Statistica Sinica Preprint No: SS-2024-0248

Title	Hybrid Denoising-screening for High-dimensional Contaminated Data
Manuscript ID	SS-2024-0248
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0248
Complete List of Authors	Liming Wang, Peng Lai, Chen Xu and Xingxiang Li
Corresponding Authors	Xingxiang Li
E-mails	lxxwlm2013@xjtu.edu.cn
Notice: Accepted author version.	

Statistica Sinica

HYBRID DENOISING-SCREENING FOR HIGH-DIMENSIONAL CONTAMINATED DATA

Liming Wang[1], Peng Lai[1], Chen Xu[2,3] and Xingxiang Li[3]

[1] *Nanjing University of Information Science & Technology,*

[2] *Peng Cheng Laboratory and* [3] *Xi'an Jiaotong University*

Abstract: Feature screening is an effective tool to eliminate irrelevant features in high-dimensional analysis. When a high-dimensional dataset is contaminated with noisy observations, the conventional screening methods may lead to a poor screening accuracy. To tackle this problem, one practical strategy is to remove noisy observations and irrelevant features simultaneously. In this paper, we propose a novel hybrid denoising-screening (HDS) procedure for high-dimensional contaminated data. The new method is built upon a dual sample-feature L_0 fitting procedure, which precisely controls both numbers of observations and features to be retained for the analysis. In the HDS process, only clean observations are selected and the joint effects between features are naturally accounted. These merits give HDS an edge to outperform the existing screening methods when faced with contaminated data. The promising performance of the method is supported by both theories and numerical examples.

Key words and phrases: Contaminated data analysis, joint feature screening, noise detection, alternating optimization strategy, sure screening property.

1. Introduction

With the rapid advances in information technology, high-dimensional datasets are ubiquitous in diverse areas of scientific research. When plenty of redundant (irrelevant) features are contained, high-dimensional data analysis may become quite challenging due to computational cost, statistical accuracy, and algorithmic stability (Fan et al., 2009). To address these challenges, one practical strategy is to screen out the redundant features in preparation for a more elaborate analysis. This pre-processing procedure is referred to as feature screening, which has gained much popularity during the past decade. As a seminal work, Fan and Lv (2008) proposed a marginal screening approach called sure independence screening (SIS) for ultrahigh-dimensional linear models. The screening operation has been extended, for example, to generalized linear models (Fan and Fan, 2008; Fan et al., 2009; Fan and Song, 2010), additive models (Fan et al., 2011), and hazard regression (Gorst-Rasmussen and Scheike, 2013). In the same spirit, various model-free screening methods have been built on marginal correlations between the response and the features (Zhu et al., 2011; Li et al., 2012). In addition, researchers have attempted to improve the marginal approaches by incorporating the joint effects among features into the screening process. In particular, Fan and Lv (2008) proposed an iterative SIS procedure (ISIS). Wang (2009) developed a forward regression screening (FRS) procedure. Xu and Chen (2014) provided the sparsity-restricted maximum likelihood estimation (SMLE) for generalized linear models. Wang and Leng (2016) introduced a high-dimensional ordinary least-squares projection (HOLP) for the screening technique. Zhou et al. (2020) proposed a model-free forward screening based on a new metric named cumulative divergence to characterize the functional dependence between predictors and the response variable.

Feature screening has been demonstrated as an attractive strategy for high-dimensional data. Most existing screening methods rely on the assumption that all observations are clean and useful. However, this ideal assumption can be unrealistic in many applications, where a dataset often contains a large number of noisy observations. Hampel et al. (2011) estimated that a routine dataset may contain about 1%–10% (or more) contaminated data. Noisy observations differ from clean ones; this leads to an inaccuracy of direct analysis with clean model assumption. For example, when making real estate market forecast analyses, some invalid and biased observations come from non-target property information, like apartments and warehouses; this may cause distorting results.

When the high-dimensional data set is contaminated, noisy observations also impose some negative impacts on classic feature screening methods, including model-based or model-free screenings. For model-based screenings, the noises seriously violate the model assumptions; this leads to the relationship between response and features becoming distorted due to the extreme values of noisy observations. As a result, the model-based screening techniques, no matter marginal or joint, tend to lose effectiveness. For some model-free screening methods, although the effect of extreme values of noisy observations may be controlled by some robust correlation measures, such as Kendall τ rank correlation

(Li et al., 2012) and fused Kolmogorov filter (Mai and Zou, 2015), their iterative joint screening counterparts remain vulnerable. This vulnerability arises because the essential model fitting performed during iterations is compromised by noises. In addition, as the number of noisy observations increases, it is more difficult to measure the true dependence between response and features due to excessive interference, even using robust screening utility. An undesirable outcome is that the signal of relevant features is weakened but that of irrelevant ones is strengthened. With the above analysis, developing a robust joint screening procedure for high-dimensional contaminated data is therefore desirable in both theory and practice.

In this paper, we aim to develop an effective joint screening approach, which is insensitive to noisy observations. We consider a dual sample-feature L_0 fitting procedure, where an objective function with two constraints that are properly designed to realize the robust and joint screening. Our purpose is to simultaneously retain most clean observations and all relevant features, which are detected by restricted but effective fitting. Thus, the new proposed feature screening method is named as hybrid denoising-screening (HDS). To be specific, HDS estimates high-dimensional model coefficients on a designated strip-shaped sample space and screens features with zero-estimated coefficients out. Since the estimation is conducted on the retained clean observations, the leading screening procedure is robust to those noisy ones. In comparison to marginal robust screening, HDS naturally accounts for the joint effects between features by jointly estimating their model coefficients; this potentially leads to a more reliable screening result. To efficiently implement HDS, we design an iterative algorithm based on an alternating optimization strategy to approximately solve the dual sample-feature L_0 fitting procedure. Each iteration under this algorithm decreases the value of least square objective via simple operations and thereby improve the retained observations and features. Under the mild assumptions, we establish the convergence of the estimating algorithm and prove that HDS enjoys the sure screening property in the sense of Fan and Lv (2008) as if the screening were conducted on a clean sample space. The promising performance of HDS is well observed in the numerical comparisons with its competitors.

The rest of this paper is organized as follows. In Section 2, we formulate the research problem and introduce the HDS procedure. In Section 3, we investigate the theoretical properties of HDS. In Section 4, we demonstrate the promising performance of HDS via Monte Carlo simulations and a real data example. Concluding remarks are given in Section 5 and the proofs of theorems are relegated to supplementary material S4.

2. Methodology

2.1 Model and Problem Setup

Consider a high-dimensional contaminated dataset consisting n observations $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$, where the response y_i and the covariate vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ are collected independently from random variables y and $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^{\mathrm{T}}$. The dimension of covariates (features) is denoted by p, which is much larger than the sample size n. In the contaminated dataset, we assume n_1 clean observations are generated by a linear model

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon_i, i \in \mathcal{I}_1, \tag{2.1}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^{\mathrm{T}}$ is the regression coefficient vector, ε_i is the random error following $N(0, \sigma^2)$, and $\mathcal{I}_1 \subset \{1, \ldots, n\}$ is the sample index set of all clean observations. While clean observations adhere to the posited linear model, the dataset contains $n_0 = n - n_1$ contaminated observations that systematically deviate from linear assumptions. Crucially, these points cannot be well-represented by any linear approximation, and their underlying distribution is typically unknown and challenging to characterize in practice. The corresponding sample index set is denoted by $\mathcal{I}_0 = \{1, \ldots, n\} \backslash \mathcal{I}_1$. Unfortunately, $\mathcal{I}_0$ often goes unnoticed, although noisy observations $\{y_i, \boldsymbol{x}_i\}$ with $i \in \mathcal{I}_0$ may have serious effects in estimation, inference, and model selection (Weisberg, 2005).

In the high-dimensional setup, it is often believed that only a small number of features are influential on the response. This amounts to assuming that $\boldsymbol{\beta}^*$ contains many zero entries (sparse); only features with non-zero coefficients are relevant to the analysis. The index set of relevant features is defined as

$$\mathcal{M} = \{j : \beta_j^* \neq 0, j = 1, \ldots, p\}.$$

The cardinality of $\mathcal{M}$ is denoted by m in the following analysis.

Obviously, massive noises and irrelevant features tend to mask the true model and impose great challenges for data analysis. If people ignore the effect of noises and irrelevant features and conduct a direct analysis with the total sample, it may lead to a distorted result. Thus, the goal of this study is to screen out the most irrelevant features with indices outside $\mathcal{M}$ by designing a data-dependent method, in which the joint effects between features and the negative influence of noises are accounted. Features after screening form a refined model, on which a subsequent analysis can be carried out with an affordable cost and improved accuracy.

2.2 Hybrid Denoising-screening

To reduce the negative impact of noisy observations, we intend to select clean observations for screening. To this end, we introduce a set of auxiliary weight parameters $\{v_i\}_{i=1}^{n}$, where each $v_i \in \{1, 0\}$. If $v_i = 1$, the corresponding $\{y_i, \boldsymbol{x}_i\}$ is regarded to be clean and selected for joint screening. Conversely, $v_i = 0$ reflects the uncertainty of $\{y_i, \boldsymbol{x}_i\}$ on its cleanness and is excluded for joint screening. Let $\boldsymbol{v} = (v_1, \dots, v_n)^{\mathrm{T}}$ and $\|\cdot\|_0$ be the L_0 norm of a vector indicating the number of non-zero elements in that vector. By this, $\|\boldsymbol{v}\|_0$ can be used to indicate the scale of clean sample. With the auxiliary $\boldsymbol{v}$, a new weighted least squared loss is defined by

$$\mathcal{L}(\boldsymbol{v},\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} v_i(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2. \tag{2.2}$$

Since the true $\{v_i\}_{i=1}^{n}$ are unknown and need to be estimated, the direct sparse regression on $\mathcal{L}(\boldsymbol{v},\boldsymbol{\beta})$ will lead to a meaningless estimate $\hat{\boldsymbol{v}} = \boldsymbol{0}$. To avoid these extreme weights, a constraint on the scale of clean sample, $\|\boldsymbol{v}\|_0 \geq L$, is imposed on $\boldsymbol{v}$ to select the clean observations for screening, where $0 < L < n$ is a user-specified integer indicating the number of clean observations to be selected. For robust and joint screening purposes, we further consider the following estimating procedure based on dual sample-feature L_0 fitting.

$$\hat{\boldsymbol{v}}, \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{v},\boldsymbol{\beta}}{\operatorname{argmin}}\ \mathcal{L}(\boldsymbol{v},\boldsymbol{\beta}) \ \text{ subject to } v_i \in \{1,0\},\ \|\boldsymbol{v}\|_0 \geq L,\ \text{ and } \|\boldsymbol{\beta}\|_0 \leq K, \tag{2.3}$$

where $m < K < \min\{L, p\}$ is another user-specified positive integer indicating the number of features to be retained after screening. The last sparsity constraint $\|\boldsymbol{\beta}\|_0 \leq K$ guarantees that $\hat{\boldsymbol{\beta}}$ would have a sparse structure with at most K non-zero elements. Thus, the index set of retained features can be defined as

$$\hat{\mathcal{M}} = \{j : \hat{\beta}_j \neq 0, j = 1, \ldots, p\},$$

where $\hat{\beta}_j$ is the jth element of $\hat{\boldsymbol{\beta}}$. In general, those K features supported most by the joint weighted squared loss in (2.2) are retained to form a reduced model from (2.1), while the rest $p - K$ features are screened out of the model. In addition to a refined feature set retained, its corresponding clean observations are determined simultaneously for the further analysis.

Through solving (2.3), L observations will obtain positive $\hat{v}_i = 1$ when their corresponding fitted loss is small enough. When L is chosen properly such that most clean observations can be well-fitted with low residuals, the joint feature screening based on these clean observations is believed to be effective. Conversely, a better feature set with a moderate cardinality K would contribute to selecting the clean observations and reducing the negative effect of noises. Obviously, for high-dimensional contaminated data, the denoising and joint screening can be realized simultaneously by solving (2.3). Thus, we call the leading procedure hybrid denoising-screening (HDS).

Remark 1. In literature, some other methods also can be used for denoising, such as least trimmed squares (LTS) (Rousseeuw and Leroy, 1987) and mean-shift-model-based outlier detection (MSMOD) (She and Owen, 2011). After careful comparison, we find some interesting relationships between HDS, LTS, and MSMOD. Due to the limited length of paper, the related discussion is placed in Section S1 of supplementary material.

Remark 2. Although HDS may be available to one component linear model in high-dimensional mixture regressions, it is improper to divide the whole sample into clean and noisy ones by HDS, since all observations are considered informative and clean in the mixture of linear models and should not be partially overlooked. Regarding the joint screening in high-dimensional mixture regressions, a sparsity-restricted expectation-approximation-maximization algorithm proposed by Jing (2023) can be referred to for a detailed description.

2.3 Algorithm

2.3.1 Alternating Optimization Strategy (AOS)

While (2.3) is conceptually simple, finding its global optimal solution can be numerically challenging, due to the complexity of combinatorial optimization on n observations and p features simultaneously. However, since our main goal is feature screening, finding the global solution to (2.3) is not necessary. In fact, it suffices if we can obtain a good local solution that retains all relevant features. Thus, an efficient algorithm is designed to solve (2.3). Since $\boldsymbol{v}$ and $\boldsymbol{\beta}$ are two different sets of parameters, we turn to use the alternating optimization strategy (AOS) for (2.3) and iterate $\boldsymbol{v}$ and $\boldsymbol{\beta}$ alternatively until they converge. The specific iterations of $\boldsymbol{v}$ and $\boldsymbol{\beta}$ are given as follows. Given the update of $\boldsymbol{\beta}$ in the t-th iteration, denoted by $\boldsymbol{\beta}^{(t)}$, the leading update of $\boldsymbol{v}$ is computed by

$$\boldsymbol{v}^{(t+1)} = \underset{\boldsymbol{v}}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n} v_i(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)})^2 \ \text{ subject to } v_i \in \{1, 0\} \text{ and } \|\boldsymbol{v}\|_0 \geq L. \tag{2.4}$$

Although the (2.4) may have multiple solutions leading to the same minimal value, the unique update of v_i can be artificially determined as

$$v_i^{(t+1)} = \begin{cases} 1, & R(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}|) \leq L \\ 0, & R(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}|) > L \end{cases}, \quad i = 1, \ldots, n, \tag{2.5}$$

where the involved function $R(a_i) = \sum_{k=1}^n I(a_k \leq a_i) - \sum_{k \neq i}^n I(a_k = a_i, i < k)$ denotes the rank of a_i given a series $a_1, a_2, \ldots, a_n$. Given $\boldsymbol{v}^{(t+1)} = (v_1^{(t+1)}, \ldots, v_n^{(t+1)})^{\mathrm{T}}$, the idealized update of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}^{(t+1)} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^n v_i^{(t+1)} (y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2 \text{ subject to } \|\boldsymbol{\beta}\|_0 \leq K. \tag{2.6}$$

Although the above $\boldsymbol{\beta}^{(t+1)}$ is conceptually defined to be the global solution of (2.6), a good local solution also suffices to retain all relevant features in general. In the following Section 2.3.2, we introduce an efficient method named iterative hard thresholding (IHT) to solve (2.6) in detail.

Obviously, the above updates (2.5) and (2.6) can be regarded as the denoising step and screening step, respectively. Together with (2.5) and (2.6), it can be found only the observations with small residuals tend to be selected for the following update of $\boldsymbol{\beta}$. Finally, the used observations may be restricted to a strip-shaped truncated region approximately centered by $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*$; it would be resistant to the noisy observations. In

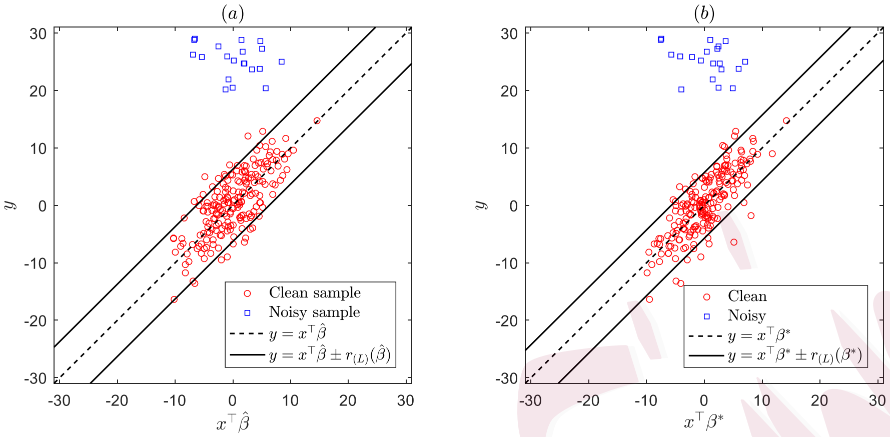

Figure 1: (a) The strip-shaped truncated area centered by $y = \boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. (b) The strip-shaped truncated area centered by $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*$, where $r_{(L)}(\cdot)$ is defined in Section S1 of supplementary material.

this section, a toy example is designed to show the above strip-shaped truncated region explicitly. Specifically, for all $\boldsymbol{x}_i$ with $i \in \mathcal{I}_1 \cup \mathcal{I}_0$, all elements x_{ij} are independently generated from $N(0, 1)$. For $i \in \mathcal{I}_1$, $y_i = 2x_{i1} + 2x_{i2} + \ldots + 2x_{i5} + \varepsilon_i$ with $\varepsilon_i \sim N(0, 4)$. For $i \in \mathcal{I}_0$, $y_i \sim U(20, 30)$, where $U(a, b)$ a uniform distribution ranging from a to b. We set $(n_1, n_0, p) = (200, 20, 1000)$ and $(L, K) = (180, 10)$. The HDS based on the AOS strategy is used to search for clean observations and remove irrelevant features. The strip-shaped truncated areas centered by $y = \boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ and $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*$ are respectively shown in Figure 1. It is seen that the estimated strip-shaped truncated area is similar to the true one since $\hat{\boldsymbol{\beta}}$ is close to $\boldsymbol{\beta}^*$. This indicates that all relevant features are effectively retained in HDS. By setting a proper L, the leading truncated area covers most clean observations with small residuals. By utilizing these clean observations, HDS mitigates the adverse effects of noises.

Remark 3. The above example illustrates an idealized scenario for using (2.3) to identify $\mathcal{M}$. When the strip-shaped truncated area centered by $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*$ contains little to no noise, the data points within this region maintain a linear relationship closely approximating the clean model. Consequently, $\boldsymbol{\beta}^*$ can be effectively recovered using L_0-based sparse least squares estimation. Some intuitive conditions contributing to the identifiability of the true model include: (1) a sufficiently large ratio n_1/n_0; (2) a sufficiently high signal-to-noise ratio $\mathrm{Var}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*)/\sigma^2$ for $i \in \mathcal{I}_1$; and (3) a high probability that the difference $|y' - y''|$ is large for any $\boldsymbol{x}$, where y' and y'' denote the clean and noisy responses, respectively. The formal assumptions for theoretical justification are presented in Section 3.

2.3.2 Iterative Hard Thresholding (IHT) Algorithm

In this section, we introduce the IHT algorithm to the update of $\boldsymbol{\beta}^{(t+1)}$ in (2.6). Since $\boldsymbol{\beta}^{(t+1)}$ is efficiently solved by IHT algorithm, we further denote $\boldsymbol{\beta}^{(t+1)}$ by $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}$. In IHT, problem (2.6) is handled by solving a series of approximated univariate problems

$$\begin{aligned}\boldsymbol{\gamma}^{(h+1)} &= \underset{\boldsymbol{\gamma}}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma}^{(h)}) + \nabla_{\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(h)}} + \frac{u}{2}\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(h)})\|_2^2 \\ &\quad \text{subject to } \|\boldsymbol{\gamma}\|_0 \le K,\ h = 0, 1, 2, \ldots, \qquad (2.7)\end{aligned}$$

where u is a positive scaling parameter. Obviously, there is an analytical solution for problem (2.7), and the whole computing process is efficient without heavy-duty operations, such as matrix inversions. The analytical solution to (2.7) can be expressed by

$$\boldsymbol{\gamma}^{(h+1)} = H(Q(\boldsymbol{\gamma}^{(h)}); K), \tag{2.8}$$

where $Q(\boldsymbol{\gamma}^{(h)}) = \boldsymbol{\gamma}^{(h)} - u^{-1}\nabla_{\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(h)}}$ and $H(\boldsymbol{a}; K)$ is a truncation operator setting the entries of $\boldsymbol{a} = (a_1, \ldots, a_p)^{\mathrm{T}}$ to zero for the entries with their rank $R(|a_j|)$ not greater than $p - K$. We can run (2.8) recursively and output $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)} = \boldsymbol{\gamma}^{(h+1)}$ when $\|\boldsymbol{\gamma}^{(h+1)} - \boldsymbol{\gamma}^{(h)}\|_2 < \epsilon_\gamma$, where ϵ_γ is a prespecified stopping threshold.

Implementation details of IHT, including the configurations of u and the initialization of $\boldsymbol{\gamma}^{(0)}$ in (2.7), are provided in Section S2 of supplementary material. This section also includes Algorithm 1, which summarizes the complete computational procedure for HDS using AOS and IHT.

2.4 The Choice of L and K

In HDS, the parameter L in (2.3) controls how many clean observations are selected for joint screening. A proper L can retain sufficient clean observations to obtain an accurate estimate of $\boldsymbol{\beta}$ for screening. However, an overly large L may lead to inaccurate screening because some noisy observations may be contained. In this paper, we suggest an extended Bayesian information criterion (EBIC) proposed by Chen and Chen (2008) based on a reduced linear model to select a proper L. The details of the reduced model can be referred to She and Owen (2011). To be specific, let $\hat{\boldsymbol{u}} = (\hat{u}_1, \ldots, \hat{u}_n)^{\mathrm{T}}$ with $\hat{u}_i = (1-\hat{v}_i)(y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})$. Furthermore, denote $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\mathbf{X}_{\hat{\mathcal{M}}} = (\boldsymbol{x}_{1,\hat{\mathcal{M}}}, \ldots, \boldsymbol{x}_{n,\hat{\mathcal{M}}})^{\mathrm{T}}$, and $\mathbf{H} = \mathbf{X}_{\hat{\mathcal{M}}}(\mathbf{X}_{\hat{\mathcal{M}}}^{\mathrm{T}}\mathbf{X}_{\hat{\mathcal{M}}})^{-1}\mathbf{X}_{\hat{\mathcal{M}}}^{\mathrm{T}}$. We consider the following EBIC,

$$\mathrm{EBIC}(L) = \mathrm{RSS} + (n - L) \cdot (\log(q) + \log(n)), \tag{2.9}$$

where $\mathrm{RSS} = \|(\mathbf{I} - \mathbf{H})(\mathbf{y} - \hat{\boldsymbol{u}})\|_2^2$ and $q = n - K$. In practice, given a series of L, we choose the optimal L with the minimal EBIC value.

The screening size K determines the number of important features to be retained in HDS. When some prior information on the total number of relevant features is available, one practical strategy is to set K to be 2-4 times larger than the anticipated number of relevant features. This contributes to increasing the chance of retaining all relevant features and reducing the interpretative difficulty and computational burden for the subsequent in-depth analysis. If there is no prior on the number of relevant features, it is suggested to select K related to n, such as $K = \lfloor n/\log(n) \rfloor$ (Fan and Lv, 2008) and $K = \lfloor n^{\alpha}\log(n) \rfloor$ with some $0 < \alpha \leq 1/2$ (Xu and Chen, 2014), where $\lfloor a \rfloor$ denotes the largest integer not greater than a. It is gratifying that we observed the performance of HDS is insensitive to a wide range of K in our numerical examples. Generally, it is not necessary to select the optimal K during the screening procedure with excessive computational cost.

3. Theoretical Properties

This section provides some theoretical justifications for the proposed screening method. In particular, we aim to answer two key questions: 1) is the updating procedure of the method convergent; and 2) how well the new method performs for feature screening with high-dimensional contaminated data.

For the convenience of presentation, we introduce a few additional notations as follows. We use s to denote an arbitrary subset of $\{1, \ldots, p\}$, which amounts to a submodel with covariates $\boldsymbol{x}_s = \{x_j, j \in s\}$ and associated coefficients $\boldsymbol{\beta}_s = \{\beta_j, j \in s\}$. Also, we use $\mathbb{N}(s)$ to indicate the size (cardinality) of set s. Clearly, the screening performance of the proposed method relies on a stable updating of model parameters in (2.2). It is thus important to first check whether the parameter updating would stabilize over the iterations. By the theorem below, we show that the procedure updating is convergent in terms of the objective value of (2.2) and parameter updates in (2.5) and (2.8).

Theorem 1. *Let $e_{max} = \max\{\lambda_{\max}(\mathbf{A}(\boldsymbol{v})), \|\boldsymbol{v}\|_0 = L\}$. If $u > e_{max}$ and $t \geq T_1$, we have*

$$\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}) \leq \mathcal{L}(\boldsymbol{v}^{(t)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}), \tag{3.10}$$

where $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}$ is a limiting point of $\boldsymbol{\gamma}^{(h)}$. If $\mathbf{A}(\boldsymbol{v}, s) = n^{-1}\sum_{i=1}^n v_i \boldsymbol{x}_{i,s}\boldsymbol{x}_{i,s}^{\mathrm{T}}$ is positive defined for any pair $(\boldsymbol{v}, s)$ satisfying $\|\boldsymbol{v}\|_0 \geq L$ and $\mathbb{N}(s) \leq K$, we have $\{\boldsymbol{v}^{(t)}, \boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}\}$ converges to a limiting point $\{\widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{\beta}}\}$. Besides, $\widetilde{\boldsymbol{\beta}}$ is a local minimum of $\mathcal{L}(\widetilde{\boldsymbol{v}}, \boldsymbol{\beta})$ subject to $\|\boldsymbol{\beta}\|_0 \leq K$.

Theorem 1 indicates that, when the scale parameter u is large enough, the proposed procedure necessarily improves the objective function. Since $L(\boldsymbol{v}, \boldsymbol{\beta})$ is lower bounded, $L(\boldsymbol{v}^{(t)}, \boldsymbol{\beta}_{\text{IHT}}^{(t)})$ will stabilize as $t \to \infty$. Besides, when $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\text{T}}$ has a proper structure, $\boldsymbol{\beta}_{\text{IHT}}^{(t)}$ leads to a local solution of (2.3) and a screening feature subset $\hat{\mathcal{M}}$.

In Theorem 1, we derive a lower bound of u to guarantee the improvement of $L(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta})$ given $\boldsymbol{v}^{(t+1)}$. This bound purely serves for the theoretical justification and may not necessarily provide a practical guidance of setting u. Our empirical experience shows that the proposed procedure usually still enjoys the increment property with a u below this bound at certain iterations. In practice, we use an adaptive tuning strategy for u, which is proposed by Zang et al. (2022) and discussed in Section S2 of supplementary material.

By Theorems 1, we have gained the insights on the convergence of the proposed procedure when it is applied to a given dataset. We now turn to evaluate the statistical performance of HDS. We first work on the AOS performance of HDS at the population level. The population counterpart of $\{v_i, \boldsymbol{x}_i, y_i\}$ is denoted as $\{v, \boldsymbol{x}, y\}$. Given a $\boldsymbol{\beta}^{(t)}$, the population update $v^{(t+1)}$ on $(y, \boldsymbol{x})$ can be expressed by

$$I(|y - \boldsymbol{x}^{\text{T}}\boldsymbol{\beta}^{(t)}| \leq \nu_\rho^{(t)}), \tag{3.11}$$

where $\nu_\rho^{(t)}$ satisfies $E[I(|y - \boldsymbol{x}^{\text{T}}\boldsymbol{\beta}^{(t)}| \leq \nu_\rho^{(t)})] = \rho$ with a given $\rho \in (0,1)$ reflecting the proportion of sample used for $\boldsymbol{\beta}$-update at the population level. Obviously, the $\nu_\rho^{(t)}$ is the

ρth population quantile of the distribution of $|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^{(t)}|$. Similarly, $v_i^{(t+1)}$ in (2.5) can be re-expressed as

$$v_i^{(t+1)} = I(|y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \leq |y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^{(t)}|_{(L)}). \tag{3.12}$$

The Lth order statistic $|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^{(t)}|_{(L)}$ can be seen as the estimate of $\nu_\rho^{(t)}$, where $L = \lfloor n\rho \rfloor + I(n\rho - \lfloor n\rho \rfloor > 0)$ is an integer close to $n\rho$. With $\nu_\rho^{(t)}$, we can define the population counterpart of $\mathcal{L}(\boldsymbol{v}^{(t+1)}, \boldsymbol{\beta})$ as $G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) = E[I(|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^{(t)}| \leq \nu_\rho^{(t)})(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})^2]$. Given $\boldsymbol{\beta}^{(t)}$, the population update of $\boldsymbol{\beta}$ is

$$\bar{\boldsymbol{\beta}}^{(t+1)} = \operatorname*{argmin}_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)}) \text{ subject to } \|\boldsymbol{\beta}\|_0 \leq K. \tag{3.13}$$

Given $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\gamma}^{(h)}$, the population IHT update can be expressed by

$$\bar{\boldsymbol{\gamma}}^{(h+1)} = H(Q_\rho(\boldsymbol{\gamma}^{(h)}); K), \tag{3.14}$$

where $Q_\rho(\boldsymbol{\gamma}^{(h)}) = \boldsymbol{\gamma}^{(h)} - u^{-1}\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)})|_{\boldsymbol{\beta}=\boldsymbol{\gamma}^{(h)}}$. Define ν_ρ^* to be the constant such that $E[I(|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*| \leq \nu_\rho^*)] = \rho$. Let $f(y|\boldsymbol{x}) = \pi_1 f_1(y|\boldsymbol{x}) + \pi_0 f_0(y|\boldsymbol{x})$ be the conditional density of y given $\boldsymbol{x}$, where $f_1(y|\boldsymbol{x}) = (\sqrt{2\pi}\sigma)^{-1}\exp\{-(y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2/(2\sigma^2)\}$ and $f_0(y|\boldsymbol{x})$ denote the conditional distributions of clean y and noisy y, respectively.

In the following theorem, we show that under the idealized condition, true parameter

$\boldsymbol{\beta}^*$ is one limiting point of population HDS updates.

Theorem 2. *Assume that* $\min_s \lambda_{\min}\left(\int \boldsymbol{x}_s\boldsymbol{x}_s^{\mathrm{T}} f(\boldsymbol{x}_s)\mathrm{d}\boldsymbol{x}_s\right) \geq c_1 > 0$ *for any model* s *with* $\mathbb{N}(s^*) \leq \mathbb{N}(s) \leq \mathbb{N}(s^*) + K$*, where* $\lambda_{\min}(\cdot)$ *denotes the smallest eigenvalue of a matrix and* $f(\boldsymbol{x}_s)$ *is the density of* $\boldsymbol{x}_s$*. For a given* $\rho \in (0, 1)$*, if* $f_0(y|\boldsymbol{x}) = 0$ *for all* $(y, \boldsymbol{x})$ *satisfying* $|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*| \leq \nu_\rho^*$*, we have*

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*) \text{ subject to } \|\boldsymbol{\beta}\|_0 \leq K.$$

Furthermore, when $\bar{\boldsymbol{\gamma}}^{(0)} = \bar{\boldsymbol{\beta}}^{(t)} = \boldsymbol{\beta}^*$*, the population IHT update satisfies*

$$\bar{\boldsymbol{\gamma}}^{(h+1)} = H(Q_\rho(\bar{\boldsymbol{\gamma}}^{(h)}); K) = \boldsymbol{\beta}^* \text{ for all } h \geq 0.$$

Theorem 2 implies that when the strip-shaped truncated region satisfying $|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*| \leq \nu_\rho^*$ does not contain any noisy observation, the population AOS updates may find a limiting point equaling $\boldsymbol{\beta}^*$. Besides, IHT is proved to be an effective tool to stabilize population $\bar{\boldsymbol{\beta}}^{(t)}$ at $\boldsymbol{\beta}^*$. Once $\boldsymbol{\beta}^*$ is obtained by using the population update, this means all relevant features also have been retained after screening.

Since $\boldsymbol{\beta}^*$ can be a limiting point at the population level with a proper ρ, we are now interested in whether $\boldsymbol{\beta}^*$ lead to a minimal dual sample-feature L_0 fitting loss in population level. If yes, the motivation of HDS is justified. Assume $\widetilde{\boldsymbol{\beta}}$ being a limiting point of population update (3.13). Its leading dual sample-feature L_0 fitting loss is $G_\rho(\widetilde{\boldsymbol{\beta}}|\widetilde{\boldsymbol{\beta}})$, in which there exists a corresponding $\widetilde{\nu}_\rho$ such that $E[I(|y-\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}| \leq \widetilde{\nu}_\rho)] = \rho$. The strip-shaped truncated area of $\widetilde{\boldsymbol{\beta}}$ is denoted by $\Lambda_\rho(\widetilde{\boldsymbol{\beta}}) = \{(y,\boldsymbol{x}) : |y-\boldsymbol{x}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}| \leq \widetilde{\nu}_\rho\}$. Let Θ_ρ be the set of all possible population limiting points and Λ_ρ be the union set of all $\Lambda_\rho(\widetilde{\boldsymbol{\beta}})$ with $\widetilde{\boldsymbol{\beta}} \in \Theta_\rho$. In the following theorem, we show that under the idealized condition, $\boldsymbol{\beta}^*$ corresponds to the lowest population objective in HDS.

Theorem 3. *For a given $\rho \in (0,1)$, assume $\boldsymbol{\beta}^* \in \Theta_\rho$ and $f_0(y|\boldsymbol{x}) = 0$ for all $(y,\boldsymbol{x}) \in \Lambda_\rho$. For any $\widetilde{\boldsymbol{\beta}} \in \Theta_\rho \setminus \{\boldsymbol{\beta}^*\}$, we have $G_\rho(\boldsymbol{\beta}^*|\boldsymbol{\beta}^*) < G_\rho(\widetilde{\boldsymbol{\beta}}|\widetilde{\boldsymbol{\beta}})$.*

Theorem 3 indicates $G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta})$ obtains the smallest objective function at true parameter $\boldsymbol{\beta}^*$; this provides a theoretical support for HDS procedure based on the minimization of dual sample-feature L_0 fitting objective (2.3). When we obtain different limiting points by using multiple initials in practice, the $\hat{\boldsymbol{\beta}}$ with the smallest $\mathcal{L}(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\beta}})$ is suggested because this $\hat{\boldsymbol{\beta}}$ is more likely to approximate $\boldsymbol{\beta}^*$ by Theorem 3.

In Theorem 2 and 3, to support the identifiability of $\boldsymbol{\beta}^*$, we assume that $f_0(y|\boldsymbol{x}) = 0$ for all $(y,\boldsymbol{x}) \in \Lambda_\rho(\boldsymbol{\beta}^*)$ or Λ_ρ. These conditions are naturally satisfied for data without any outliers (i.e. $f(y|\boldsymbol{x}) = (\sqrt{2\pi}\sigma)^{-1}\exp\{-(y-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*)^2/(2\sigma^2)\}$). This means that HDS is applicable whether the data is clear or contaminated. We conjecture that HDS would still be effective under more general $\pi_0 f_0(y|\boldsymbol{x})$. However, the corresponding theoretical justification is likely to be lengthy. Here, we aim to provide some theoretical understanding of HDS and do not intend to make these conditions weakest possible.

The above Theorem 2 and 3 show that the desired performance of HDS on $G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)$. Now, a core question is that whether the proposed algorithm can search a limiting point close to $\boldsymbol{\beta}^*$. Theorem 4 proves that using the AOS strategy, the update sequence iteratively approaches $\boldsymbol{\beta}^*$ with high probability. When $\boldsymbol{\beta}^{(t)}$ is sufficiently close to $\boldsymbol{\beta}^*$, the sure screening property of HDS can be conducted in Theorem 5. The remaining theoretical investigations are based on the following technical conditions.

C1 Let $\mathcal{B}(R;\boldsymbol{\beta}^*) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq R\}$, where R is a radius around $\boldsymbol{\beta}^*$. For an appropriately small parameter $C_1 \geq 0$, we have

$$\|\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}) - \nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)\| \leq C_1 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \text{ for any } \boldsymbol{\beta} \in \mathcal{B}(R;\boldsymbol{\beta}^*),$$

where the gradient is taken with respect to the first variable of $G_\rho(\cdot|\cdot)$.

C2 The function $G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)$ is Lipschitz-smooth with modulus C_2 and strongly-convex with modulus C_3 in $\mathcal{B}(R;\boldsymbol{\beta}^*)$.

C3 Each feature x_j is bounded, i.e. $\|\boldsymbol{x}\|_\infty \leq d$, where d is a positive constant.

C4 There exist positive constants $\omega_1, \omega_2, \omega_3$ and some constants τ_1, τ_2, τ_3 such that

$$\omega_1 n^{-\tau_1} \leq \min_{j \in \mathcal{M}} |\beta_j^*| \leq \|\boldsymbol{\beta}^*\|_2 \leq \omega_2 n^{\tau_2}, \quad \text{and} \quad m < K \leq \omega_3 n^{\tau_3},$$

where $\tau_1, \tau_3 > 0$ and $\tau_2 > -\tau_1$. In addition, $\log p = O(n^a)$ for some $0 \leq a < 1$.

Condition C1 requires that the difference between $\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta})$ and $\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)$ can be restricted by a sufficiently small parameter C_1 when $\boldsymbol{\beta}$ is in the neighborhood of $\boldsymbol{\beta}^*$. Condition C2 indicates that, the function $G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)$ is sandwiched between two quadratic functions when $\boldsymbol{\beta}^*$ is fixed. The constants C_2 and C_3 in Condition C2 can be chosen by the largest and smallest eigenvalues of matrix $E[I(|y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}^*| \leq \nu_\rho^*)\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}]$. Conditions C1 and C2 play core roles in establishing the desired geometric convergence of $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}$ in population level. In detail, Condition C2 ensures that the gradient-based updates on $G_\rho(\cdot|\boldsymbol{\beta}^*)$ can converge geometrically to $\boldsymbol{\beta}^*$ (Nesterov, 2013). It is worth noting that our analysis is mainly conducted on $G_\rho(\cdot|\boldsymbol{\beta})$, not only on $G_\rho(\cdot|\boldsymbol{\beta}^*)$. Thus, to establish the population convergence, we attempt to quantify the difference between $\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta})$ and $\nabla_{\boldsymbol{\beta}} G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)$. If the above difference can be controlled in a small extent, the performance of $G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta})$ tend to mimic that of $G_\rho(\boldsymbol{\beta}|\boldsymbol{\beta}^*)$. Hence, the geometric convergence of HDS in population level can be established. When the statistical error between the sample and population updates is further controlled, the convergence region of HDS can be obtained. This ingenious analytical strategy was first proposed by Balakrishnan et al. (2017) for EM algorithm. In order to make the proof more concise, Condition C3 restricts the support of $\boldsymbol{x}$ into a compact region. This condition can be extended to more general light-tailed distributions. To establish the sure screening property of HDS, Condition C4 states a few requirements. The first one is the minimal non-zero β_j^* does not degenerate too fast, so that the signal is detectable in the asymptotic sequence. Meanwhile, Condition C4 confines an appropriate order of K that guarantees the identifiability of $\mathcal{M}$. In addition, we assume that p diverges with n up to an exponential rate, which implies that p can be substantially larger than n.

Based on Conditions C1-C3, we show that the proposed update $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t)}$ converges geometrically to $\boldsymbol{\beta}^*$ with an overwhelming probability in the following theorem.

Theorem 4. *For a given $\rho \in (0,1)$, assume $\boldsymbol{\beta}^* \in \Theta_\rho$. Then, we assume Conditions C1-C3 hold with $1-2\cdot(C_3-C_1)/(C_1+C_2) \in (0,1)$ and $\boldsymbol{\beta}^{(0)} \in \mathcal{B}(R;\boldsymbol{\beta}^*)$ with $\|\boldsymbol{\beta}^{(0)}\|_0 \le K$ and $R = \kappa \cdot \|\boldsymbol{\beta}^*\|_2$ for some $\kappa \in (0,1)$. Furthermore, the stepsize is set to $1/u = 2/(C_2+C_3)$. Meanwhile, we assume that*

$$K = \lceil C \cdot \max\left\{ \frac{16}{\{1/[1-2\cdot(C_3-C_1)/(C_2+C_3)]-1\}^2}, \frac{4\cdot(1+\kappa)^2}{(1-\kappa)^2} \right\} \cdot m \rceil, \tag{3.15}$$

$$\left(\sqrt{K} + C'\sqrt{m/(1-\kappa)}\right)\cdot \epsilon \le \min\left\{ \left(1-\sqrt{1-2\cdot\frac{C_3-C_1}{C_2+C_3}}\right)^2 \cdot R, \frac{(1-\kappa)^2}{2\cdot(1+\kappa)} \cdot \|\boldsymbol{\beta}^*\|_2 \right\}, \tag{3.16}$$

where $C \ge 1$ and $C' > 0$ are positive constants and $\lceil a \rceil$ denotes the smallest integer value being greater than or equal to a. For each $\epsilon > 0$, when $n > (8d\nu_\rho^{(t)})/u\epsilon$, we have

$$\|\boldsymbol{\beta}_{\mathrm{IHT}}^{(t)} - \boldsymbol{\beta}^*\|_2 \le \left(1-2\cdot\frac{C_3-C_1}{C_3+C_2}\right)^{t/2} \cdot R + \frac{\left(\sqrt{K}+C'\sqrt{m/(1-\kappa)}\right)\cdot\epsilon}{1-\sqrt{1-2(C_3-C_1)/(C_3+C_2)}} \tag{3.17}$$

holds with probability at least $1-t\cdot\delta$, where $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)} = H(Q(\boldsymbol{\beta}_{\mathrm{IHT}}^{(t)});K)$ and $\log\delta = O(-cn\epsilon^2+$

$\log p$) *with some positive constant* c.

It is worth noting that $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}$ in (3.17) involves only one step of truncation in IHT. For the general $\boldsymbol{\beta}_{\mathrm{IHT}}^{(t+1)}$ with sufficient iterations, the proof tend to be similar by using a variant of Condition C1, like (5.5) in Balakrishnan et al. (2017) or (3.3) in Wang et al. (2014). In (3.15), K is chosen to be sufficiently large and the same order as m; this is used to bound the error incurred by the truncation step in IHT. The error term ϵ in (3.16) and (3.17) is used to bound $\|Q(\boldsymbol{\gamma}^{(0)}) - Q_\rho(\boldsymbol{\gamma}^{(0)})\|_2$ and thus dependent on sample size n. When n is sufficient large, that is ϵ can be set sufficiently small, (3.16) would be easily satisfied and the second term in (3.17) can be controlled under a low level. Theorem 4 illustrate that, the upper bound of the proposed estimation error in (3.17) can be decomposed into two terms. The first term is the optimization error bound in population level. The optimization error bound decreases to zero at a geometric rate of convergence. Meanwhile, the second term is statistical error bound reflecting the random vibration of $\|Q(\boldsymbol{\gamma}^{(0)}) - Q_\rho(\boldsymbol{\gamma}^{(0)})\|_2$. Theorem 4 shows that with some mild conditions, the above two error terms can be well controlled with an overwhelming probability; this means the proposed AOS strategy with IHT performs effectively in estimation of $\boldsymbol{\beta}$.

Next, we justify the sure screening property of HDS using the following theorem with the support of Theorem 4.

Theorem 5. *Denote* $\mathcal{M}^{(t)} = \{j : \beta_j^{(t)} \neq 0, j = 1, \ldots, p\}$ *the screened features in the* t*-th iteration. Let* $\varrho = 1-2(C_3-C_1)/(C_3+C_2) \in (0,1)$ *and* $t_0 = \lceil 2\log_\varrho(\omega_1/(2\kappa\omega_2)^{-1}\cdot n^{-\tau_1-\tau_2})\rceil$.

When Condition C4 is further assumed with $(a+\tau_3)/2+\tau_1 < 1/2$, following the notations and conditions in Theorem 4, there exists a constant $\zeta \in (\tau_3/2+\tau_1, (1-a)/2)$ such that

$$\lim_{n\to\infty} P\left(\mathcal{M} \subset \mathcal{M}^{(t)}\right) \to 1$$

for each $t = t_0, t_0+1, \ldots t_0+T$, where T is a finite positive integer.

Theorem 5 indicates that the proposed HDS enjoys sure screening property in the sense of Fan and Lv (2008), even when the number of relevant features m, the number of retained features K and the dimension of features p are all diverging with n. That is, when n is large, HDS 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 Theorem 5 requires that m, K and p can not diverge too fast with n. In addition, the signal of minimal nonzero β_j^* can not be too small to be detected. The above requirements is reflected by $(a+\tau_3)/2+\tau_1 < 1/2$. Theorem 5 also shows that within finite iteration steps, HDS can effectively retain all relevant features; this means the proposed algorithm can be stopped in advance when difference between successive estimators is small.

4. Numerical Studies

In our numerical studies, we consider four simulation examples and a real data analysis, where Example 2 further conducts a detailed sensitivity analysis and Example 3 evaluates the effectiveness and robustness of HDS with some more challenging setups. Due to the limited length of the paper, we place Example 2 and 3 in the supplementary material S3.

4.1 Simulations

Example 1. Influence of Noises on Classic Joint Screening

For the simulation study, we first generate the observations of features $\{\boldsymbol{x}_i\}_{i=1}^n$. For $i \in \mathcal{I}_1$, $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where covariance matrix $\boldsymbol{\Sigma} = [\rho_{j,k}]_{j,k=1,\ldots,p}$ with $\rho_{j,j} = 1$, $\rho_{j,k} = 0.5$ for $j \neq k$. For $\boldsymbol{x}_i$ with $i \in \mathcal{I}_0$, we consider two different settings, which are respectively introduced in the following (1a) and (1b) setups. Given $\boldsymbol{x}_i$, the corresponding y_i is generated by the distribution $f(y_i|\boldsymbol{x}_i) = \pi_1 f_1(y_i|\boldsymbol{x}_i) + \pi_0 f_0(y_i|\boldsymbol{x}_i)$. We set $f_1(y_i|\boldsymbol{x}_i) = (\sqrt{2\pi})^{-1}\exp\{-(y_i - \sum_{j\in\mathcal{M}} x_{ij}\beta_j)^2/2\}$. The remaining settings, including non-zero regression coefficients, different distributions of noise $f_0(y_i|\boldsymbol{x}_i)$, an (n, p, K), are specifically given in the following setups (1a)-(1c). The index set of relevant features is given by $\mathcal{M} = \{1, \ldots, 5\}$.

(1a) **Noisy y_i.** The noisy $\boldsymbol{x}_i$ follow the same distribution of clean ones, and $f_0(y_i|\boldsymbol{x}_i) = 0.5 f_{U(-40,-20)}(y_i) + 0.5 f_{U(20,40)}(y_i)$, where $f_{U(a,b)}$ is the density function of distribution $U(a, b)$. The nonzero values of coefficients are generated by $\beta_j^* = (-1)^W(4\text{log}n_1/\sqrt{n_1}+$

$|V|$) and $\beta^*_j = 0$ otherwise, where $W \sim$ Bernoulli(0.4) and $V \sim N(0,1)$. We set $(n,p,K) = (150, 2000, 20)$.

(1b) **Noisy $\boldsymbol{x}_i$.** The noisy covariate vector is generated by $\boldsymbol{x}_i = \boldsymbol{a}_i + \boldsymbol{b}_i$, where the signal $\boldsymbol{a}_i$ follows the same distribution of clean $\boldsymbol{x}_i$ and the disturbance $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{ip})^{\mathrm{T}}$ with each $b_{ij} \sim U(5,10)$. Furthermore, $f_0(y_i|\boldsymbol{x}_i) = (\sqrt{2\pi})^{-1}\exp\{-[y_i - \sum_{j\in\mathcal{M}}(x_{ij} - b_{ij})\beta^*_j]^2/2\}$ with $\boldsymbol{\beta}^*_{\mathcal{M}} = (1.5, 1.5, 1.5, 1.5, -1.5)^{\mathrm{T}}$. We set $(n,p,K) = (180, 3000, 10)$.

(1c) **Noisy $(y_i, \boldsymbol{x}_i)$.** The noisy $\boldsymbol{x}_i$ follows (1b), and the noisy y_i follows (1a). We set $\boldsymbol{\beta}^*_{\mathcal{M}} = (-3, 3, 2.5, -2, 2)^{\mathrm{T}}$ and $(n,p,K) = (200, 5000, 15)$.

In the above setups, (1a) and (1b) respectively consider the effects of contaminated response y_i and features $\boldsymbol{x}_i$. In (1c), both response and features are likely to contain noises.

In this example, we test different noisy-sample-to-clean-sample ratios (NCR $= \pi_0/\pi_1$), which are 0%, 2.5%, 5%, 10%, 20%, 30%, 50%, respectively. For comparison purposes, we also conduct joint feature screening using ISIS, FRS, SMLE, and HOLP. The evaluation criterion for the involved methods is a sucessful screening rate (SSR) based on $T = 100$ repetitions. Specifically, let $\hat{\mathcal{M}}(t)$ denote the index set of the retained features at the t-th repetition and $\mathbb{N}(\cdot)$ denote the cardinality of a set. The SSR is computed by

$$\mathrm{SSR} = \frac{1}{T}\sum_{t=1}^{T} I(\mathcal{M} \subset \hat{\mathcal{M}}(t)),$$

where $\mathbb{N}(\hat{\mathcal{M}}(t)) = K$. In addition to the classic joint screening methods, we also show the

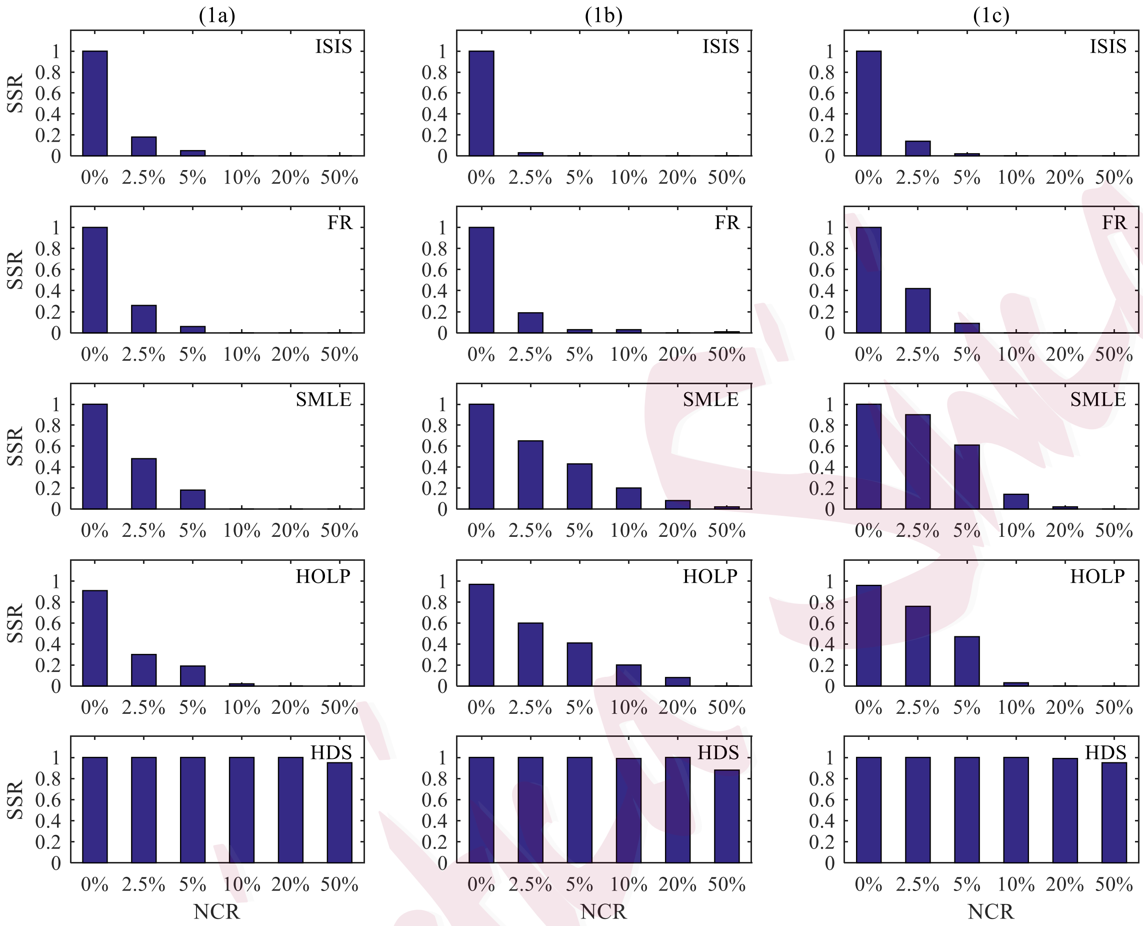

Figure 2: The SSRs of all joint screeners in setups (1a)-(1c), where the NCR is ranging from 0% to 50%.

SSR of HDS with $L = \lfloor 0.85n_1 \rfloor$ for comparison. The simulation results of all methods are summarized in Figure 2. Moreover, to demonstrate the impact of L on HDS, we consider 5 different Ls for each NCR. Specifically, $L = \lfloor \alpha\pi_1 n \rfloor$, where $\alpha = 0.95, 0.9, 0.85, 0.8, 0.75$ is a scale parameter. Table 1. shows the SSR corresponding to each pair of (NCR,α).

For all four tested screening procedures, it is easily found that all methods work well when NCR = 0% in setup (1a)-(1c), even though there are strong correlations between

Table 1: The SSRs of HDS with different NCR and α on setup (1a).

Setup	NCR=0%	NCR=2.5%	NCR=5%	NCR=10%	NCR=20%	NCR=30%	NCR=50%
$\alpha = 0.95$	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$\alpha = 0.90$	1.00	1.00	1.00	1.00	1.00	1.00	0.99
$\alpha = 0.85$	1.00	1.00	1.00	1.00	1.00	0.98	0.95
$\alpha = 0.80$	1.00	0.99	1.00	1.00	0.98	0.89	0.86
$\alpha = 0.75$	1.00	0.97	0.99	0.96	0.94	0.76	0.61

features. As NCR increases, SSRs of classic methods become more and more inaccurate. This is due to they are sensitive to the extreme values of noisy observations. This finding is a major motivation for our HDS. In comparison, although HDS encounters a slight drop of SSR when NCR $\geq$ 20%, HDS still has pronounced advantages in terms of joint screening accuracy. Whether the fully clean data (NCR $=$ 0%) or three different types of contaminated data in (1a)-(1c) are considered, HDS shows great robustness and accuracy. Since HDS aims to retaining clean observations in a strip-shaped truncated area, almost all noisy observations are ruled out of analysis with a proper L. As a result, the extreme values of the response or features do not cause great challenges on HDS in this example.

Regarding the effect of L reflected by the α in Table 1, we find that a larger number of clean observations tends to be beneficial to improve the screening accuracy. In practice, we suggest choosing a satisfactory L when the computational cost is affordable. The related simulations can be found in Example 4.

Example 4. EBIC-based Selection of L

In the results of Example 1, we find that a larger number of clean observations tend to be beneficial to improve the screening accuracy. In this section, we test the proposed EBIC-based selection procedure for L (2.9) in the following experiments, where the contaminated data are generated by setups (1a)-(1c) in Example 1. To ensure fairness for each setup, we set $K = 2\mathbb{N}(\mathcal{M})$, and $(n, p) = (200, 2000)$. The candidates of L are given by $n_u, n_u - 5, n_u - 10, \ldots, 5\lfloor 0.16\pi_1 n\rfloor$ with $n_u = 5\min\{n, \lfloor 0.22\pi_1 n\rfloor\}$. We test HDS with EBIC under four different NCRs, which are 0%, 10%, 30%, and 50%, respectively. In addition to the evaluation criteria on screening performance (PSR and FDR can be found in Example 3 of supplementary material S2), five new indices are designed to measure the denoising performance based on $T = 100$ repetitions. They are selection rate (SR_o), positive selection rate (PSR_o) and false discovery rate (FDR_o) on observations. To evaluate the estimating accuracy (EA) of HDS, we consider two indices, which are denoted by EA_1 and EA_2 for short. To be specific, these indices are computed by

$$\mathrm{SR}_o = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{N}(\hat{\mathcal{I}}_1(t))}{n_1(t)},\quad \mathrm{PSR}_o = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{N}(\mathcal{I}_1(t)\cap\hat{\mathcal{I}}_1(t))}{\mathbb{N}(\mathcal{I}_1(t))},\quad \mathrm{FDR}_o = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{N}(\hat{\mathcal{I}}_1(t)-\mathcal{I}_1(t))}{\mathbb{N}(\hat{\mathcal{I}}_1(t))},$$

$$\mathrm{EA}_1 = \frac{1}{T}\sum_{t=1}^{T}\frac{\|\hat{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}^*(t)\|_2^2}{\|\boldsymbol{\beta}^*(t)\|_2^2},\quad \mathrm{EA}_2 = \frac{1}{T}\sum_{t=1}^{T}\frac{\|(\hat{\boldsymbol{\beta}}(t)-\boldsymbol{\beta}^*(t))_{\mathcal{M}}\|_2^2}{\|(\boldsymbol{\beta}^*(t))_{\mathcal{M}}\|_2^2},$$

where $n_1(t)$, $\mathcal{I}_1(t)$, and $\hat{\mathcal{I}}_1(t)$ are n_1, $\mathcal{I}_1$ and $\hat{\mathcal{I}}_1 = \{i : \hat{v}_i = 1, i = 1, \ldots, n\}$ in the t-th repetition. The results are summarized in Table 2.

It is seen that the adaptive EBIC can select an effective L and retains the overwhelming majority of clean observations and a minority of noisy ones; this is indicated by the

Table 2: The simulation result of setups in Example 1.

Setup	NCR	SSR	PSR	FDR	SR_o	PSR_o	FDR_o	EA_1	EA_2
(1a)	0%	1.00	1.00	0.50	1.00	1.00	0.000	0.019	0.003
	10%	1.00	1.00	0.50	0.99	0.99	0.000	0.019	0.003
	30%	1.00	1.00	0.50	0.99	0.99	0.000	0.023	0.005
	50%	1.00	1.00	0.50	0.99	0.99	0.001	0.027	0.005
(1b)	0%	1.00	1.00	0.50	1.00	1.00	0.000	0.043	0.007
	10%	1.00	1.00	0.50	0.99	0.99	0.000	0.048	0.008
	30%	0.99	1.00	0.50	0.99	0.99	0.002	0.068	0.013
	50%	1.00	1.00	0.50	0.99	0.99	0.000	0.070	0.016
(1c)	0%	1.00	1.00	0.50	1.00	1.00	0.000	0.015	0.002
	10%	1.00	1.00	0.50	1.00	0.99	0.011	0.018	0.003
	30%	1.00	1.00	0.50	1.01	0.97	0.036	0.026	0.005
	50%	1.00	1.00	0.50	1.01	0.95	0.060	0.039	0.009

results that $\mathrm{SR}_o \approx 1$, $\mathrm{PSR}_o \approx 1$, and $\mathrm{FDR}_o \approx 0$. Although very few noisy observations have not been removed due to their small distances to the true linear model, they have not create great challenges on HDS; this is reflected by the ideal SSRs, PSRs, and FDRs. Based on these high-quality observations, HDS finally obtains very sharp performances in terms of screening and estimating accuracy. The later is supported by the obtained small EA_1 and EA_2.

Figure 3 shows the EBIC values of the given Ls in the first simulation. It can be found that the optimal selection of L is very close to n_1, which is the true number of clean observations. To be more specific, when the data set is uncontaminated, all observations are retained by EBIC-based HDS; this is supported by the first subplot with NCR=0%. In addition, we observe that when $L > n_1$, the corresponding EBIC value exhibits a sharp increase. This prompts the EBIC-based HDS to select the maximum candidate

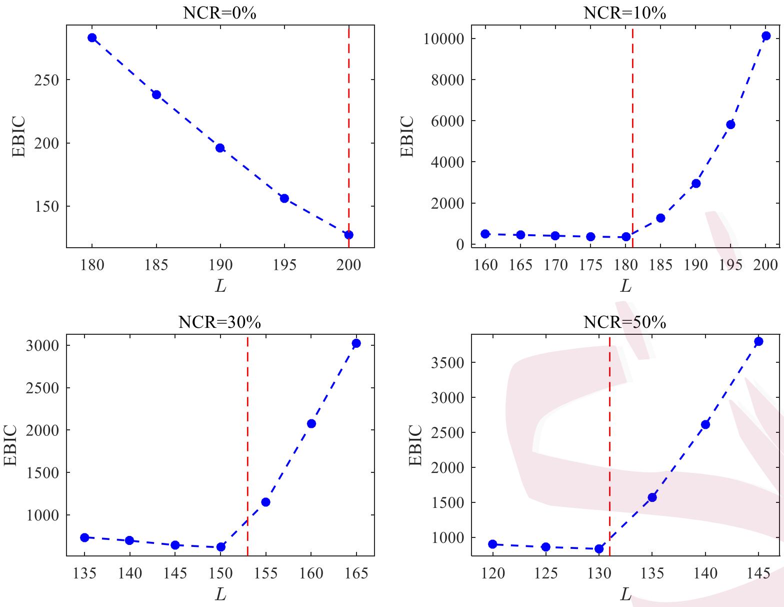

Figure 3: The EBIC values corresponding to the given Ls in the first simulation. The dash line denotes the true number of clean observations (n_1).

value for L that does not exceed n_1 ($L \leq n_1$). Consequently, the proposed EBIC-based HDS maximizes the retention of clean observations.

4.2 Real Data Analysis

We apply the proposed HDS to a real dataset, which contains 124 community indices (features) extracted from 2215 communities in the USA along with the associated number of murders per 100,000 population (response). The goal of this paper is to explore the relationship between murder and community indices and select a few community indices that may mainly contribute to the murder rate. More details of this data are available at `http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized`.

We first remove 22 features, each containing at least 1872 missing values. Then, the data entries containing the missing value are deleted, and $n = 2214$ observations are eventually left for analysis based on a linear model. It is observed that the condition number of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is more than 2.69×10^{16}, there exists strong correlations and collinearity among features. It indicates that the dataset likely contains many redundant features. As a result, we consider conducting feature screening to obtain a reduced model with a moderate model size and better interpretability.

Next, we investigate whether the dataset contains noisy observations. For this purpose, HDS is implemented with $K = 8$ and $L = n - N_0$ with N_0, where N_0 is selected from $\mathcal{N}_0 = \{0, 50, 100, 150, 200, 250\}$. Let $\hat{\boldsymbol{\beta}}_d$ and $\hat{\mathcal{M}}_d$ denote the estimates of $\boldsymbol{\beta}$ and $\mathcal{M}$ based on the d-th candidate value in $\mathcal{N}_0$. The resulting mean squared error (MSE) and model similarity ratio (SR) between successive estimates are defined as

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_d\right)^2 \quad \text{and} \quad \mathrm{SR}_d = \frac{\mathbb{N}(\hat{\mathcal{M}}_d \cap \hat{\mathcal{M}}_{d+1})}{K}.$$

These metrics are summarized in Figure 4. It is seen that when 50-100 plausible noisy observations are excluded from the joint screening procedure, the MSE encounters a drastic decline. In addition, as N_0 becomes larger, the estimated models tend to be more similar. By these findings, it is reasonable to suspect that the dataset contains some “noisy” observations that are not appropriate to be trained together with other observations by using

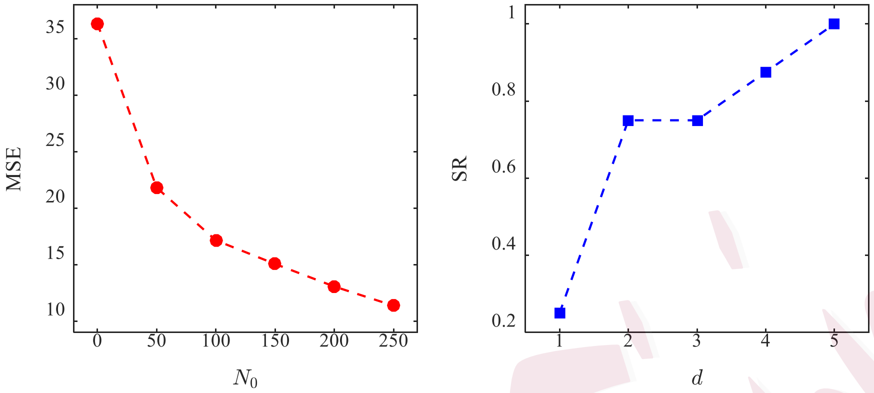

Figure 4: MSE and SR of the selected screeners with $N_0 = 0, 50, 100, 150, 200, 250$.

a unified linear model. When "noisy" observations are ruled out, the estimated model becomes more stable. By investigation, most of these "noisy" observations correspond to the communities with very high murder rates. These communities may have different crime models from other communities.

The performance of HDS is further evaluated by its prediction in this example. To this end, the dataset is divided into a training set and a testing set. We consider a contaminated training set containing the "top N_0 noisiest" observations, which correspond to the top N_0 largest fitting residuals in HDS with $K = 8$ and $L = n - N_0$. The N_0 is selected from $\{25, 50, 75, 100\}$. After N_0 "noises" are selected, we randomly choose $1000 - N_0$ observations in the remaining data; this leads to a training set of size $n_{train} = 1000$. The remaining 1214 observations are treated as a testing set. The proposed HDS with $L = n - N_0$ is considered on the training set to retain a set of $K = 8$ key features, based on which a linear model is trained. The predicting MSE is assessed on the testing set.

Table 3: Prediction MSE for different screening methods.

Method	$N_0 = 25$	$N_0 = 50$	$N_0 = 75$	$N_0 = 100$
HDS	29.52 (1.66)	25.34 (1.46)	22.99 (1.06)	20.73 (1.16)
RRCS	32.26 (2.03)	29.27 (1.70)	26.98 (1.80)	25.46 (1.68)
ISIS	32.44 (2.13)	28.64 (1.83)	26.91 (2.31)	25.27 (2.44)
FRS	33.93 (4.43)	28.52 (1.54)	26.87 (1.89)	24.84 (1.59)
SMLE	37.87 (7.48)	32.99 (2.41)	30.20 (2.56)	28.72 (2.22)
HOLP	37.05 (6.27)	33.24 (2.44)	30.96 (2.78)	29.35 (3.08)

For comparison, we repeat the analysis with other screening methods RRCS, ISIS, FRS, SMLE, and HOLP. All results are given in Table 3, which shows the median of prediction MSE with its robust standard deviation in the parentheses based on 100 repetitions.

It is clear that HDS does a good job in this example, as the associated linear model enjoys decent prediction accuracy and stability. As the number of “noises” increases in the training set, more clean data entries (easier to predict) are left in the testing set. This leads to an MSE drop for each method, but HDS still has the significant superiority among all methods. The out-performance of HDS is due to that the estimated model obtained by HDS is trained based on the most relevant features and clean observations and therefore more suitable to the testing set.

5. Concluding Remarks

In this paper, we propose a novel hybrid denoising-screening (HDS) method for analyzing high-dimensional contaminated data. The HDS framework originates from solving a dual L_0-regularized optimization problem addressing both sample and feature spaces, which simultaneously eliminates noisy observations and irrelevant features. Through removing noisy samples, the accuracy of our joint screening method can be effectively guaranteed. Furthermore, we designed an effective algorithm based on alternating optimization strategy, which leads the computation procedure to be efficiently implemented. The promising performance of the method is supported by both theory and extensive numerical examples.

In HDS, to avoid the bad local optimal solution without sure screening property, we suggest using lasso initial for IHT update in the first T_1 iterations. In fact, some other strategies also can be employed. The first strategy is to execute HDS with multiple initializations, then selecting the optimal solution from the resulting set of locally optimal solutions. This procedure is easily implemented but often leads to excessive computational costs. Another strategy is to test whether the retained observations are from a linear model, where the residual approximately follows a symmetrically truncated normal distribution. Since this strategy has a vast space for research, we leave this interesting topic for future research.

Moreover, our current work focuses on the contaminated linear regression. It would be promising to explore the possibility of using a HDS in more general regression or classification models.

Supplementary Material

The supplementary material contains the comparison of HDS with LTS and MSMOD, the implementation details of IHT and the final algorithm for HDS, Example 2 and 3 in simulations, and the proofs of all theoretical results in the main text.

Acknowledgement

Xu's research was supported by National Key R&D Program of China (2022Y-FA1003804) and Major Key Project of Peng Cheng Laboratory (PCL2024AS103). Li's research supported by the Natural Science Foundation of China (12401394), Postdoctoral Fellowship of CPSF (GZB20240611), China Postdoctoral Science Foundation (2024M752549), Shaanxi Province Postdoctoral Research Project Funding (2024BSHSDZZ145), and Fundamental Research Funds for the Central Universities (xzy012024035). Wang's research was supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (24KJD110002), Postgraduate Research and Practice Innovation Program of Jiangsu Province (No. KYCX23_1288) and Qing Lan Project of Jiangsu Province (2022).

References

Balakrishnan, S., M. J. Wainwright, and B. Yu (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics 45*(1), 77–120.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Fan, J. and Y. Fan (2008). High dimensional classification using features annealed independence rules. *Annals of statistics 36*(6), 2605–2637.

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

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

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

Fan, J. and R. Song (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics 38*(6), 3567–3604.

Gorst-Rasmussen, A. and T. Scheike (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society Series B: Statistical Methodology 75*(2), 217–245.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (2011). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

Jing, K. (2023). *Joint feature screening and subsampling in analysis of massive data*. Ph. D. thesis, Université d'Ottawa/University of Ottawa.

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

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

Mai, Q. and H. Zou (2015). The fused kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics 43*(4), 1471–1497.

Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, Volume 87. Springer Science & Business Media.

Rousseeuw, P. J. and A. M. Leroy (1987). *Robust regression and outlier detection*, Volume 589. John wiley & sons.

She, Y. and A. B. Owen (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association 106*(494), 626–639.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association 104*(488), 1512–1524.

Wang, X. and C. Leng (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(3), 589–611.

Wang, Z., Q. Gu, Y. Ning, and H. Liu (2014). High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*.

Weisberg, S. (2005). *Applied linear regression*, Volume 528. John Wiley & Sons.

Xu, C. and J. Chen (2014). The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association 109*(507), 1257–1269.

Zang, Q., C. Xu, and K. Burkett (2022). Smle: An r package for joint feature screening in ultrahigh-dimensional glms. *arXiv preprint arXiv:2201.03512*.

Zhou, T., L. Zhu, C. Xu, and R. Li (2020). Model-free forward screening via cumulative divergence. *Journal of the American Statistical Association 115*(531), 1393–1405.

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

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing, China

E-mail: (wanglimingnuist@163.com)

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing, China

E-mail: (laipeng@nuist.edu.cn)

Department of Mathematics and Fundamental Research, Peng Cheng Laboratory, Shenzhen, China

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China

E-mail: (cx3@xjtu.edu.cn)

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China

E-mail: (lxxwlm2013@xjtu.edu.cn.)