_{\mathbb{X}_j(k),\text{AVS}} - \mathcal{R}^2_{\mathbb{X}_j} = \Delta_{xb} + \Delta_{xv}\{1 + o_p(1)\}$. Here, $\Delta_{xb} = c_1 n^{-1} + c_2 \max\{n^{-2}, N^{-1}\}\{1 + o(1)\}$ is the bias term, where c_1 and c_2 are finite constants. In addition, $\Delta_{xv} = O_p(n^{-1/2})$ with $E(\Delta_{xv}) = 0$ is the leading term for variance.

The proof of Lemma 1 is given in Appendix C.2.

Remark 5. According to Lemma 1, the leading bias term is of order $O(n^{-1})$ and the leading variance term is of order $O_p(n^{-1/2})$. By repeated sampling, we can manage to reduce the variance order; however, the bias will remain. If the signal strengths of all important features are strong, then it is easy to distinguish them from nonimportant features. However, when the signal strength is small, for example, $\min_{j \in \mathcal{M}_T^\beta} \mathcal{R}^2_{\mathbb{X}_j} = O(N^{-\nu})$, we must increase the subsample size n to ensure that $\Delta_{xb} \ll \min_{j \in \mathcal{M}_T^\beta} \mathcal{R}^2_{\mathbb{X}_j}$. Otherwise, the screening procedure will have poor performance.

To alleviate the bias issue of the AVS measure, we use the jackknife debiased procedure. The bias reduction effect is established in the following lemma.

Lemma 2. Under Assumption 1, we have $R^2_{\mathbb{X}_j(k),\text{DAS}} - \mathcal{R}^2_{\mathbb{X}_j} = \Delta_{xb2} + \Delta_{xv}\{1 + o_p(1)\}$. Here, $\Delta_{xb2} = c\max\{n^{-2}, N^{-1}\}\{1 + o(1)\}$ is the bias term, where c is a finite constant. In addition, $\Delta_{xv} = O_p(n^{-1/2})$ with $E(\Delta_{xv}) = 0$ is the leading variance term.

The proof of Lemma 2 is given in Appendix C.4. As shown by Lemma 2, the leading bias of the DAS measure is $O(n^{-2})$, which is much smaller than that of the AVS measure. Therefore, this approach allows for a smaller sample size to achieve competitive screening accuracy. To further obtain the uniform convergence property, we require the following condition.

Assumption 3 (Dimensionality for the DAS measure under RAS).

(a) (Quantitative covariates) Assume $\log p_1 + \log B \ll \min\{n^{1/3}, N^{1/4}\}$. Let $\log p_1 \ll \min\{nBN^{-2\nu}, n^{1/2}BN^{-\nu}\}$ for some $\nu \in [0, 1/2)$.

(b) (Qualitative covariates) Assume $\log p_2 + \log B + \max_j \log l_j \ll \min\{(n\pi_{\min})^{1/2}, n^{1/3}, N^{1/4}\}$. Further assume $\log p_2 + \max_j \log l_j \ll \min_j\{nBN^{-2\nu}\pi^2_{\min} l_j^{-2}, n^{1/2}BN^{-\nu}\pi_{\min} l_j^{-1}\}$ for some $\nu \in [0, 1/2)$.

The above dimensionality requirement is slightly more restrictive than Assumption 2. That is because it uses higher-order Taylor expansion for the theoretical properties, which increases the difficulty in analyzing the convergence property. Similarly, we establish the convergence property of the DAS measure, as follows.

Theorem 2. Under Assumptions 1 and 3, the following conclusions hold.

(a) It holds $\max_j |R^2_{\mathbb{X}_j,\mathrm{DAS}} - \mathcal{R}^2_{\mathbb{X}_j} - \Delta_{xb}| = O_p(N^{-\nu})$, where $\Delta_{xb} = O(n^{-2})$.
(b) It holds $\max_j |R^2_{\mathbb{Z}_j,\mathrm{DAS}} - \mathcal{R}^2_{\mathbb{Z}_j} - \Delta_{zb}| = O_p(N^{-\nu})$, where $\Delta_{zb} = O(n^{-2} \max_j \sum_l \pi_{jl}^{-1})$.

The proof of Theorem 2 is provided in Appendix C.5. The theoretical properties are established based on the AVS measure. For theoretical completeness, we discuss the convergence property of the AVS measure in Appendix B. Compared to that of the AVS measure, the bias order is reduced for quantitative variables. For qualitative variables, the bias reduction effect is also related to π_{jl}. Specifically, the features should not be too sparse to be detected. In summary, as long as Δ_{xb} and Δ_{zb} are of a smaller order than $N^{-\nu}$, the DAS measure can approximate $\mathcal{R}^2_{\mathbb{X}_j}$ and $\mathcal{R}^2_{\mathbb{Z}_j}$ uniformly on the order of $O_p(N^{-\nu})$.

3.2. Convergence Properties Under SAS

Subsequently, we discuss the theoretical properties for the screening method under the SAS setting. Under this sampling scheme, we no longer have the conditional independence of all sampling points given $\mathbb{Z}$, as in RAS. Instead, we have conditional independence of $\{\mathbb{X}_{(k)}, \mathbb{Z}_{(k)j}, \mathbb{Y}_{(k)}\}$ for $k = 1, \dots, B$. This scenario requires slightly more restrictive conditions on the subsampling size n. For the AMS measure, we assume the following condition.

Assumption 4 (Dimensionality for the AMS measure under SAS).

(a) (Quantitative covariates). There exists $\delta \in (0, 1/2)$ such that $\log p_1 \ll \min\{n^{1-2\delta} BN^{-2\nu}, Bn^{1/2-\delta} N^{-\nu}\}$ and $n^{2\delta} \gg \log p_1 + \log N$, where $\nu \in [0, 1/2)$.
(b) (Qualitative covariates). There exists $\delta \in (0, 1/2)$ such that $\log p_2 + \max_j \log l_j \ll \min_j\{n^{1-2\delta} BN^{-2\nu} l_j^{-2}, Bn^{1/2-\delta} N^{-\nu} l_j^{-1}, n^{1/2-\delta} B\pi_{\min}\}$ and $n^{2\delta} \gg \log p_2 + \max_j \log l_j + \log N$, where $\nu \in [0, 1/2)$.

Compared to the random sampling setting for the AMS measure, Assumption 4 imposes more restrictive assumptions on the subsample size n. Specifically, n should be sufficiently large to allow for high-dimensional features (i.e., $\log p_1 \ll n^{2\delta}$ and $\log p_2 \ll n^{2\delta}$).

Theorem 3. Assume Assumptions 1 and 4; then, the following hold (a) $\max_j |R^2_{\mathbb{X}_j,\mathrm{AMS}} - \mathcal{R}^2_{\mathbb{X}_j}| = O_p(N^{-\nu})$ and (b) $\max_j |R^2_{\mathbb{Z}_j,\mathrm{AMS}} - \mathcal{R}^2_{\mathbb{Z}_j}| = O_p(N^{-\nu})$.

The proof of Theorem 3 is given in Appendix C.6. The AMS measure under SAS enjoys uniform convergence, and no further bias correction procedure is required as long as B is sufficiently large. To establish the uniform convergence for the DAS measure under SAS, we require the following condition.

Assumption 5 (Dimensionality for the DAS measure under SAS).

(a) (Quantitative Covariates) There exists $\delta \in (0, 1/3)$ such that $\log p_1 \ll \min\{n^{1-2\delta} BN^{-2\nu}, Bn^{1/2-\delta} N^{-\nu}\}$ and $\log p_1 + \log N \ll \min\{n^{2\delta}, n^{2(1-3\delta)}, n^{3/2-3\delta}\}$ hold. In addition assume $\log p_1 + \log B \ll \min\{n^{1/3}, N^{1/4}\}$.
(b) (Qualitative Covariates) There exists $\delta \in (0, 1/3)$ such that $\log p_2 + \max_j \log l_j \ll \min_j\{n^{1-2\delta} BN^{-2\nu} l_j^{-2}, n^{1/2-\delta} BN^{-\nu} l_j^{-1}\}$ hold. In addition assume $\log p_2 + \max_j \log l_j + \log N \ll \min\{n^{2\delta}, n^{2(1-3\delta)}, n^{3/2-3\delta}, n^{3/2+2\delta}\pi_{\min}, (n\pi_{\min})^{1/2}, n^{5/8+\delta/4}\pi_{\min}^{1/4}\}$ and $\log p_2 + \max_j \log l_j + \log B \ll \min\{n^{1/3}, N^{1/4}\}$.

Similar to the RAS setting, in order for the DAS measure to work, we place additional restrictions on the subsampling size n under the SAS sampling scheme. We establish the theoretical properties in the following theorem.

Theorem 4. Assume Assumptions 1 and 5; then, the following conclusions hold.
(a) It holds $\max_j |R^2_{\mathbb{X}_j,\mathrm{DAS}} - \mathcal{R}^2_{\mathbb{X}_j} - \Delta_{xb}| = O_p(N^{-\nu})$, where $\Delta_{xb} = O(n^{-1})$.
(b) It holds $\max_j |R^2_{\mathbb{Z}_j,\mathrm{DAS}} - \mathcal{R}^2_{\mathbb{Z}_j} - \Delta_{zb}| = O_p(N^{-\nu})$, where $\Delta_{zb} = O(n^{-2} \sum_l \pi_{jl}^{-1})$.

The proof of Theorem 4 is given in Appendix C.8 and the results are consistent with Theorem 2 under the RAS setting.

3.3. Statistical Inference Under RAS and DC

Using the same computational procedure, both DAS and AMS can be obtained under the DC setting. In the view of estimation, the AMS under the DC setting is equivalent to the global estimator (when $nB = N$), thus it is optimal. However, the AMS method cannot support automatic statistical inference. On the other hand, both AVS and DAS estimators are able to provide a relatively complete toolbox for automatic statistical inference, which includes standard error estimation, quantile estimation, confidence interval construction, and many others. For illustration propose, we compare the statistical inferences based on the AVS method for both RAS and DC settings.

Take the standard error (SE) of $R^2_{\mathrm{AVS}}(\mathbb{X}_j)$ as an example. We could estimate the SE as follows. Let $\widehat{\theta}_{(k)} = (\widehat{\theta}_{xy(k)}, \widehat{\theta}_{x(k)}, \widehat{\theta}_{y(k)})^\top \in \mathbb{R}^3$, where $\widehat{\theta}_{xy(k)} = \mathbb{X}_{(k)j}^\top \mathbb{Y}_{(k)}/n$, $\widehat{\theta}_{x(k)} = \mathbb{X}_{(k)j}^\top \mathbb{X}_{(k)j}/n$, $\widehat{\theta}_{y(k)} = ||\mathbb{Y}_{(k)} - \overline{\mathbb{Y}}_{(k)}||^2/n$. As a result, we have $R^2_{\mathrm{AVS}}(\mathbb{X}_j) = B^{-1} \sum_{k=1}^B g(\widehat{\theta}_{(k)})$, where $g(\widehat{\theta}_{(k)}) = (\widehat{\theta}_{y(k)})^{-1} (\widehat{\theta}_{x(k)})^{-1} (\widehat{\theta}_{xy(k)})^2$. The SE^2 of $R^2_{\mathrm{AVS}}(\mathbb{X}_j)$ under the RAS setting can be estimated as follows:

$$\widehat{\mathrm{SE}}^2 = \frac{n}{B}\left(\frac{1}{nB} + \frac{1}{N}\right) \sum_{k=1}^{B} \left\{g(\widehat{\theta}_{(k)}) - R^2_{\mathrm{AVS}}(\mathbb{X}_j)\right\}^2. \quad (3.1)$$

Similarly, under the DC setting we estimate SE^2 by $\widehat{\mathrm{SE}}^2 = B^{-2} \sum_{k=1}^B \{g(\widehat{\theta}_{(k)}) - R^2_{\mathrm{AVS}}(\mathbb{X}_j)\}^2$. To investigate the theoretical properties of $\widehat{\mathrm{SE}}^2$, we present the following conclusion.

Theorem 5. Define $\Sigma_\theta = \mathrm{cov}(D_i)$ with $D_i = (X_{ij}Y_i, X_{ij}^2, (Y_i - \mu_y)^2)^\top$ and $\mu_y = E(Y_i)$. Let $\tau = \dot{g}(\theta)^\top \Sigma_\theta \dot{g}(\theta)$. Assume Assumption 1 and it holds,

(a) under RAS we have $\mathrm{SE}^2 = \tau(1/nB + 1/N)\{1 + o(1)\}$;
(b) under DC with $nB \le N$, we have $\mathrm{SE}^2 = \tau(nB)^{-1}\{1 + o(1)\}$;
(c) under both settings we have $\widehat{\mathrm{SE}}^2 = \mathrm{SE}^2\{1 + O_p(1/B^{1/2} + 1/N^{1/2})\}$.

The proof of Theorem 5 is given in Appendix C.9. By Theorem 5, we can conclude that $\widehat{\mathrm{SE}}^2$ is a consistent estimator and a reliable estimation of $\widehat{\mathrm{SE}}^2$ requires relatively large number of B. Under the DC setting, it must satisfy $B \le M = N/n$, while for RAS we could allow for a much larger subsampling rounds B which is not restricted by N and n. As a result, the statistical inference under RAS setting is more reliable than DC setting. We verify our findings in an extensive numerical study given in Appendix F.1.

3.4. Screening Consistency

The uniform convergence of the screening measures guarantees that the important features have higher ranks (of R-squared) than the nonimportant ones as long as the signal is sufficiently strong. This enables us to select the important features consistently using the R-squared screening measure. To clarify this statement, we next establish the sure screening property of $R^2_{\mathbb{X}_j}$ and $R^2_{\mathbb{Z}_j}$. To this end, we require the following conditions.

Assumption 6 (Correlation). Let $\mathcal{Z}^*_{ijl} = \mathcal{Z}_{ijl}/\sqrt{\pi_{jl}}$ and $\mathcal{Z}^*_i = (\mathcal{Z}^*_{ijl} : 1 \le j \le p_2, 1 \le l \le l_j - 1)$. In addition, define $\mathcal{X}^*_i = (X_i^\top, \mathcal{Z}_i^{*\top})^\top \in \mathbb{R}^q$ and $\Sigma = \mathrm{cov}(\mathcal{X}^*_i) \in \mathbb{R}^{q\times q}$, where $q = p_1 + \sum_j l_j - p_2$. Assume $\lambda_{\max}(\Sigma) \le \tau_{\max}$, where $\tau_{\max}$ is a positive constants. In addition, let $\lambda_{\min}(\Sigma^{(\mathcal{M}_T)}) > 0$, where $\mathcal{M}_T$ is the set of indexes of nonzero coefficients and $\Sigma^{(\mathcal{M}_T)}$ is the sub-matrix of Σ of important variables.

Assumption 7 (Minimum Signal). Let $R_{\min} \overset{\mathrm{def}}{=} \min\{\min_{j\in\mathcal{M}_T^\beta} \mathcal{R}^2_{\mathbb{X}_j}, \min_{j\in\mathcal{M}_T^\gamma} \mathcal{R}^2_{\mathbb{Z}_j}\} > 2c_\theta$, where $c_\theta = \max\{c_\beta, c_\gamma\}$. For the AVS measure, let $\min\{c_\theta, \max_j n^{-1}l_j\} \gg N^{-\nu}$. For the DAS measure, let $\min\{c_\theta, \max_j n^{-2}\sum_{l=1}^{l_j-1}\pi_{jl}^{-1}\} \gg N^{-\nu}$. For the AMS measure, let $c_\theta \gg N^{-\nu}$.

Assumption 6 restricts the correlations among the covariates; therefore, the eigenvalues of the covariance matrix behave properly. We require $\lambda_{\min}(\Sigma^{(\mathcal{M}_T)}) > 0$ to ensure the model identification. Next, Assumption 7 requires that the minimal signal strength of all important features is not too weak to be detected. Specifically, a lower ν implies a weaker signal, which places more restrictive conditions on the uniform convergence. To understand this assumption more intuitively, we further discuss in Lemma 3 a lower bound of $R_{\min}$ in a specific scenario, which relates to minimum nonzero coefficients. In addition, under the AVS and DAS settings, this value is also related to the subsample size n. Therefore, a sufficient number of subsamples should be used in each round to guarantee the detection of weak signals. Based on the above assumptions, we can obtain the following screening properties.

Theorem 6. (Screening Consistency under RAS) Let $m_{\max} = 4\tau_{\max}\sigma_y^2/R_{\min}$. Then under Assumptions 1–3, 6–7, we have

$$P\Big(\mathcal{M}_T^\beta \subset \widehat{\mathcal{M}}^\beta \text{ and } \mathcal{M}_T^\gamma \subset \widehat{\mathcal{M}}^\gamma\Big) \to 1, \qquad (3.2)$$

$$P\Big(\max\Big\{\big|\widehat{\mathcal{M}}^\beta\big|, \big|\widehat{\mathcal{M}}^\gamma\big|\Big\} < m_{\max}\Big) \to 1. \qquad (3.3)$$

The proof of Theorem 6 is given in Appendix D.1. First, (3.2) implies that all important features can be consistently selected under appropriate conditions. Subsequently, Equation (3.3) indicates the model size is be well controlled. Specifically, the model size is closely related to $\tau_{\max}$, σ_y^2, and $R_{\min}$. First, if $\tau_{\max}$ is large, the dependence among the features is higher; thus, it is more difficult to screen all relevant features. Next, if σ_y^2 is high, the signal-to-noise ratio will be low, which increases the difficulty of the screening task. Last, a lower $R_{\min}$ indicates weaker signal strengths. As a result, we need to include more features in $\widehat{\mathcal{M}}^\beta$ and $\widehat{\mathcal{M}}^\gamma$ to guarantee the screening consistency property. In the following Lemma, we further relate $R_{\min}$ to the minimum regression coefficient for important features under a specific scenario.

Lemma 3. Define $\mathcal{M}_T$ as the set of nonzero coefficients corresponding to $\mathcal{X}^*_i$ defined in Assumption 6. Suppose the elements of $\Sigma^{(\mathcal{M}_T)}$ are nonnegative and the nonzero model coefficients are positive. Then we have $R_{\min} \ge \min_{j\in\mathcal{M}_T}(\sum_{i\in\mathcal{M}_T}\sigma_{ij})^2 \theta^2_{\min}/\sigma_y^2$, where $\theta_{\min} = \min\{|\beta_{\min}|, |\gamma_{\min}|\}$ with $\beta_{\min} = \min_{j\in\mathcal{M}_T^\beta}|\beta_j|$ and $\gamma_{\min} = \min_{j\in\mathcal{M}_T^\gamma}\min_{1\le l\le l_j-1}\pi_{jl}^{1/2}\gamma_{jl}$.

The proof of Lemma 3 is given in Appendix D.2. In Lemma 3, we consider a special scenario that $\Sigma^{(\mathcal{M}_T)}$ has nonnegative elements and all nonzero parameters are positive. In this case, $R_{\min}$ is related to two important factors, i.e., $\theta_{\min}$ and $\min_{j\in\mathcal{M}_T}(\sum_{i\in\mathcal{M}_T}\sigma_{ij})^2/\sigma_y^2$. First, $\theta_{\min}$ is the minimum absolute nonzero model parameters. If $\theta_{\min}$ is large, the separation between zero and nonzero coefficients will be large therefore the screening power is increased. If $\min_{j\in\mathcal{M}_T}(\sum_{i\in\mathcal{M}_T}\sigma_{ij})^2/\sigma_y^2$ is high, then the correlation among important features is high. Therefore, it makes it easier to detect important features. Assumption 7 can be guaranteed if $\theta_{\min} \gg N^{-\nu/2}/\min_{j\in\mathcal{M}_T}(\sum_{i\in\mathcal{M}_T}\sigma_{ij})^2$ under this scenario, which implies the weakest signal approaches zero slowly as the sample size N increases.

Under the SAS sampling scheme, we can also establish the screening consistency result, as follows.

Theorem 7. (Screening Consistency under SAS) Let $m_{\max} = 4\tau_{\max}\sigma_y^2/R_{\min}$. Then, under Assumptions 1 and 4–7, we have

$$P\Big(\mathcal{M}_T^\beta \subset \widehat{\mathcal{M}}^\beta \text{ and } \mathcal{M}_T^\gamma \subset \widehat{\mathcal{M}}^\gamma\Big) \to 1, \qquad (3.4)$$

$$P\Big(\max\Big\{\big|\widehat{\mathcal{M}}^\beta\big|, \big|\widehat{\mathcal{M}}^\gamma\big|\Big\} < m_{\max}\Big) \to 1. \qquad (3.5)$$

The proof of Theorem 7 is given in Appendix D.3. As implied by the result, the SAS can perform as well as RAS but with much lower computational complexity. In the next section, we evaluate the finite sample performances and computational costs.

Remark 6. Theoretically, $m_{\max}$ is in the order of $O(N^{\nu})$ by using the Assumption 6–7, while it relies on known parameters. In practice, to implement the proposed method, we follow Fan and Lv (2008), choosing $\widehat{\mathcal{M}}^{\beta} = \{1 \le j \le p_1 : R^2_{\mathbb{X}_j}$ is the first d_β largest among all$\}$ and $\widehat{\mathcal{M}}^{\gamma} = \{1 \le j \le p_2 :$ $R^2_{\mathbb{Z}_j}$ is the first d_γ largest among all $\}$. By further assuming that $n/\log n \gg N^{\nu}$, we set $d_\beta = d_\gamma = [n/\log n]$ in our simulation, where $[r]$ denotes the integer part of r.

4. Numerical Study

4.1. Simulation Models and Settings

To evaluate the finite sample performance of the proposed method, we present two examples in this section. The first example focuses on quantitative covariates X_i, i.e., $Y_i = X_i^{\top}\beta + \varepsilon_i$, and the second focuses on qualitative covariates $\mathcal{Z}_{ij}$, i.e., $Y_i = \sum_{j=1}^{p} \mathcal{Z}_{ij}^{\top}\gamma_j + \varepsilon_i$. In each example, the feature dimension p is fixed at 1000. We also consider a more challenging case with larger dimension $p = 5 \times 10^4$. The main findings are similar and the details are given in Appendix F.2. In addition, the noise term ε_i is generated independently and identically from a normal distribution $N(0, \sigma^2)$. In each example, we perform feature screening procedures under both the RAS and SAS settings. The examples are given as follows.

Example 1 (Quantitative covariates). In this example, we consider an autoregressive-type correlation structure on covariates. Specifically, we generate the covariate X_{ij} from a multivariate normal distribution with mean $\mathbf{0}_p$ and $\mathrm{Cov}(X_{ij_1}, X_{ij_2}) = \rho^{|j_1 - j_2|}$ with $\rho \in [0, 1)$ for $(1 \le j_1, j_2 \le p)$. A larger ρ implies higher dependence among covariates, which in turn increases the difficulty of the feature screening task. The important feature set is given as $\mathcal{M}_T^{\beta} = \{1, \ldots, 50\}$ with $|\mathcal{M}_T^{\beta}| = 50$. Correspondingly, the regression coefficients of important features are set as $\alpha(-1)^{U_{1j}} U_{2j}$, where U_{1j} is generated from a Bernoulli distribution $B(0.4)$, U_{2j} is sampled from uniform distribution $U[1, 2]$, and α is a parameter controlling the signal strength. We set $\alpha = 0.03$, $\rho = 0.8$ and $\sigma = 1$. In addition, for this example, we also evaluate a case when the signal is relatively stronger by setting $\alpha = 0.04$, $\rho = 0.1$, $\sigma = 0.4$. The results for stronger signal case are given in Appendix F.3.

Example 2 (Qualitative covariates). In this example, we generate the qualitative covariates based on the setting in Example 1. Specifically, we first generate a set of quantitative covariates X_{ij} from a multivariate normal distribution with mean $\mathbf{0}_p$ and $\mathrm{Cov}(X_{ij_1}, X_{ij_2}) = \rho^{|j_1 - j_2|}$ with $\rho = 0.8$ for $(1 \le j_1, j_2 \le p)$, moreover, we let $\sigma = 1$ as in Example 1. Given the important feature set $\mathcal{M}_T^{\gamma} = \{1, \ldots, 5\}$, we generate the qualitative variables $\mathcal{Z}_{ij}$, as follows. For the important variables, we set $l_j = j + 1$ with $1 \le j \le 5$. For the nonimportant covariates, we set $l_j = 3$ with $j \ge 6$. For the jth variable, define $\pi_j = (P(\mathcal{Z}_{ijk} = 1) : 1 \le k \le l_j)^{\top}$. We consider two typical cases for π_j. First, for $j = 1, 2$, we evaluate the case where the levels of the qualitative variables are relatively balanced. Specifically, we set $\pi_1 = (0.5, 0.5)^{\top}$ and $\pi_2 = (0.3, 0.3, 0.4)^{\top}$. Next, for the three other important variables we evaluate the unbalanced situations, that is, $\pi_3 = (0.1, 0.2, 0.3, 0.4)^{\top}$, $\pi_4 = (0.1, 0.1, 0.3, 0.35, 0.15)^{\top}$, and $\pi_5 = (0.1, 0.1, 0.1, 0.2, 0.2, 0.5)^{\top}$. Last, for the remaining covariates, we set $\pi_j = (0.1, 0.2, 0.7)^{\top}$. The qualitative variables are then generated as $\mathcal{Z}_{ijk} = 1$ $(1 \le k \le l_j - 1)$ if $q_{j,k-1} \le X_{ij} < q_{j,k}$, where $q_{j,k}$ is the $\alpha_k = \sum_{m=1}^{k} \pi_{jk}$th quantile of the standard normal distribution. As a consequence, the dependence of the qualitative variables is embedded through $\{X_{ij}\}$. Finally, the regression coefficients of the important features are set to $\nu(-1)^{U_{1jk}} U_{2jk}$ for $1 \le k \le l_j - 1$, where $U_{1jk} \sim B(0.4)$, $U_{2jk} \sim U[1, 2]$, and $\nu = 0.03$ in the simulation.

Table 1. Critical model parameters (i.e. $\tau_{\max}$, σ_y^2 and $R_{\min}$) of Example 1 and Example 2.

	$\tau_{\max}$	σ_y^2	$R_{\min}$
Example 1	9.00	1.24	2.84×10^{-5}
Example 2	51.27	1.01	3.84×10^{-3}

To better understand the simulation setting, we present several critical model parameters (i.e. $\tau_{\max}$, σ_y^2 and $R_{\min}$) in Table 1. In the first example $R_{\min}$ is smaller, hence, the signal of the nonzero coefficients is weaker. In the second example $\tau_{\max}$ is larger, which implies a higher dependence level among covariates.

4.2. Performance Measurements and Simulation Results

For each example, we set the sample size as $N = 10^5$ and 10^6. Once the whole dataset is generated, they are placed as a single file. The sizes of the files vary from 0.9 to 19.7 GB on the hard drive, and can hardly be read into computer memory as a whole. For a reliable evaluation, we replicate the experiment a total of $M = 100$ times. All computations are performed using Python 3.7.

To gauge the finite-sample performance, we employ the following measurements. First, we evaluate the numerical performance of parameter estimation. Specifically, for the mth replication, we calculate $R^{2(m)}_{\mathrm{AVS}}$, $R^{2(m)}_{\mathrm{DAS}}$, and $R^{2(m)}_{\mathrm{AMS}}$. Take the AVS measure as an example. The bias and the standard error (SE) are estimated as $\mathrm{Bias}_{\mathrm{AVS}} = |M^{-1}\sum_{m=1}^{M}(R^{2(m)}_{\mathrm{AVS}} - R^2)|$ and $\mathrm{SE}_{\mathrm{AVS}} = |M^{-1}\sum_{m=1}^{M}(R^{2(m)}_{\mathrm{AVS}} - \overline{R}^2_{\mathrm{AVS}})^2|^{1/2}$, respectively, where $\overline{R}^2_{\mathrm{AVS}} = M^{-1}\sum_{m=1}^{M} R^{2(m)}_{\mathrm{AVS}}$. In addition, the root-mean-squared error (RMSE) is calculated as $\mathrm{RMSE}_{\mathrm{AVS}} = \{M^{-1}\sum_{m=1}^{M}(R^{2(m)}_{\mathrm{AVS}} - R^2)^2\}^{1/2}$. The measurements can be defined similarly for the DAS and AMS measures.

Second, we evaluate the accuracy of the feature screening procedure according to the ranking consistency. In this work, ranking consistency is evaluated by the AUC measure, which is calculated as follows. Take the AVS measure as an example. For the m-th replication, we divide all the features into two sets: the positive set, which includes all the important features (i.e. $\mathcal{M}_T^{\beta}$), and the negative set, defined as $\widetilde{\mathcal{M}}_T^{\beta}$, which includes all the nonimportant features (i.e., $\widetilde{\mathcal{M}}_T^{\beta} = \{j : j \notin \mathcal{M}_T^{\beta}\}$). Then the AUC measure (in the mth experiment) is expressed as follows:

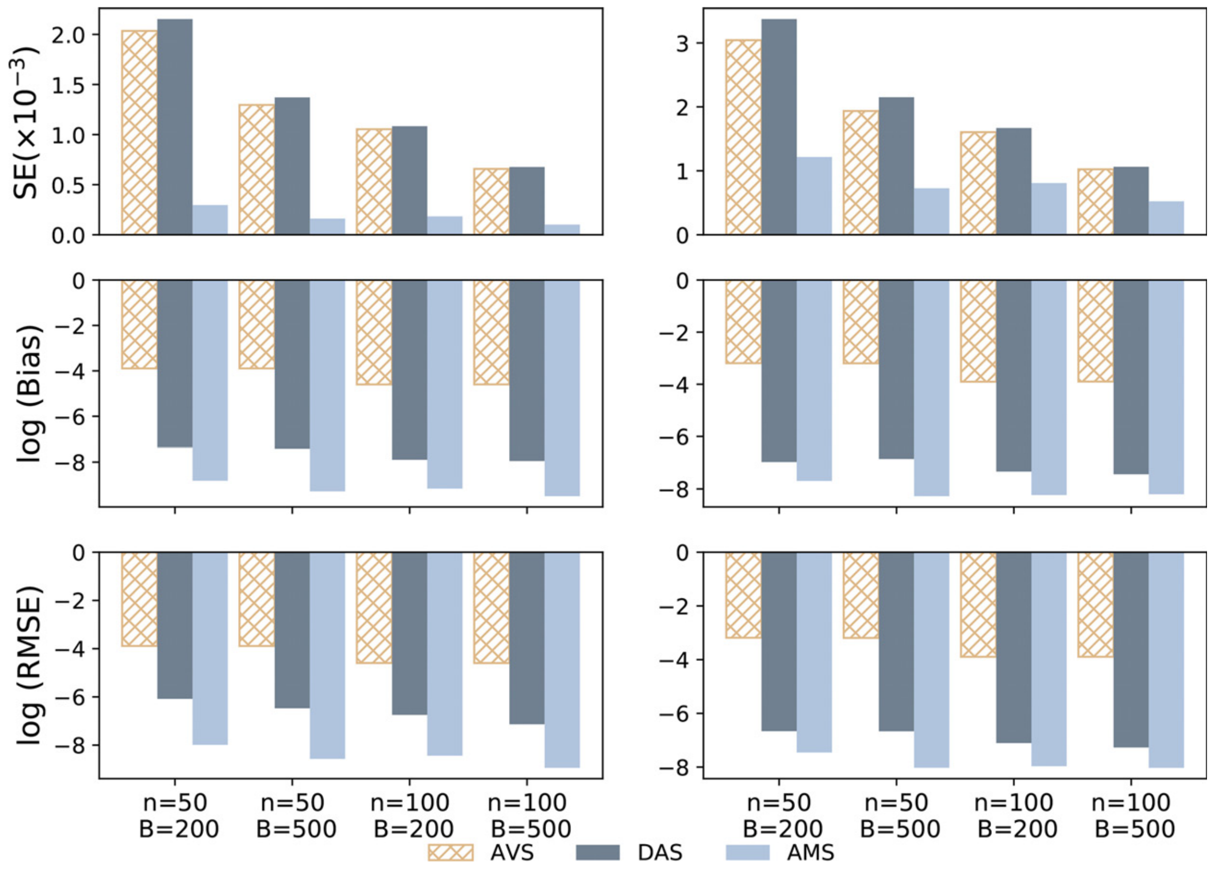

Figure 2. Bar chart of SE, log(Bias) and log(RMSE) values for the AVS, DAS and AMS measures for different (n, B) under the SAS sampling scheme for Example 1 (left panels) and Example 2 (right panels). The sample size N is fixed to $N = 10^5$.

$$
\begin{aligned}
\mathrm{AUC}_{\mathrm{AVS}}^{(m)} = 1 - \Bigg(& \frac{1}{|\mathcal{M}_T^{\beta}||\widetilde{\mathcal{M}}_T^{\beta}|} \sum_{j^+ \in \mathcal{M}_T^{\beta}} \sum_{j^- \in \widetilde{\mathcal{M}}_T^{\beta}} \\
& \times \Big[\mathbf{I}\left\{ R_{\mathrm{AVS}}^{2(m)}\left(\mathbb{X}_{j^+}\right) < R_{\mathrm{AVS}}^{2(m)}\left(\mathbb{X}_{j^-}\right) \right\} \\
& + 0.5\mathbf{I}\left\{ R_{\mathrm{AVS}}^{2(m)}\left(\mathbb{X}_{j^+}\right) = R_{\mathrm{AVS}}^{2(m)}\left(\mathbb{X}_{j^-}\right) \right\} \Big] \Bigg) \overset{\mathrm{def}}{=} 1 - Q_{\mathrm{AVS}}^{(m)}. \quad (4.1)
\end{aligned}
$$

Equation (4.1) was proposed by Hanley and McNeil (1982), and it has been widely used to measure the ranking quality of algorithms (Herlocker et al. 2004; Elith* et al. 2006; McKinney et al. 2020). In fact, for each pair in the positive set and negative set, if the rank of the screening measure is inconsistent with that of the true features (i.e., $R_{\mathrm{AVS}}^{2(m)}(\mathbb{X}_{j^+}) \le R_{\mathrm{AVS}}^{2(m)}(\mathbb{X}_{j^-})$), the rank loss $Q_{\mathrm{AVS}}^{(m)}$ will increase by one or a half unit as a penalty. Consequently, when the ranks of the screening measures and true features are completely consistent, we have AUC $= 1$; otherwise, AUC decreases as the ranking consistency decreases. Further define AUC of the AVS measure as $\mathrm{AUC}_{\mathrm{AVS}} = M^{-1} \sum_{m=1}^{M} \mathrm{AUC}_{\mathrm{AVS}}^{(m)}$. Similar measurements are reported for the two other screening measures. Last, the average sampling time costs are recorded under the RAS and SAS sampling schemes, respectively, as TC_{RAS} and TC_{SAS} to evaluate the computational efficiency.

We summarize the Bias, SE and RMSE in Figure 2 and present the rank consistency results in Table 2. Since the statistical performance under SAS is comparable to that under RAS in all settings, we report the results only for SAS in Figure 2. The detailed results for RAS are given in Appendix F.3. First, for a fixed sample size N, SE decreases when either n or B increases (as long as $N \ge nB$). Among the three screening measures, AMS has smaller SE values than the AVS and DAS methods. Next, as expected, the Bias decreases as the subsample size n increases but is not related to B. In addition, after the jackknife debiasing procedure, the bias of the DAS measure is much smaller than that of the AVS measure. Consequently, the RMSE value of AMS is the smallest, while that of AVS is the largest. This result is consistent with our theoretical findings in Lemma 1 and 2.

Subsequently, we evaluate the screening accuracy with respect to the AUC criterion. The AMS measure outperforms the other two screening measures across all settings. The performance of the DAS measure is comparable to that of the AMS measure, especially for Example 2 (i.e., the qualitative case). For instance, for $N = 10^5$, $n = 50$, and $B = 500$, the AUCs of the AMS and DAS measures are larger than 99.98% under the both RAS and SAS sampling schemes, while the AUC of the AVS measure is only 80%. Last, in terms of the computational efficiency, the procedure is less time consuming under SAS than under RAS especially when n and B are large. For instance, with $N = 10^6$, $n = 100$, and $B = 500$ in Example 1, the average computational time under SAS is 44.43s, while that under RAS is 158.80 sec.

4.3. Real Data Analysis With Airline Dataset

In this section, we use a U.S. airline dataset to illustrate the proposed method. The dataset is available on the official website

Table 2. Simulation results for Example 1 and 2 under the RAS and SAS sampling schemes. The numerical performance is evaluated for different sample sizes $N(\times 10^5)$, subsample sizes n and numbers of subsamples B. For the AVS, DAS, and AMS measures, the AUC values are reported as the screening accuracy criterion. Finally, the average time cost of sampling is also reported.

N	n	B	$\mathrm{AUC_{RAS}}(\%)$			TC_{RAS}	$\mathrm{AUC_{SAS}}(\%)$			TC_{SAS}
			AVS	DAS	AMS		AVS	DAS	AMS	
					Example 1					
1	50	200	96.60	96.61	98.80	45.63	95.66	95.74	98.86	16.68
		500	97.26	97.40	98.93	116.38	96.67	96.83	98.99	41.28
	100	200	97.57	97.63	98.95	63.70	97.35	97.44	98.91	17.32
		500	98.28	98.38	98.90	158.80	97.96	98.04	98.91	44.43
10	50	200	96.65	96.57	99.05	53.92	96.65	96.61	98.88	15.75
		500	97.48	97.48	99.05	121.93	97.83	97.88	99.08	39.39
	100	200	97.70	97.68	99.25	56.26	97.88	97.85	99.10	17.68
		500	98.46	98.43	99.17	139.87	98.65	98.62	99.30	44.20
					Example 2					
1	50	200	80.05	98.78	100.00	55.77	80.13	98.70	100.00	30.69
		500	80.00	99.98	100.00	140.59	80.00	99.98	100.00	77.26
	100	200	88.43	100.00	100.00	67.27	88.75	100.00	100.00	33.64
		500	87.39	100.00	100.00	167.88	88.56	100.00	100.00	85.53
10	50	200	80.18	98.89	100.00	74.99	80.15	99.13	100.00	30.60
		500	80.00	99.98	100.00	172.32	80.00	99.99	100.00	76.98
	100	200	89.87	100.00	100.00	79.44	89.70	100.00	100.00	34.29
		500	88.83	100.00	100.00	200.82	90.11	100.00	100.00	86.67

Table 3. Estimation results for airline data under the RAS and SAS sampling schemes. The numerical performance is evaluated for different subsample sizes n and numbers of subsamples B. For the AVS, DAS, and AMS measures, the SE, Bias, RMSE, and AUC are reported. Finally, the average time cost of sampling is also reported.

Sch	B	SE $(\times 10^{-3})$			Bias $(\times 10^{-2})$			RMSE $(\times 10^{-2})$			AUC(%)			TC
		R^2_{AVS}	R^2_{DAS}	R^2_{AMS}	R^2_{AVS}	R^2_{DAS}	R^2_{AMS}	R^2_{AVS}	R^2_{DAS}	R^2_{AMS}	R^2_{AVS}	R^2_{DAS}	R^2_{AMS}	
							$n = 50$							
RAS	100	11.46	12.62	10.53	4.261	0.446	0.513	4.456	1.489	1.176	87.86	89.08	89.39	8.54
	200	7.751	8.604	7.009	4.141	0.344	0.361	4.235	1.092	0.792	87.95	89.26	89.79	17.70
	500	5.029	5.573	4.567	4.156	0.331	0.370	4.196	0.809	0.591	88.13	89.42	90.89	43.40
SAS	100	11.47	12.71	10.43	4.193	0.363	0.453	4.392	1.484	1.141	87.86	89.04	89.31	1.17
	200	7.973	8.831	7.331	4.148	0.341	0.357	4.249	1.118	0.818	88.07	89.42	89.96	2.39
	500	5.179	5.726	4.613	4.191	0.345	0.407	4.233	0.824	0.618	88.16	89.44	90.91	6.02
							$n = 100$							
RAS	100	7.485	7.890	6.864	2.172	0.173	0.199	2.347	0.871	0.719	88.88	89.13	89.95	9.36
	200	5.534	5.824	5.107	2.165	0.145	0.177	2.265	0.670	0.542	89.19	89.41	90.65	18.49
	500	3.293	3.478	2.989	2.200	0.146	0.212	2.235	0.445	0.368	89.47	89.75	91.68	46.44
SAS	100	7.670	8.100	7.100	2.168	0.166	0.200	2.352	0.891	0.741	89.00	89.31	89.88	1.29
	200	5.363	5.658	4.959	2.170	0.148	0.186	2.263	0.652	0.532	89.24	89.43	90.59	2.68
	500	3.576	3.769	3.292	2.191	0.139	0.207	2.233	0.472	0.390	89.41	89.62	91.71	6.54

of the American Statistical Association at *http://stat-computing.org/dataexpo/2009*. The airline dataset contains information about commercial flights in the United States from 1987 to 2008. After basic data cleaning, we keep the flight information from 2004 to 2008 with 12 variables. The variables are: ArrTime (actual arrive time), Year, Month, DayofMonth, DayofWeek, CRSElapsedTime (scheduled elapsed time), CRSArrTime, Actual ElapsedTime, Distance, UniqueCarrier, Dest and Origin.

For the analysis, we use the ArrTime as the response and generate corresponding covariates from the other variables. First we split the cities in Origin according to the states. Next, we keep the top 10 states and group cities in the other states together as one state. Then, for each state, we keep the top 2 cities and group the others as one city. Via this procedure, we generate 10 categorical variables from the variable Origin. Similar procedures are applied to the variable Dest. Finally, we generate the 1-40th lags of the response ArriveTime and predictor ActualElapsedTime to detect lag effects. As a result, we derive 109 predictors of $N = 3.27 \times 10^7$ records and the size of the dataset is 27.6 GB. Detailed variable information is provided in Appendix F.4; see Table F.3 (supplementary material), and all the numerical variables are standarized to a mean of 0 and a variance of 1.

To evaluate the screening accuracy, we treat the R-squared screening measure using the whole dataset as the gold standard. The comparison is performed with the three screening measures (AVS, DAS, and AMS) using the same procedure as in the simulation study. Since in the real data analysis we do not know the ground truth (i.e., $\mathcal{M}_T^{\beta}$ and $\mathcal{M}_T^{\gamma}$), we revise the AUC measure as follows (for the AVS measures, for example),

$$\mathrm{AUC_{AVS}} = 1 - \left(\frac{1}{p^2-p}\sum_{\substack{j_1\neq j_2\\ j_1=1}}^{p}\sum_{j_2=1}^{p}\mathbf{I}\left\{R^2_{\mathrm{AVS}}(\mathbb{X}_{j_1}) < R^2_{\mathrm{AVS}}(\mathbb{X}_{j_2})\right\} \times \mathbf{I}\left\{R^2(\mathbb{X}_{j_1}) > R^2(\mathbb{X}_{j_2})\right\}\right),$$

where $R^2(\mathbb{X}_j)$ is the R-squared measure for the jth feature using the whole dataset. The detailed results are summarized in Table 3.

The performance under RAS and SAS is comparable in all settings, while the time cost of SAS is approximately 1/8 that of the RAS. Second, the SE decreases when either n or B increases, while the Bias decreases only when n increasing. Comparing the three measures, AMS has the smallest SE and Bias. In addition, the Biases of DAS and AMS measures are substantially smaller than that of the AVS measure. In terms of the screening accuracy, both the AMS and DAS measures have higher AUC values than the AVS measure, with AMS having the largest AUC. For instance, in the case of $B = 500, n = 50$ (under SAS), the AUC of the AMS measure is 90.91%, which is larger than that of the DAS (89.44%) and AVS (88.16%) methods.

5. Discussion

In this article, we develop a subsampling method for feature screening with massive datasets. Three R-squared-based screening measures are investigated. According to both theoretical and empirical studies, the DAS and AMS measures show advantages in terms of reducing biases. To reduce the sampling cost, we further consider a novel sequential sampling method in place of the simple random sampling. The theoretical properties are rigorously established. In practice, the accuracy of the feature screening methods is comparable, and the sequential sampling approach is more computationally efficient.

To conclude the article, we provide several topics for future study. First, other famous feature screening methods such as the forward screening measure (Wang 2009), RRCS (Li et al. 2012a), and DC-SIS (Li, Zhong, and Zhu 2012b) can be incorporated into our subsampling approach. Second, variable selection is another important topic in the regime of high-dimensional modeling, subsampling methods for variable selection should be designed and investigated. Last, the proposed subsampling approach cannot be applied for dependent data (e.g., time series and spatial data). As a consequence, a subsampling method that preserves the dependence structure of the data with low computational cost should be developed.

Supplementary Materials

Supplementary_Material.pdf: This document provides the extensions of the proposed method, the proofs of the theoretical results in the main text, and some additional simulation results. Appendix A reports some extensions and discussions of the proposed method. Appendix B contains the detailed proofs of the theoretical results of the AVS measure. Appendix C contains the detailed proofs of the main theorems and Lemmas developed in section 3.1-3.3 of the main text. In particular, it contains the proofs of theorems 1, 2, 3, 4, and 5 and Lemmas 1 and 2 of the main text. Appendix D contains the detailed proofs of screening consistency developed in sections 3.4 of the main text. In particular, it contains the proofs of theorems 5 and 6 and Lemma 3 of the main text. Appendix E provides technical lemmas which are useful to prove the results in section 3 of the main text. Finally, Appendix F contains some additional numerical results.

Code.zip: This file is the python code for the proposed method. Please see the "README.md"ž in the file for using the code.

Funding

Xuening Zhu is supported by the National Natural Science Foundation of China (nos. 11901105, 71991472, U1811461), and the Shanghai Sailing Program for Youth Science and Technology Excellence (19YF1402700). The research of Rui Pan is supported by National Natural Science Foundation of China (NSFC, 11601539, 11631003), and the Emerging Interdisciplianry Project of Central University of Finance and Economics. Hansheng Wang's research is partially supported by National Natural Science Foundation of China (No. 11831008).

References

Barut, E., Fan, J., and Verhasselt, A. (2016), "Conditional Sure Independence Screening," *Journal of the American Statistical Association*, 111, 1266–1277. [1892]

Chang, X., Lin, S.-B., and Wang, Y. (2017), "Divide and Conquer Local Average Regression," *Electronic Journal of Statistics*, 11, 1326–1350. [1892]

Cho, H. and Fryzlewicz, P. (2012), "High Dimensional Variable Selection via Tilting," *Journal of the Royal Statistical Society*, Series B, 74, 593–622. [1895]

Elith*, J., H. Graham*, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., et al. (2006), "Novel Methods Improve Prediction of Species Distributions from Occurrence Data," *Ecography*, 29, 129–151. [1900]

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557. [1894]

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020), *Statistical Foundations of Data Science*, Boca Raton, FL: CRC press. [1892]

Fan, J. and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [1892,1895,1896,1899]

Fan, J., Song, R., et al. (2010), "Sure Independence Screening in Generalized Linear Models with NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1895]

Fan, J., Wang, D., Wang, K., and Zhu, Z. (2019), "Distributed Estimation of Principal Eigenspaces," *Annals of Statistics*, 47, 3009. [1892]

Fan, Y., Kong, Y., Li, D., and Lv, J. (2016), "Interaction Pursuit With Feature Screening and Selection," arXiv:1605.08933. [1895]

Hanley, J. A., and McNeil, B. J. (1982), "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, 143, 29–36. [1900]

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004), "Evaluating Collaborative Filtering Recommender Systems," *ACM Transactions on Information Systems (TOIS)*, 22, 5–53. [1900]

Jordan, M. I., Lee, J. D., and Yang, Y. (2019), "Communication-Efficient Distributed Statistical Inference," *Journal of the American Statistical Association*, 526, 668–681. [1892]

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014), "A Scalable Bootstrap for Massive Data," *Journal of the Royal Statistical Society*, Series B, 76, 795–816. [1892,1895]

Li, G., Peng, H., Zhang, J., Zhu, L., et al. (2012a), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, 40, 1846–1877. [1892,1902]

Li, R., Zhong, W., and Zhu, L. (2012b), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [1892,1895,1902]

Li, X., Li, R., Xia, Z., and Xu, C. (2020), "Distributed Feature Screening via Componentwise Debiasing." *Journal of Machine Learning Research*, 21, 1–32. [1892,1893,1895]

Ma, P., Mahoney, M. W., and Yu, B. (2015), "A Statistical Perspective on Algorithmic Leveraging," *The Journal of Machine Learning Research*, 16, 861–911. [1892]

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al. (2020), "International Evaluation of an AI System for Breast Cancer Screening," *Nature*, 577, 89–94. [1900]

Pan, R., Wang, H., and Li, R. (2016), "Ultrahigh-Dimensional Multiclass Linear Discriminant Analysis by Pairwise sure Independence Screening," *Journal of the American Statistical Association*, 111, 169–179. [1892]

Pan, R., Zhu, Y., Guo, B., Zhu, X., and Wang, H. (2020), "A Sequential Addressing Subsampling Method for Massive Data Analysis With Memory Constraint," Working Paper. [1892,1893]

Sengupta, S., Volgushev, S., and Shao, X. (2016), "A Subsampled Double Bootstrap for Massive Data," *Journal of the American Statistical Association*, 111, 1222–1232. [1892,1895]

Shamir, O., Srebro, N., and Zhang, T. (2014), "Communication-Efficient Distributed Optimization Using an Approximate Newton-Type

Method," in *International Conference on Machine Learning*, pp. 1000–1008. [1892]

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [1892,1896,1902]

——— (2012), "Factor Profiled Sure Independence Screening," *Biometrika*, 99, 15–28. [1892]

——— (2019), "More Efficient Estimation for Logistic Regression with Optimal Subsamples," *Journal of Machine Learning Research*, 20, 1–59. [1892]

Wang, H., Yang, M., and Stufken, J. (2019), "Information-Based Optimal Subdata Selection for Big Data Linear Regression," *Journal of the American Statistical Association*, 114, 393–405. [1892]

Wang, H., Zhu, R., and Ma, P. (2018), "Optimal Subsampling for Large Sample Logistic Regression," *Journal of the American Statistical Association*, 113, 829–844. [1892]

Wang, L., Kim, Y., and Li, R. (2013), "Calibrating Non-Convex Penalized Regression in Ultra-High Dimension," *Annals of Statistics*, 41, 2505. [1896]

Wu, Y. and Yin, G. (2015), "Conditional Quantile Screening in Ultrahigh-Dimensional Heterogeneous Data," *Biometrika*, 102, 65–76. [1892]

Yu, J., Wang, H., Ai, M., and Zhang, H. (2020), "Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators with Massive Data," *Journal of the American Statistical Association*, 1–29, DOI: 10.1080/01621459.2020.1773832. [1892]

Zhou, M., Dai, M., Yao, Y., Liu, J., Yang, C., and Peng, H. (2019), "BOLT-SSI: A Statistical Approach to Screening Interaction Effects for Ultra-High Dimensional Data," arXiv:1902.03525. [1895]

Zhou, T., Zhu, L., and Li, R. (2020), "Model-Free Forward Regression via Cumulative Divergence," *Journal of the American Statistical Association*, 531, 1393–1405. [1892,1895]