Statistica Sinica **33** (2023), 1169-1191
doi:https://doi.org/10.5705/ss.202020.0421

FEATURE SCREENING VIA DISTANCE CORRELATION FOR ULTRAHIGH DIMENSIONAL DATA WITH RESPONSES MISSING AT RANDOM

Linli Xia[1,2] and Niansheng Tang[1]

[1] *Yunnan University and* [2] *Tongren University*

Abstract: This study examines the feature screening problem for ultrahigh-dimensional data with responses missing at random. A two-step procedure is proposed to screen important features. The first step screens the significant covariates associated with the missing indicators via the fused mean-variance filter. The second step screens the important predictors associated with the response by fusing the distance correlation and a nonparametric imputation technique. The proposed feature screening procedure has the following merits: (i) it is model free, because it does not depend on a special model structure or distribution assumption; (ii) it avoids resampling on the conditional function of the missing value because a kernel smoothing technique is adopted to implement the nonparametric conditional mean imputation; (iii) it is not sensitive to a misspecification of the propensity score function because it does not impose a special model on the respondent probability. Under some regularity conditions, the sure screening property is shown. A modified maximum ratio criterion is proposed to select the tuning parameter. Simulation studies are conducted to investigate the finite-sample performance of the proposed feature screening procedure. Finally, an example is used to illustrate the proposed methodologies.

Key words and phrases: Distance correlation, missing at random, nonparametric imputation, sure screening property, ultrahigh dimensional data.

1. Introduction

Ultrahigh-dimensional data are often encountered in fields of modern scientific research such as signal processing, biomedical imaging and functional magnetic resonance imaging, and finance. Here, the number of candidate predictors p may increase at an exponential rate of the sample size n, while only a small number of predictors contribute to the response when there is sparsity among the candidate predictors. Under the "larger p smaller n" data framework, various penalized variable selection procedures have been developed to reduce the dimensionality to a number below the sample size by effectively distinguishing

Corresponding author: Niansheng Tang, Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kun-ming 650500, China. E-mail: nstang@ynu.edu.cn.

important predictors. For example, see the lasso (Tibshirani (1996)), smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)), adaptive lasso (Zou (2006)), and minimax concave penalties (Zhang (2010)). The aforementioned penalized variable selection methods may not perform well for ultrahigh-dimensional data, owing to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability (Fan, Samworth and Wu (2009)).

To accommodate these challenges, various feature screening procedures have been developed for ultrahigh-dimensional data. For instance, Fan and Lv (2008) proposed a sure independence screening (SIS) procedure and an iterated sure independence screening (ISIS) procedure to select active predictors by ranking the marginal correlations for a linear model with Gaussian covariates and responses. Fan and Song (2010)presented a more general version of independent learning by ranking the maximum marginal likelihood estimates or the maximum marginal likelihood itself in generalized linear models. Fan, Feng and Song (2011) developed a nonparametric independence screening (NIS) method for sparse ultrahigh-dimensional additive models. Chang, Tang and Wu (2013) presented a marginal empirical likelihood-based independence feature screening procedure for linear regression models. The aforementioned methods depend on the considered models. To this end, several model-free feature screening approaches have been proposed in recent years, including the sure independent ranking and screening (Zhu et al. (2011)), robust ranking correlation-based screening (Li et al. (2012)), distance correlation-based screening (DC-SIS) (Li, Zhong and Zhu (2012); Zhong et al. (2016)), mean-variance sure independence screening (MV-SIS) (Cui, Li and Zhong (2015)), fused Kolmogorov filter screening (Mai and Zou (2015)), conditional quantile sure independence screening (Wu and Yin (2015)), conditional sure independence screening based on the Cramer-von Mises statistic (Wang et al. (2017)), and fused mean-variance filter (Yan et al. (2018)). The aforementioned feature screening procedures focus mainly on complete-data problems. However, in many fields, such as medical, social, and economic studies, some subjects have missing responses or predictors, perhaps owing to an unwillingness of some sampled subjects to answer sensitive questions, a loss of information caused by uncontrollable factors, or subjects that schedule intermittent visits or drop out of the study (Little and Rubin (2002)).

To address these issues, there is a growing body of literature on feature screening for ultrahigh-dimensional data with missing data. For example, Lai et al. (2017) studied a model-free feature screening procedure for ultrahigh-dimensional data with responses missing at random (MAR), based on the inverse probability weighted (IPW) method. Here, the Kolmogorov filter method

is adopted to screen the active predictors under an unknown propensity score function assumption. However, their method depends heavily on the specification of the unknown propensity score function, which is rather difficult to specify correctly or estimate efficiently in the presence of ultrahigh-dimensional predictors. To address this issue, Wang and Li (2018) proposed the missing indicator imputation screening approach and a Venn diagram-based screening procedure for ultrahigh-dimensional data with responses MAR. However, their method may not work well in some cases because missing values are imputed using the missing indicator value. Recently, Tang, Xia and Yan (2019) developed a feature screening method based on the profile marginal kernel-assisted estimating equations imputation technique in ultrahigh-dimensional partially linear models with responses MAR, which is a model-based method. The MV-SIS method was developed for the case in which the response variable is fully observed categorical data, rather than missing indicators, and the predictors are continuous. The DC-SIS method was proposed for a completely observed response and predictors, and cannot be used directly to screen features in the presence of missing responses and ultrahigh-dimensional predictors, owing to the "curse of dimensionality" issue in estimating the distribution function of a missing response. To the best of our knowledge, few studies apply the MV-SIS and DC-SIS methods to the missing data problem when simultaneously screening important features associated with missing indicators and responses.

To solve the aforementioned problems, we propose a novel two-step feature screening procedure by incorporating the ideas of the MV-SIS and DC-SIS methods for ultrahigh-dimensional data with responses MAR. The first step is to screen the significant covariates associated with the missing indicators using the fused mean-variance filter (Cui, Li and Zhong (2015) to measure the dependence between the missing indicators and the covariates by regarding the former as responses. Based on these selected significant covariates, the second step screens the important predictors associated with the response by developing a modified distance correlation between the marginal distributions of the predictor and the response variable with a missing value, and nonparametrically imputing the conditional distribution function of the response. This step differs significantly from the DC-SIS procedure because it requires that we nonparametrically impute the missing responses, which may suffer from the well-known "curse of dimensionality" issue. To address this issue, a modified maximum ratio criterion for selecting the tuning parameter is proposed. The proposed tuning parameter selection procedure can choose a smaller model size than the procedures do in Huang, Li and Wang (2014) and Ni and Fang (2016). The proposed feature screening procedure

has the following merits. First, the proposed procedure is model free, because it does not require specifying a regression model of the response on the predictors. Second, it works well without resampling on the conditional distribution of the missing value in that the kernel smoothing technique is applied to implement the nonparametric conditional mean imputation. Third, it is robust to a heavy-tailed distribution of the response variable or to outliers of the response, because only the marginal distribution function of the response is used to construct the utility for screening the significant predictors. Fourth, it is robust to a misspecification of the propensity score function in that a special model for respondent probability is not required. Fifth, the modified tuning parameter selection procedure effectively addresses the "curse of dimensionality" issue. Sixth, it possesses the sure screening properties under some regularity conditions.

The rest of this paper is organized as follows. Section 2 introduces a new feature screening approach by fusing the nonparametric conditional mean imputation and the MV-SIS and DC-SIS procedures. A modified tuning parameter selection procedure is also presented in this section. Section 3 investigates the theoretical properties for the proposed feature screening approach under some regularity conditions. Simulation studies are conducted to investigate the finite-sample performance of the proposed methodologies in Section 4. An example is used to illustrated the proposed methodologies in Section 5. A brief discussion is given in Section 6. All technical details are provided in the Supplementary Material.

2. Feature Screening Approach

2.1. Distance correlation in the presence of responses MAR

Let Y be a response variable with support Ω_y, and $X = (X_1, \ldots, X_p)^\top$ be a p-dimensional predictor vector. Let $F(y|X) = \Pr(Y \leq y|X)$ be the conditional distribution function of Y given X. Without specifying a regression model, we define the index sets of the active and inactive predictors as

$$\begin{aligned}\mathcal{M} &= \{k : F(y|X) \text{ functionally depends on } X_k, \text{ for some } y \in \Omega_y\},\\ \mathcal{I} &= \{k : F(y|X) \text{ does not functionally depend on } X_k, \text{ for any } y \in \Omega_y\},\end{aligned}$$

respectively. The above definition indicates that X_k is an active predictor if $k \in \mathcal{M}$, whereas X_k is an inactive predictor if $k \in \mathcal{I}$. Denote $X_{\mathcal{M}} = \{X_k : k \in \mathcal{M}\}$ and $X_{\mathcal{I}} = \{X_k : k \in \mathcal{I}\}$ as the sets of important and unimportant predictors, respectively. Here, it is assumed that the dimensionality satisfies $p = o\{\exp(n^\alpha)\}$, for some constant $\alpha > 0$, but the cardinality of $\mathcal{M}$, denoted

as $|\mathcal{M}|$, satisfies $|\mathcal{M}| = o(n)$. In this framework, our main purpose is to screen important predictors $X_{\mathcal{M}}$ using some appropriate method.

The DC-SIS procedure (Li, Zhong and Zhu (2012)) is equivalent to the marginal Pearson correlation learning for a normal linear regression with normally distributed predictors, it is more effective than the marginal Pearson correlation learning in the presence of a nonlinear relationship between two random vectors, it is model free and possesses the sure screening property, and its implementation does not involve any numerical optimization algorithm. As such, we consider screening features associated with the response using the DC-SIS procedure. Following Li, Zhong and Zhu (2012) and Zhong et al. (2016), the marginal distance correlation between the marginal distribution functions $F_k(X_k)$ of X_k and $F(Y)$ of Y can be defined as

$$\omega_k = \mathrm{dcorr}^2\{F_k(X_k), F(Y)\} = \frac{\mathrm{dcov}^2\{F_k(X_k), F(Y)\}}{\mathrm{dcov}\{F_k(X_k), F_k(X_k)\}\mathrm{dcov}\{F(Y), F(Y)\}}, \quad (2.1)$$

for $k = 1, \ldots, p$, where $F_k(x) = E\{I(X_k \le x)\}$, for $k = 1, \ldots, p$, and $F(y) = E\{I(Y \le y)\}$, in which $I(\cdot)$ denotes an indicator function. Furthermore, $\mathrm{dcov}(u, v)$ represents the marginal distance covariance between two random variables u and v, and is defined as $\mathrm{dcov}^2(u, v) = S_1 + S_2 - 2S_3$, where $S_1 = E\{||u - \tilde{u}|| \cdot ||v - \tilde{v}||\}$, $S_2 = E\{||u - \tilde{u}||\}E\{||v - \tilde{v}||\}$, and $S_3 = E\{E(||u - \tilde{u}||\ |u)E(||v - \tilde{v}||\ |v)\}$, in which $(\tilde{u}, \tilde{v})$ is an independent copy of (u, v). In general, $F_k(X_k)$ and $F(Y)$ are unknown and can be empirically estimated by $\hat{F}_k(X_k) = (1/n)\sum_{i=1}^n I(X_{ik} \le X_k)$ and $\hat{F}_n(Y) = n^{-1}\sum_{i=1}^n I(Y_i \le Y)$, respectively, for the collected sample $\{(X_i, Y_i) : i = 1, \ldots, n\}$, where $X_i = (X_{i1}, \ldots, X_{ip})^\top$. In this case, ω_k can be estimated by the following sample distance correlation between $F_k(X_k)$ and $F(Y)$:

$$\hat{\omega}_k = \widehat{\mathrm{dcorr}}^2\{F_k(X_k), F(Y)\} = \frac{\widehat{\mathrm{dcov}}^2\{F_k(X_k), F(Y)\}}{\widehat{\mathrm{dcov}}(F_k(X_k), F_k(X_k))\widehat{\mathrm{dcov}}\{F(Y), F(Y)\}}, \quad (2.2)$$

for $k = 1, \ldots, p$, where $\widehat{\mathrm{dcov}}^2\{F_k(X_k), F(Y)\} = \hat{S}_{k1} + \hat{S}_{k2} - 2\hat{S}_{k3}$, in which $\hat{S}_{k1} = n^{-2}\sum_{i=1}^n\sum_{j=1}^n |\hat{F}_k(X_{ik}) - \hat{F}_k(X_{jk})| \cdot |\hat{F}_n(Y_i) - \hat{F}_n(Y_j)|$, $\hat{S}_{k2} = n^{-4}\{\sum_{i=1}^n\sum_{j=1}^n |\hat{F}_k(X_{ik}) - \hat{F}_k(X_{jk})|\} \cdot \{\sum_{i=1}^n\sum_{j=1}^n |\hat{F}_n(Y_i) - \hat{F}_n(Y_j)|\}$, and $\hat{S}_{k3} = n^{-3}\sum_{i=1}^n\sum_{j=1}^n \sum_{l=1}^n |\hat{F}_k(X_{ik}) - \hat{F}_k(X_{lk})| \cdot |\hat{F}_n(Y_j) - \hat{F}_n(Y_l)|$. Li, Zhong and Zhu (2012) suggested selecting a set of important predictors with large $\hat{\omega}_k$. Thus, the important predictors can be indexed as

$$\hat{\mathcal{M}} = \{k : \hat{\omega}_k \ge cn^{-\upsilon} \quad \text{for } 1 \le k \le p\}, \quad (2.3)$$

for some prespecified thresholds $c > 0$ and $0 \leq \upsilon < 1/2$.

When Y_i is subject to missingness, the above DC-SIS procedure cannot be used directly to select the important predictors. To this end, a new feature screening approach is developed to simultaneously select the important predictors associated with Y and the important covariates associated with the missingness data mechanism, as follows.

Let $\delta_i = 1$ if Y_i is observed, and $\delta_i = 0$ if Y_i is missing. Throughout this paper, it is assumed that δ_i and δ_j are independent, for any $i \neq j$, and δ_i depends only on X_i, such that $\Pr(\delta_i = 1|X_i, Y_i) = \Pr(\delta_i = 1|X_i)$, for $i = 1, \ldots, n$, which indicates that the missingness data mechanism is MAR (Little and Rubin (2002)). In the ultrahigh-dimensional setting, we further assume that only a small number of covariates in X_i contribute to the missing indicator δ_i. Again, without specifying a parametric model for $\Pr(\delta = 1|X)$, we define the index sets of the important and unimportant covariates associated with $\Pr(\delta = 1|X)$ as

$$\begin{aligned}\mathcal{M}_\delta &= \{k : \Pr(\delta = 1|X) \text{ functionally depends on } X_k\},\\ \mathcal{I}_\delta &= \{k : \Pr(\delta = 1|X) \text{ does not functionally depend on } X_k\},\end{aligned}$$

respectively, which shows that X_k has an important effect on the missing Y when $k \in \mathcal{M}_\delta$, whereas X_k is not associated with the missing Y when $k \in \mathcal{I}_\delta$. Similarly, $X_{\mathcal{M}_\delta} = \{X_k : k \in \mathcal{M}_\delta\}$ and $X_{\mathcal{I}_\delta} = \{X_k : k \in \mathcal{I}_\delta\}$ represent the sets of important covariates associated with the indicator δ and the unimportant covariates that are not associated with the indicator δ, respectively. Thus, the propensity score function can be written as

$$\pi(Z) = \Pr(\delta = 1|X, Y) = \Pr(\delta = 1|X) = \Pr(\delta = 1|Z), \tag{2.4}$$

where $Z = (Z_1, \ldots, Z_{s_n^\delta})^\top, Z_j = X_k$ for some $k \in \mathcal{M}_\delta$, and s_n^δ is the cardinality of $\mathcal{M}_\delta$.

Note that when the response variable Y is subject to MAR, the sample distance correlation defined in Equation (2.2) is not available, because $\hat{F}_n(Y)$ cannot be evaluated for the missing Y. To address the issue, in what follows, a nonparametric mean imputation approach is developed to estimate $F(Y)$ under the assumption given in Equation (2.4).

Following Cheng and Chu (1996), if the missingness data mechanism model (2.4) is correctly specified, a nonparametric estimator of $F(y)$ is given as

$$\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i I(Y_i \leq y) + (1 - \delta_i)\hat{F}_{1n}(y|Z_i)\right\}, \tag{2.5}$$

where $Z_i = (Z_{i1}, \ldots, Z_{is_n^\delta})^\top$, in which $Z_{ij} = X_{ik}$ for some $k \in \mathcal{M}_\delta$, and $\hat{F}_{1n}(y|Z_i)$ is a nonparametric regression estimator of $F_{1n}(y|Z_i) = \Pr(Y \le y|Z_i)$, which is defined as

$$\hat{F}_{1n}(y|Z_i) = \frac{\sum_{j=1}^n \delta_j I(Y_j \le y) K((Z_{j1} - Z_{i1})/h_1, \ldots, (Z_{js_n^\delta} - Z_{is_n^\delta})/h_{s_n^\delta})}{\sum_{j=1}^n \delta_j K((Z_{j1} - Z_{i1})/h_1, \ldots, (Z_{js_n^\delta} - Z_{is_n^\delta})/h_{s_n^\delta}},$$

where $K(\cdot)$ is a s_n^δ-dimensionality kernel function, $h_j = C_j h$ is the bandwidth, C_j is a fixed positive constant, for $j = 1, \ldots, s_n^\delta$, and $h = h_n^0 \to 0$ as $n \to \infty$. It follows from Theorem 2.1 of Cheng and Chu (1996) that the estimator $\hat{F}_n(y)$ defined in Eq. (2.5) is a consistent estimator of $F(y)$. When s_n^δ is large, the preceding nonparametric estimator of $F_{1n}(y|Z_i)$ may perform poorly. In this case, we can simply use a product of s_n^δ univariate kernel functions with independent smoothing parameters to replace $K(\cdot)$. In particular, when s_n^δ is diverging, we can adopt an adaptive kernel estimation to replace $\hat{F}_{1n}(y|Z_i)$ (e.g., see Bouř, Kůs and Franc (2017)). It is impossible to evaluate $\hat{F}_n(y)$ using Eq. (2.5) when the index set $\mathcal{M}_\delta$ involved is unknown. Hence, it is necessary to identify the index set $\mathcal{M}_\delta$ using an appropriate screening approach before evaluating $\hat{F}_n(y)$ via Eq. (2.5). Because δ is a Bernoulli variable, identifying the index set $\mathcal{M}_\delta$ is equivalent to a binary discriminant analysis problem. In what follows, a fused mean-variance filter (Cui, Li and Zhong (2015)) is adopted to distinguish $\mathcal{M}_\delta$.

Denote $P_0 = \Pr(\delta = 0)$, $P_1 = \Pr(\delta = 1)$, and $F_k(x) = \Pr(X_k \le x)$. Let $F_{k0}(x) = \Pr(X_k \le x|\delta = 0)$ and $F_{k1}(x) = \Pr(X_k \le x|\delta = 1)$ be the conditional cumulative distribution function of X_k given $\delta = 0$ and $\delta = 1$, respectively. Following Cui, Li and Zhong (2015), we define

$$\omega_k^\delta = MV(X_k|\delta) = P_0 \int \{F_{k0}(x) - F_k(x)\}^2 dF_k(x) + P_1 \int \{F_{k1}(x) - F_k(x)\}^2 dF_k(x),$$

which is zero if X_k is independent of δ. The sample version of ω_k^δ has the form

$$\hat{\omega}_k^\delta = \frac{1}{n}\sum_{j=1}^n \hat{P}_0 \left\{\hat{F}_{k0}(X_{jk}) - \hat{F}_k(X_{jk})\right\}^2 + \frac{1}{n}\sum_{j=1}^n \hat{P}_1 \left\{\hat{F}_{k1}(X_{jk}) - \hat{F}_k(X_{jk})\right\}^2,$$

where $\hat{P}_0 = n^{-1}\sum_{i=1}^n I(\delta_i = 0)$, $\hat{P}_1 = n^{-1}\sum_{i=1}^n I(\delta_i = 1)$, $\hat{F}_k(x) = n^{-1}\sum_{i=1}^n I(X_{ik} \le x)$, $\hat{F}_{k0}(x) = n^{-1}\sum_{i=1}^n I(X_{ik} \le x, \delta_i = 0)/\hat{P}_0$, and $\hat{F}_{k1}(x) = n^{-1}\sum_{i=1}^n I(X_{ik} \le x, \delta_i = 1)/\hat{P}_1$. Thus, $\hat{\omega}_k^\delta$ can be used to select the important covariates indexed by

$$\hat{\mathcal{M}}_\delta = \{k : \hat{\omega}_k^\delta \ge c_\delta n^{-\tau_\delta} \text{ for } 1 \le k \le p\},$$

where c_δ and τ_δ are the predetermined thresholds defined in Condition (C2). However, in many practical applications, it is quite difficult to give the thresholds c_δ and τ_δ. To this end, we consider the following reduced subset:

$$\hat{\mathcal{M}}_\delta^* = \{k : \hat{\omega}_k^\delta \text{ is among the top } d_\delta \text{ largest of all}\}, \tag{2.6}$$

where $d_\delta < n$ is a prespecified positive integer. The above feature screening procedure is called the MV-SIS procedure, according to which, the number of ultrahigh-dimensional covariates is reduced to d_δ. Under conditions (C1) and (C2), it follows from Theorem 2.1 of Cui, Li and Zhong (2015) that $\Pr(\mathcal{M}_\delta \subseteq \hat{\mathcal{M}}_\delta) \to 1$, which shows that using the MV-SIS procedure to select the important covariates associated with δ has the desirable sure screening property.

Based on the selected subset $\hat{\mathcal{M}}_\delta^*$, Eq. (2.5) can be rewritten as

$$\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i I(Y_i \le y) + (1-\delta_i)\hat{F}_{1n}(y|\hat{Z}_i)\right\}, \tag{2.7}$$

where $\hat{Z}_i = (Z_{i1}, \ldots, Z_{i\hat{s}_n^\delta})^\top$, $Z_{ij} = X_{ik}$, for some $k \in \hat{\mathcal{M}}_\delta^*$, and $\hat{s}_n^\delta$ is the cardinality of $\hat{\mathcal{M}}_\delta^*$.

With the estimated $\hat{F}_n(y)$ and $\hat{\omega}_k$ defined in Eq. (2.2), we can obtain the selected subset $\hat{\mathcal{M}}$ using Eq. (2.3). Similarly, it is quite difficult to specify the thresholds c and υ. Again, we use the following criterion to select the significant predictors:

$$\hat{\mathcal{M}}^* = \{k : \hat{\omega}_k \text{ is among the top } d_n \text{ largest of all}\}, \tag{2.8}$$

where $d_n < n$ is a prespecified threshold. The preceding feature screening procedure is referred to as the nonparametric mean imputation-based DC-SIS procedure (denoted as 'DCNI-SIS').

Remark 1. When the sample values of the mean-variance utility corresponding to important covariates (i.e., large $\hat{\omega}_k^\delta$) are always ranked beyond those corresponding to unimportant covariates with a relatively high probability, the following modified tuning parameter selection algorithm can select a relatively small model size that includes all important covariates. In this case, it is not necessary to employ the penalized likelihood methods to further reduce the number of covariates. Thus, the modified tuning parameter selection algorithm can effectively address the "curse of dimensionality" issue. However, when $\hat{\omega}_k^\delta$ corresponding to important covariates are not always ranked beyond those for inactive covariates, the cardinal number of subset $\widehat{\mathcal{M}}_\delta^*$ may still be quite large, which implies that there is still a "curse of dimensionality" problem. To this end, the penalized like-

lihood approaches (e.g., the SCAD and adaptive Lasso methods) may be adopted to further select important covariates. Thus, the estimated subset $\widehat{\mathcal{M}}_\delta^*$ may be reduced further to the subset $\widetilde{\mathcal{M}}_\delta^*$ satisfying $\Pr(\mathcal{M}_\delta \subseteq \widetilde{\mathcal{M}}_\delta^*) \to 1$ as $n \to \infty$.

2.2. The selection of the tuning parameter d_δ

In many practical applications, we need to determine an optimal tuning parameter d_δ in Equation (2.6). For simplicity, d_δ is abbreviated as d, and d_0 represents the true size of the considered model. Fan and Lv (2008) suggested taking $d = \lfloor n/\log n \rfloor$, where the notation $\lfloor a \rfloor$ denotes the integer less than or equal to a. In this case, for our considered nonparametric mean imputation, there is still the "curse of dimensionality" problem. To address the issue, Huang, Li and Wang (2014) proposed an approach for selecting the tuning parameter based on the maximum ratio criterion, but their method may lead to a much larger d than d_0. To solve this problem, Ni and Fang (2016) modified their approach to select the tuning parameter. However, this approach may be unstable.

To address the aforementioned issue, we develop a modified approach of Ni and Fang (2016). Let $\{t_1, \ldots, t_p\}$ be a permutation of the set $\{1, \ldots, p\}$ such that $\hat{\omega}_{t_1} \geq \cdots \geq \hat{\omega}_{t_p}$, where $\hat{\omega}_{t_k}$ may represent $\hat{\omega}_k$ or $\hat{\omega}_k^\delta$, for $k = 1, \ldots, p$. In general, an optimal value of d should be selected based on the following two conditions: (i) $(\hat{\omega}_{t_k} + \hat{\omega}_{t_{k+1}})/(\hat{\omega}_{t_{k+1}} + \hat{\omega}_{t_{k+2}})$ should be $O_p(1)$ for $k \neq d_0$, and (ii) $(\hat{\omega}_{t_k} + \hat{\omega}_{t_{k+1}})/(\hat{\omega}_{t_{k+1}} + \hat{\omega}_{t_{k+2}}) \xrightarrow{P} \infty$ for $k = d_0$, where $\xrightarrow{P}$ represents convergence in probability. Thus, similarly to Ni and Fang (2016), we take

$$d = \underset{1 \leq k \leq p-2}{\operatorname{argmax}} \frac{\hat{\omega}_{t_k} + \hat{\omega}_{t_{k+1}}}{\hat{\omega}_{t_{k+1}} + \hat{\omega}_{t_{k+2}}}. \tag{2.9}$$

In this case, d can be selected such that $d_{\min} \leq d \leq d_{\max}$, where $d_{\min} \geq 1$ and $d_{\max} < p$ are two user-specified positive integers. The algorithm for implementing the aforementioned approach is as follows.

Step 1: Calculate $d^{(1)} = \operatorname{argmax}_{1 \leq k \leq d_{\max}} \{(\hat{\omega}_{t_k} + \hat{\omega}_{t_{k+1}})/(\hat{\omega}_{t_{k+1}} + \hat{\omega}_{t_{k+2}})\}$.

Step 2: For $m = 1, 2, \ldots$, if $d^{(m)} < d_{\min}$, calculate $d^{(m+1)} = \operatorname{argmax}_{d^{(m)}+1 \leq k \leq d_{\max}}$ $\{(\hat{\omega}_{t_k} + \hat{\omega}_{t_{k+1}})/(\hat{\omega}_{t_{k+1}} + \hat{\omega}_{t_{k+2}})\}$.

Step 3: Repeat step 2 until $d^{(m)} \geq d_{\min}$, and choose $d = d^{(m)}$.

Intuitively, the above procedure is expected to work well, because the modified algorithm is more robust than that given in Ni and Fang (2016). In practical applications, similarly to Ni and Fang (2016), we may take $d_{\max}$ as n or $O(n/\log n)$, which is a commonly used value in the feature screening literature.

To be conservative, we may take $d_{\min} = 2, 5$, or even larger. However, the above procedure may select some inactive predictors. To solve this problem, one can again use the regularization method to select the active predictors.

3. Theoretical Properties

Now, we discuss the theoretical properties of the proposed DCNI-SIS procedure. To this end, we first introduce some conditions.

(C1) There are two positive constants c_1 and c_2 such that $c_1 \leq \min\{P_0, P_1\} \leq \max\{P_0, P_1\} \leq c_2$, where $P_t = \Pr(\delta = t)$ for $t = 0, 1$.

(C2) There are two positive constants $c_\delta > 0$ and $0 \leq \tau_\delta < 1/2$ such that $\min_{k \in \mathcal{M}_\delta} \omega_k^\delta \geq 2c_\delta n^{-\tau_\delta}$.

(C3) The probability density function $f(Z)$ is bounded away from ∞ on the support $\mathbb{Z}$ of Z, and has bounded partial derivatives (within the compact support of Z) up to order two.

(C4) The propensity score function $\pi(Z) = \Pr(\delta = 1|Z)$ satisfies $\min_i \pi(Z_i) \geq D_0 > 0$ a.s. for some positive constant D_0, and has bounded partial derivatives (within the compact support of Z) up to order two.

(C5) Function $K(\cdot)$ is an s_n^δ-dimensionality kernel function of order κ satisfying $\int K(u_1, \ldots, u_{s_n^\delta}) du_1 \cdots du_{s_n^\delta} = 1$, $\int u_j^l K(u_1, \ldots, u_{s_n^\delta}) du_1 \cdots du_{s_n^\delta} = 0$ for any $1 \leq l < \kappa$, and $\int u_j^\kappa K(u_1, \ldots, u_{s_n^\delta}) du_1 \cdots du_{s_n^\delta} \neq 0$ for any $j = 1, \ldots, s_n^\delta$, where s_n^δ is the dimension of Z. In particular, consider $K(\cdot) = \prod_{i=1}^{s_n^\delta} K_i(\cdot)$. For $\forall i \in \{1, \ldots, s_n^\delta\}$, the kernel function $K_i(\cdot)$ is the probability density function such that (i) it is bounded and has compact support; (ii) it is symmetric with $\sigma^2 = \int \omega^2 K_i(\omega) d\omega < \infty$; and (iii) $K_i(\omega) \geq D_1$ for some constant $D_1 > 0$ in some closed interval centered at zero.

(C6) For $j = 1, \ldots, s_n^\delta$, $h_j = C_j h$ is the bandwidth in which C_j is some fixed positive constant, and $h = O\{n^{-\theta/(s_n^\delta + 2)}\}$ for some $0 < \theta \leq 1$.

(C7) There are two positive constants, $c > 0$ and $0 < \upsilon < 1/3$, such that $\min_{k \in \mathcal{M}} \omega_k \geq 2cn^{-\upsilon}$.

Condition (C1) ensures that the proportions of $\delta = 0$ and $\delta = 1$ are neither too small nor too large, and is used in Shao, Yu and Zhou (2016). Condition (C2) requires that the minimum true signal cannot be too small and its order is $n^{-\tau_\delta}$, which allows the minimum true signal to degenerate to zero as $n \to \infty$, and

is used in Cui, Li and Zhong (2015). Conditions (C1) and (C2) ensure that all the important covariates associated with the missing indicator δ can be selected by the DCNI-SIS procedure with probability tending to one. Conditions (C3)–(C6) are widely adopted in the nonparametric and missing data literature (Tang, Zhao and Zhu (2014)). Condition (C7) is employed to ensure that the values of the marginal DCNI-SIS utility corresponding to the active predictors cannot be too small. Condition (C7) is used in the ultrahigh-dimensional data analysis literature (Li, Zhong and Zhu (2012)).

Based on the aforementioned conditions, we obtain the following theorem.

Theorem 1. *Suppose that Conditions* (C1)–(C7) *and* $\mathcal{M}_\delta \subseteq \widehat{\mathcal{M}}_\delta$ *hold. For any* $\upsilon/2 < \alpha < 1/2 - \upsilon$ *and* $0 \le \upsilon - \alpha < \theta \le 1$, *and* $c_2 \le c_3 n^{-\alpha}$ *with* $c_3 > 0$, *there exist positive constants* m_1, m_2, *and* m_3 *such that*

$$\Pr\left(\max_{1\le k\le p} |\hat{\omega}_k - \omega_k| \ge cn^{-\upsilon}\right) \le O\left(p\mathbb{J}_n\right), \tag{3.1}$$

where $\mathbb{J}_n = \exp\left\{-m_1 n^{1-2(\upsilon+\alpha)}\right\} + n^{2-\alpha}\exp\left\{-m_2 n^{1-2(\upsilon-\alpha)}\right\} + n\exp(-m_3 n^{\alpha})$. *In particular, we have*

$$\Pr(\mathcal{M} \subseteq \widehat{\mathcal{M}}) \ge 1 - O\Big(s_n \mathbb{J}_n\Big), \tag{3.2}$$

which indicates that $\Pr(\mathcal{M} \subseteq \widehat{\mathcal{M}}) \to 1$ *as* $n \to \infty$, *where* s_n *is the cardinality of* $\mathcal{M}$.

Theorem 1 shows that the DCNI-SIS procedure has the desirable sure screening property. Thus, we have extended the DC-SIS procedure to ultrahigh-dimensional data with responses MAR.

4. Simulation Studies

In this section, simulation studies are conducted to investigate the finite-sample performance of the proposed feature screening procedure. The following criteria are used to evaluate the performance of the procedure: AMS (i.e., the average number of model sizes); CF (i.e., the proportion that all the true active predictors are exactly selected among 500 replications); OF (i.e., the proportion that all the true active covariates are exactly selected, and at least one inactive covariate is selected among 500 replications); UF (i.e., the proportion that the true active covariates are not completely selected among 500 replications); CP_j (i.e., the proportion that the true active predictor X_j is selected into the submodel with size $\lfloor n/\log(n)\rfloor$ among 500 replications); CP_a (i.e., the proportion that all

the true active predictors are selected into the submodel with size $\lfloor n/\log(n) \rfloor$ among 500 replications); and MMS (i.e., the minimum model size needed to include all the true active predictors). In general, an SIS procedure is regarded as better than other SIS procedures if the MMS value of the former is closer to the true model size than are those of the latter, the CF, CP_j, and CP_a values of the former are larger than those of the latter, and the AMS, OF, and UF values of the former are smaller than those of the latter. We consider the following three experiments with $(p, n) = (1000, 200)$, and set the screening size as $d_n = \lfloor n/\log(n) \rfloor$.

For comparison, we evaluate the results for the SIS (Fan and Lv (2008)), SIRS (Zhu et al. (2011)), DC-SIS (Li, Zhong and Zhu (2012)), DF-SIS (Wu and Yin (2015)), MC-SIS (Wang et al. (2017)), BMI procedures (e.g., MI-I, MI-S, and V-D methods) applying the DC-SIS method to select the predictors associated with Y (Wang and Li (2018)), and the complete-case (CC) analysis method with the fixed screening size d_n under several different missingness data mechanism models.

Experiment 1. This experiment is designed to investigate the finite-sample performance of the proposed DCNI-SIS method under the assumption that $E(Y|X)$ is linear in some components of $X = (X_1, \ldots, X_p)^\top$. Here, we generate the response Y from the following linear regression model: $Y = X^\top \beta + \varepsilon$, where $\beta = (\beta_1, \ldots, \beta_p)^\top = (2.5, 2.0, 4.0, 2.0, 0.0, \ldots, 0.0)^\top$, which indicates that $X_1, \ldots, X_4$ are active predictors and $X_5, \ldots, X_p$ are inactive predictors, and $X = (X_1, \ldots, X_p)^\top$ is generated from a multivariate normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$, in which $\sigma_{ij} = 0.5^{|i-j|}$. To demonstrate that the proposed DCNI-SIS procedure does not depend on the distribution of the predictors, X_1 and X_4 are replaced by those generated from the normal distribution $N(0, 2)$. This is robust to outliers of the response. We add 5% of outliers generated from the distribution $U(70, 90)$ for the response. In addition, to illustrate that the proposed DCNI-SIS procedure is independent of the distribution of ε, we consider the following three settings: (E1) $\varepsilon \sim N(0, 1)$; (E2) $\varepsilon \sim t(3)$, where $t(3)$ represents the t-distribution with three degrees of freedom; and (E3) $\varepsilon \sim \mathrm{Cauchy}(0, 1)$, where $\mathrm{Cauchy}(0, 1)$ denotes the Cauchy distribution with location and scatter parameters being zero and one, respectively. It is assumed that X is completely observed, while Y is subject to missingness.

To create missing data for Y, the missing indicator δ is generated from a Bernoulli distribution with probability $\pi = \Pr(\delta = 1|Z)$, specified by

(i) $\mathrm{logit}(\pi) = \alpha_0 + \alpha_1 X_3$, where $(\alpha_0, \alpha_1) = (-1.0, 4.0)$, $\mathrm{logit}(a) = \log\{a/(1-a)\}$;

(ii) $\pi = 0.15$ if $X_2 + X_3 < 0$, and 0.85 otherwise;

(iii) $\text{logit}(\pi) = \alpha_0 + \alpha_1 X_3 + \alpha_2 Y$, where $(\alpha_0, \alpha_1, \alpha_2) = (-1.1, 4.0, 0.01)$.

Scenario (i) describes a MAR mechanism specified by a logistic regression model, and indicates $Z = X_3$, that is, the number of important covariates associated with the missing indicator δ is one. Scenario (ii) represents a MAR mechanism nonparametrically specified, and implies $Z = (Z_1, Z_2)^\top = (X_2, X_3)^\top$, that is, the number of important covariates associated with the missing indicator δ is two. Scenario (iii) denotes a nonignorable missing data mechanism and prescribes a selection bias case, which is used to investigate the robustness of the proposed DCNI-SIS procedure to the misspecification of π. This scenario implies $Z = X_3$, that is, the number of important covariates associated with the missing indicator δ is one. The average missing proportions among 500 replications corresponding to the three missingness data mechanisms (i), (ii) and (iii) are about (56.3%, 55.8%, 56.2%), (50.4%, 49.9%, 49.9%), and (56.6%, 56.2%, 56.7%) for error distributions (E1), (E2), and (E3), respectively.

For the 500 data sets generated from each of the above nine settings (i.e., three error distributions × three missingness data mechanisms), the MV-SIS and DCNI-SIS procedures are used to screen the important covariates associated with δ and the active predictors in the considered linear regression model, respectively. To estimate the distribution function of Y in the presence of missing Y, we take the kernel function $K(\cdot) = \prod_{j=1}^{s_n^\delta} K_j(\cdot)$ with $K_j(z_j) = 0.75(1 - z_j^2)_+$, and set the bandwidth as $h_j = C_j \hat{\sigma}_{z_j} n^{-1/5}$, where $C_j = 1$, and $\hat{\sigma}_{z_j}$ is the standard deviation of the observations $\{Z_{ij} : i = 1, \ldots, n; j = 1, \ldots, s_n^\delta\}$.

The results for screening the important covariates in the propensity score function are given in Table 1, where 'MV-TM' and 'MV-T' represent the MV-SIS approaches with $d_{\min} = 1$ for scenarios (i) and (iii) or $d_{\min} = 2$ for scenario (ii), respectively, and d_δ is determined by the modified tuning parameter selection method introduced in Section 2.2 and the algorithm introduced in Ni and Fang (2016). Table 1 shows that the MV-TM method behaves better than the MV-T method, because the former has larger CF values and smaller OF values than the latter, and the AMS values for the former are closer to the true model size than are those of the latter, regardless of the error distributions and missingness data mechanisms.

The results for identifying active predictors are presented in Table 2, where 'DCNI-TM' and 'DCNI-T' represent the DCNI-SIS procedures with $d_{\min} = 4$ and d_n determined by the modified tuning parameter selection method introduced in Section 2.2 and the algorithm introduced in Ni and Fang (2016), respectively.

Table 1. Performance of MV-SIS procedure for screening covariates in propensity score function in Experiment 1.

		$\varepsilon \sim N(0,1)$				$\varepsilon \sim t(3)$				$\varepsilon \sim \text{Cauchy}(0,1)$			
Case	Method	AMS	CF	OF	UF	AMS	CF	OF	UF	AMS	CF	OF	UF
(i)	MV-TM	1.002	0.998	0.002	0.000	1.008	0.992	0.008	0.000	1.004	0.996	0.004	0.000
	MV-T	1.048	0.952	0.048	0.000	1.060	0.940	0.060	0.000	1.034	0.966	0.034	0.000
(ii)	MV-TM	2.000	1.000	0.000	0.000	2.000	1.000	0.000	0.000	2.000	1.000	0.000	0.000
	MV-T	2.004	0.996	0.004	0.000	2.002	0.998	0.002	0.000	2.006	0.994	0.006	0.000
(iii)	MV-TM	1.010	0.992	0.008	0.000	1.010	0.990	0.010	0.000	1.010	0.990	0.010	0.000
	MV-T	1.060	0.942	0.058	0.000	1.052	0.948	0.052	0.000	1.066	0.934	0.066	0.000

Note: "MV-TM" and "MV-T" represent the MV-SIS approaches with d_δ determined by the modified tuning parameter selection method introduced in Section 2.2 and the algorithm introduced in Ni and Fang (2016), respectively.

Table 2. Performance of the DCNI-SIS procedure in identifying active predictors in Experiment 1.

		$\varepsilon \sim N(0,1)$				$\varepsilon \sim t(3)$				$\varepsilon \sim \text{Cauchy}(0,1)$			
Case	Method	AMS	CF	OF	UF	AMS	CF	OF	UF	AMS	CF	OF	UF
(i)	DCNI-TM	5.066	0.506	0.274	0.220	4.860	0.508	0.220	0.272	5.582	0.322	0.282	0.396
	DCNI-T	5.312	0.438	0.352	0.210	5.258	0.450	0.306	0.244	5.952	0.270	0.372	0.358
(ii)	DCNI-TM	4.526	0.704	0.186	0.110	4.514	0.718	0.164	0.118	4.844	0.508	0.228	0.264
	DCNI-T	4.718	0.660	0.244	0.096	4.600	0.678	0.226	0.096	5.080	0.462	0.310	0.228
(iii)	DCNI-TM	5.038	0.474	0.272	0.254	4.910	0.478	0.232	0.290	5.510	0.274	0.268	0.458
	DCNI-T	5.362	0.426	0.328	0.246	5.126	0.428	0.308	0.264	6.064	0.236	0.350	0.414

Note: "DCNI-TM" and "DCNI-T" represent the DCNI-SIS approaches with d_n determined by the modified tuning parameter selection method introduced in Section 2.2 and the algorithm introduced in Ni and Fang (2016), respectively.

Table 2 shows that the DCNI-TM method behaves better than the DCNI-T method in that the former has larger CF values and smaller OF values than the latter, and the AMS values for the former are closer to the true model size than are those of the latter, regardless of the error distributions and missingness data mechanisms.

The results for $CP_1, \ldots, CP_4$ and CP_a in identifying active predictors with fixed $d_n = \lfloor n/\log(n) \rfloor$ are reported in Table 3. Table 3 indicates that the DCNI-SIS procedure outperforms the other feature screening procedures, because the former has the largest CP_a value of the 10 feature screening methods, regardless of the error distributions and missingness data mechanisms.

The results for the 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model sizes for the various feature screening methods are presented in Table 4. Once again, the DNCI-SIS procedure outperforms the other feature screening procedures in identifying active predictors because all the quantiles of the mini-

Table 3. Performance of CP_j and CP_a for various screening methods in identifying active predictors in Experiment 1.

		$\varepsilon \sim N(0,1)$					$\varepsilon \sim t(3)$					$\varepsilon \sim \mathrm{Cauchy}(0,1)$				
Case	Method	CP_1	CP_2	CP_3	CP_4	CP_a	CP_1	CP_2	CP_3	CP_4	CP_a	CP_1	CP_2	CP_3	CP_4	CP_a
	MI-I	0.994	0.996	1.000	0.882	0.874	0.992	0.998	1.000	0.904	0.896	0.952	0.998	1.000	0.786	0.748
	MI-S	0.044	0.998	1.000	0.040	0.002	0.036	0.996	1.000	0.020	0.000	0.030	0.994	1.000	0.034	0.000
	V-D	0.996	0.818	0.940	0.884	0.718	0.992	0.788	0.944	0.906	0.690	0.952	0.660	0.872	0.780	0.470
	MC-SIS	1.000	0.924	0.982	0.990	0.902	1.000	0.910	0.978	0.988	0.876	1.000	0.840	0.958	0.958	0.770
(i)	SIS	0.648	0.362	0.538	0.474	0.122	0.660	0.302	0.498	0.468	0.092	0.452	0.232	0.340	0.350	0.048
	SIRS	1.000	0.936	0.988	0.990	0.918	1.000	0.902	0.982	0.988	0.872	0.994	0.842	0.966	0.950	0.776
	DF-SIS	0.870	0.264	0.510	0.578	0.088	0.908	0.220	0.440	0.550	0.058	0.684	0.194	0.326	0.420	0.030
	DC-SIS	1.000	0.818	0.940	0.942	0.756	1.000	0.786	0.944	0.972	0.738	0.984	0.660	0.872	0.864	0.520
	DCNI-SIS	1.000	0.990	1.000	0.980	0.970	1.000	0.990	1.000	0.974	0.964	0.990	0.990	1.000	0.918	0.902
	CC	1.000	0.894	0.972	0.980	0.858	1.000	0.856	0.968	0.982	0.818	0.996	0.790	0.928	0.932	0.686
	MI-I	0.998	1.000	1.000	0.948	0.946	0.998	1.000	1.000	0.948	0.946	0.966	1.000	1.000	0.876	0.848
	MI-S	0.032	1.000	1.000	0.032	0.000	0.022	1.000	1.000	0.032	0.000	0.038	1.000	1.000	0.034	0.002
	V-D	0.998	0.722	0.988	0.950	0.686	0.998	0.670	0.964	0.948	0.614	0.966	0.522	0.936	0.876	0.438
	MC-SIS	1.000	0.852	0.994	1.000	0.848	1.000	0.792	0.992	0.996	0.780	1.000	0.706	0.976	0.988	0.680
(ii)	SIS	0.664	0.334	0.538	0.502	0.094	0.686	0.290	0.544	0.522	0.088	0.472	0.214	0.372	0.374	0.048
	SIRS	1.000	0.878	0.998	0.998	0.876	1.000	0.852	0.994	0.998	0.844	1.000	0.780	0.980	0.992	0.760
	DF-SIS	0.896	0.168	0.550	0.592	0.054	0.908	0.144	0.540	0.628	0.048	0.740	0.112	0.370	0.482	0.028
	DC-SIS	0.998	0.722	0.988	0.976	0.698	1.000	0.670	0.964	0.992	0.65	0.99	0.522	0.936	0.930	0.470
	DCNI-SIS	1.000	1.000	1.000	0.990	0.990	1.000	1.000	1.000	0.994	0.994	1.000	1.000	1.000	0.966	0.966
	CC	1.000	0.792	0.992	0.994	0.780	1.000	0.736	0.986	0.996	0.722	1.000	0.638	0.958	0.976	0.602
	MI-I	0.986	0.996	1.000	0.852	0.836	0.988	0.998	1.000	0.868	0.858	0.936	0.994	1.000	0.722	0.680
	MI-S	0.036	0.998	1.000	0.026	0.000	0.032	0.996	1.000	0.020	0.002	0.044	0.99	1.000	0.042	0.000
	V-D	0.986	0.758	0.904	0.856	0.622	0.988	0.732	0.922	0.872	0.594	0.936	0.604	0.778	0.718	0.380
	MC-SIS	1.000	0.914	0.978	0.982	0.882	1.000	0.882	0.970	0.986	0.842	1.000	0.800	0.912	0.964	0.728
(iii)	SIS	0.570	0.280	0.396	0.418	0.060	0.584	0.246	0.360	0.430	0.052	0.406	0.194	0.238	0.298	0.026
	SIRS	1.000	0.916	0.980	0.982	0.886	1.000	0.894	0.978	0.986	0.862	1.000	0.830	0.930	0.960	0.760
	DF-SIS	0.832	0.212	0.384	0.524	0.044	0.858	0.178	0.318	0.510	0.048	0.642	0.168	0.222	0.330	0.010
	DC-SIS	0.996	0.758	0.904	0.934	0.674	1.000	0.732	0.922	0.958	0.668	0.976	0.604	0.778	0.836	0.454
	DCNI-SIS	1.000	0.988	1.000	0.960	0.950	1.000	0.992	1.000	0.970	0.962	0.996	0.994	1.000	0.920	0.910
	CC	1.000	0.866	0.960	0.970	0.816	1.000	0.834	0.944	0.980	0.774	0.998	0.764	0.876	0.952	0.672

mum model sizes for the former are closest to the true model size of the 10 feature screening methods. That is, the DCNI-SIS procedure is more efficient in selecting active predictors than are other feature screening procedures.

Experiment 2. This experiment is designed to investigate the performance of the proposed DCNI-SIS under the assumption that $E(Y|X)$ is nonlinear in some components of X. In this experiment, we generate the response Y from the following regression model:

$$Y = \beta_1 I(X_1 > 0) + \beta_2 X_2 X_4 + \beta_3 |\sin(X_3)| + \beta_4 X_4^2 + \varepsilon,$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top = (4, 4.5, 3.5, 2.5, 0.0, \ldots, 0.0)^\top$, which implies that $X_1, \ldots, X_4$ are active predictors and $X_5, \ldots, X_p$ are inactive predictors, and $X = (X_1, \ldots, X_p)^\top$ is generated from a multivariate normal distribution with

Table 4. The 5%, 25%, 50%, 75%, and 95% quantiles of MMS for various screening methods in identifying active predictors in Experiment 1.

		$\varepsilon \sim N(0,1)$					$\varepsilon \sim t(3)$					$\varepsilon \sim \text{Cauchy}(0,1)$				
Case	Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
	MI-I	4	4	6	17	101.4	4	4	6	14	73	4	5	10	38	177.1
	MI-S	219.0	495.5	694.5	872.3	978	257.9	500.3	698.5	864.3	975.2	252.0	493.8	728.5	874.3	971
	V-D	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	MC-SIS	4	4	5	13	78.1	4	4	6	15	77.2	4	5	10	32.3	165.2
(i)	SIS	15	93.8	272	630.3	904.6	21	113	284.5	625.8	905.2	42.0	230.5	521.5	776	958.1
	SIRS	4	4	5	11.3	71	4	4	5	14.3	88.1	4	5	10	32	177.1
	DF-SIS	23.0	125	324.5	613.3	917	26.9	175.5	366.5	644.8	914.1	55.0	244.5	482	748.3	949.1
	DC-SIS	4	5	11	36.3	203.5	4	5	13	40	188.2	4	10	34	96	427.3
	DCNI-SIS	4	4	4	5	26	4	4	4	5.3	26.1	4	4	5	9.3	69.1
	CC	4	4	6	19	126.5	4	4	7	21	125.2	4	6	16	48.3	234.1
	MI-I	4	4	4	8	39.1	4	4	4	8	39	4	4	6	18	95.1
	MI-S	186.5	495.8	701.5	860	976	236.9	510.5	703	863.3	974.1	221.8	505.5	684.5	874.5	974
	V-D	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	MC-SIS	4	4	7	18	148.1	4	4	8	31	175.1	4	6	15	49	281
(ii)	SIS	22.9	107.8	293.5	595	915.3	23	114	288.5	584.3	892.2	40.0	232.3	517.5	796.3	963.2
	SIRS	4	4	5	15	94.2	4	4	6	18	149.3	4	5	10	35	241.3
	DF-SIS	34.9	167.3	372	640	926.1	41.0	201	415	684.5	941.2	68.0	264.5	489	749.3	934.1
	DC-SIS	4	5.8	14	54	289	4	6	17	70.3	328.1	4	12	41.5	142.3	550.3
	DCNI-SIS	4	4	4	4	9	4	4	4	4	11	4	4	4	6	24.2
	CC	4	5	9	33	220.8	4	5	11	45	240.2	4	7	21.5	86	390.1
	MI-I	4	4	7	21.3	135.1	4	4	7	18	80.2	4	6	15	53	191.1
	MI-S	214.8	481	705	870.8	977	223.0	513.5	722	869	976.1	249.9	478.8	684	853.5	974.1
	V-D	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	MC-SIS	4	4	6	15	100.8	4	4	7	21	108.2	4	6	12	42	206.3
(iii)	SIS	31.9	154	390	703.3	940	36.0	187	409	661.5	931.4	73	346.3	605.5	814	972.1
	SIRS	4	4	6	15	108	4	4	6	18	121.1	4	5	12	36	238
	DF-SIS	45.0	198.5	414.5	714.3	947.1	41.0	239.3	467.5	740	946.2	95.7	349	596	838.5	974.2
	DC-SIS	4	6	16	57	243.1	4	7	18.5	57.3	237.2	5	16	45	144.8	397.1
	DCNI-SIS	4	4	4	6	35.2	4	4	4	6	25.1	4	4	5	12	59.2
	CC	4	4	8	23.3	157.3	4	5	9	32	153.1	4	7	19	63.5	344.5

zero mean and covariance matrix $\Sigma = (\sigma_{ij})$, with $\sigma_{ij} = 0.5^{|i-j|}$, and ε is independently generated as in Experiment 1. It is assumed that X is completely observed, while Y is subject to missingness.

Similarly to Experiment 1, the missing indicator δ for Y is generated from a Bernoulli distribution with probability $\pi = \Pr(\delta = 1|Z)$ specified by

(i) $\pi = 0.2$ if $|X_3 - 1| \leq 1$, and 0.8 otherwise;

(ii) $\text{logit}(\pi) = \alpha_0 + \alpha_1 X_2 + \alpha_2 X_3$ with $(\alpha_0, \alpha_1, \alpha_2) = (-1.0, -4.0, -5.0)$;

(iii) $\pi = 0.2$ if $|X_3 - 0.02Y - 0.9| \leq 1.5$, and 0.8 otherwise.

Scenarios (i) and (ii) represent MAR mechanisms for the missing response Y. Scenario (iii) is a nonignorable missing mechanism, and is designed to investigate

the robustness of the DCNI-SIS procedure to a misspecification of the missingness data mechanism. Thus, the numbers of the important covariates associated with δ for Scenarios (i), (ii), and (iii) are one, two and one, respectively. The average missing proportions for Scenarios (i), (ii), and (iii) are about (48.5%, 48.5%, 48.5%), (51.1%, 51.1%, 50.8%), and (60.8%, 60.8%, 60.4%) for the error distributions (E1), (E2), and (E3), respectively.

For the 500 data sets generated from each of the three error distributions and missingness data mechanisms, the proposed DCNI-SIS procedure and the other methods are used to identify the active predictors in the considered nonlinear regression model and propensity score function. To estimate $F(Y)$ for missing Y, we take the same kernel function $K(u)$ as in Experiment 1. The same value of $d_{\min}$ is taken as that given in Experiment 1. To save space, the results for identifying the active covariates associated with the missing indicator and important predictors associated with the response are reported in Tables S1–S4 of the Supplementary Material. Tables S1–S4 yield similar observations to those given in Experiment 1.

Experiment 3. This experiment is designed to investigate the finite-sample performance of the proposed DCNI-SIS procedure under the assumption that $E(Y|X)$ has interaction terms on some components of X, and $\mathcal{M}_\delta \cap \mathcal{M}$ is empty or $\mathcal{M}_\delta \supset \mathcal{M}$. In this experiment, we generate the response from the following regression model:

$$Y = \exp(\beta_1 X_1^2) + \exp(\beta_2 X_1 X_2) + \exp(\beta_3 X_2 X_3) + \varepsilon,$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top = (3.0, 3.0, 3.0, 0.0, \ldots, 0.0)^\top$, and $X = (X_1, \ldots, X_p)^\top$ and ε are independently generated as in Experiment 2. It is assumed that X is completely observed, while Y is subject to missingness.

Similarly to Experiment 1, the missing indicator δ for Y is generated from a Bernoulli distribution with probability $\pi = \Pr(\delta = 1|Z)$, specified by

(i) $\text{logit}(\pi) = \alpha_0 + \alpha_1 X_3 + \alpha_2 X_{600}$, where $(\alpha_0, \alpha_1, \alpha_2) = (-1.0, -4.0, -4.0)$;

(ii) $\text{logit}(\pi) = \alpha_0 + \alpha_1 X_4 + \alpha_2 X_5$, where $(\alpha_0, \alpha_1, \alpha_2) = (-1.0, -3.0, -3.0)$;

(iii) $\text{logit}(\pi) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4$, where $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) = (-1.0, -4.0, -4.0, -4.0, -4.0)$.

Scenarios (i)–(iii) describe a MAR mechanism for the missing response Y. Scenario (i) shows that $\mathcal{M}_\delta \cap \mathcal{M}$ is not empty, Scenario (ii) indicates that $\mathcal{M}_\delta \cap \mathcal{M}$

is empty, and Scenario (iii) implies $\mathcal{M}_\delta \supset \mathcal{M}$. The average missing proportions corresponding to the three missingness data mechanisms (i), (ii), and (iii) are about (53.3%, 52.9%, 53.2%), (53.9%, 53.9%, 53.9%), and (49.4%, 49.2%, 48.9%) for the error distributions (E1), (E2), and (E3), respectively.

Again, for the 500 data sets generated from each of the three distributions for the random error, together with the three missingness data mechanisms, the proposed DCNI-SIS procedure and the other methods are used to screen active predictors in the considered nonlinear model and propensity score function. In estimating $F(Y)$ for missing Y, we take the same kernel function $K(u)$ as in Experiment 1, set the bandwidth as $h_j = C_j \hat{\sigma}_{z_j} n^{-1/5}$ and $C_j = 1$ for Scenarios (i) and (ii), and set $h_j = \hat{\sigma}_{z_j} n^{-1/6}$ for Scenario (iii). For the MV-SIS method, we take $d_{\min} = 2$ for scenarios (i) and (ii), and $d_{\min} = 4$ for scenario (iii). For the DCNI-SIS method, we take $d_{\min} = 3$. To save space, the results are given in Tables S5–S8 of the Supplementary Material. Tables S5–S8 yield similar observations to those in Experiment 1, which implies that the proposed DCNI-SIS procedure can be used to identify linear or nonlinear relationships between a response and predictors.

Experiment 4. This experiment is designed to investigate the finite-sample performance of the proposed DCNI-SIS method in the presence of discrete responses. To this end, we generate the response Y from a binomial distribution Binomial$(5, \theta)$ with the probability $\theta = \exp(a)/(1 + \exp(a))$, where $a = \beta_1 X_1 + \sin(\beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)$, and $\beta = (\beta_1, \ldots, \beta_p)^\top = (2.0, 3.0, 2.5, 2.8, 0.0, \ldots, 0.0)^\top$, which indicates that $X_1, \ldots, X_4$ are active predictors and $X_5, \ldots, X_p$ are inactive predictors. Here, $X = (X_1, \ldots, X_p)^\top$ is generated from a multivariate normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$, with $\sigma_{ij} = 0.5^{|i-j|}$. It is assumed that X is completely observed, while Y is subject to missingness.

To create missing data for Y, the missing indicator δ is generated from a Bernoulli distribution with probability $\pi = \Pr(\delta = 1|Z)$, specified by

(i) $\text{logit}(\pi) = \alpha_0 + \alpha_1 X_3 + \alpha_1 X_4$, where $(\alpha_0, \alpha_1, \alpha_2) = (-1.0, -4.0, -4.0)$;

(ii) $\text{logit}(\pi) = \alpha_0 + \alpha_1 X_4 + \alpha_2 X_5$, where $(\alpha_0, \alpha_1, \alpha_2) = (-1.0, -4.0, -4.0)$.

Scenarios (i) and (ii) describe a MAR mechanism for the missing response Y. Scenario (i) shows $\mathcal{M}_\delta \subset \mathcal{M}$, and Scenario (ii) indicates that $\mathcal{M}_\delta \cap \mathcal{M}$ is not empty. The average missing proportions among 500 replications corresponding to the two missingness data mechanisms (i) and (ii) are about 52.0% and 51.9%, respectively. To estimate $F(Y)$ for missing Y, we take the same kernel function $K(u)$ as in Experiment 1. The same value of $d_{\min}$ is taken as that given in

Experiment 3, regardless of the MV-SIS and DCNI-SIS methods. To save space, the results are given in Tables S9–S12 of the Supplementary Material. Tables S9–S12 show that the proposed DCNI-SIS procedure can identify relationships between a response (or indicators) and predictors well in the presence of discrete responses.

5. An Example

In this section, the cardiomyopathy microarray data set analyzed by Li, Zhong and Zhu (2012), Zhong et al. (2016), and Wang and Li (2018) is used to illustrate the proposed DCNI-SIS approach. The main purpose of this study is to identify the most influential genes for overexpression of a G protein-coupled receptor (Ro1). As an illustration, we take the Ro1 expression level as the response variable Y, and the other 6,319 gene expression levels are taken as predictors, that is, $X = (X_1, \ldots, X_{6319})^\top$. Only 30 specimens are observed (i.e., $n = 30$). Thus, n is rather small and $p = 6{,}319$ is sufficiently large. Li, Zhong and Zhu (2012) identified the most important two genes (i.e., Msa.2134.0 and Msa.2877.0) using the DC-SIS procedure. In this data set, Y and X are completely observed. To demonstrate the application of the DCNI-SIS procedure, we artificially assume that Y is MAR, and missing values for Y are generated by the following propensity score function:

$$\Pr(\delta = 1|Y, X) = \Pr(\delta = 1|Z) = \frac{\exp(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_{268})}{1 + \exp(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_{268})},$$

where $Z = (X_1, X_{268})^\top$, $\alpha_0 = 0.2$, $\alpha_1 = 9.0$, and $\alpha_2 = 5.0$, X_1 is the gene labeled Msa.1.0, and X_{268} is the gene labeled Msa.11254.0. The average missing proportion among the 500 replications corresponding to the missingness data mechanism is about 48.5%.

For this artificially created data set, the proposed DCNI-SIS procedure is used to identify the significant genes. For comparison, we also consider the SIS (Fan and Lv (2008)), SIRS (Zhu et al. (2011)), DC-SIS (Li, Zhong and Zhu (2012)), DF-SIS (Wu and Yin (2015)), MC-SIS (Wang et al. (2017)), BMI procedures (e.g., MI-I, MI-S, and V-D methods) (Wang and Li (2018)), and CC method.

Similarly to Experiment 1, we take the kernel function $K(\cdot) = \prod_{j=1}^{s_n^\delta} K_j(\cdot)$ as $K_j(z_j) = 0.75(1-z_j^2)_+$, and set the bandwidth as $h_j = C_j \hat{\sigma}_{z_j} n^{-1/5}$, where $C_j = 1$ and $\hat{\sigma}_{z_j}$ is the standard deviation of the observations $\{Z_{ij} : i = 1, \ldots, n; j = 1, \ldots, s_n^\delta\}$. For the given model size $d_1 = \lfloor n/\log(n) \rfloor = 8$ and $d_2 = 2\lfloor n/\log(n) \rfloor =$

Table 5. The selected predictors among 500 repetitions in the real example.

	$d_1 = 8$			$d_2 = 16$		
Method	Msa.2134.0	Msa.2877.0	CP_a	Msa.2134.0	Msa.2877.0	CP_a
MI−I	0.672	0.588	0.588	0.674	0.682	0.644
MI−S	0.000	0.000	0.000	0.000	0.000	0.000
V−D	0.672	0.588	0.588	0.674	0.682	0.644
MC−SIS	0.118	0.966	0.100	0.200	1.000	0.200
SIS	0.012	1.000	0.012	0.064	1.000	0.064
SIRS	0.474	1.000	0.474	0.734	1.000	0.734
DF−SIS	0.000	0.000	0.000	0.000	0.002	0.000
DC−SIS	0.356	0.684	0.244	0.722	0.974	0.706
DCNI−SIS	0.810	0.876	0.690	0.950	0.964	0.914
CC	0.040	0.812	0.030	0.070	0.938	0.070

16, the results for CP_1, CP_2, and CP_a in identifying the active predictors Msa.2134.0 and Msa.2877.0 are given in Table 5. Table 5 indicates that the DCNI-SIS procedure outperforms the other feature screening procedures in that the former has larger CP_a values than those of the latter.

6. Conclusion

In the missing data literature, it is common to artificially assume a parametric or semiparametric model with some prespecified covariates for the considered missingness data mechanism. However, the plausibility of the posited model is doubtful. To address this issue, we investigate the feature screening problem for ultrahigh-dimensional data in the presence of responses MAR. A new feature screening procedure is proposed to simultaneously select important predictors associated with the response variable and significant covariates associated with missing indicators. The procedure uses the nonparametric conditional mean imputation technique and the distance correlation and mean-variance indexes, measuring the dependence between two random variables. A modified iterative algorithm is developed to find the optimal tuning parameter, which works well for the missingness data mechanism model. The proposed feature screening procedure has the following merits. First, it is model-free, because it does not assume a regression model for the relationship between the response and the predictors, and does not require a missingness data mechanism model for the missing indicator. Second, it can be used to screen the nonlinear relationships between the response and the predictors, and the missing indictor and the covariates. Third, it is computationally feasible. Fourth, it is robust to a misspecification of the

propensity score function.

Under some regularity conditions, we show the sure screening property of the proposed screening procedure. Several simulation studies are conducted to investigate the finite-sample performance of the proposed screening procedure. Empirical results show that the proposed screening procedure works well, even if the missingness data mechanism model is misspecified and the proportion is high, and is robust to heavy-tailed distributions for the response. An example from the cardiomyopathy microarray data set illustrates the proposed screening procedure.

Although we consider the case that the response is MAR, the proposed feature screening procedure can be extended to missing not at random for the response and/or the covariates, which is quite challenging, because the propensity score function depends on the missing response. In addition, we did not consider the asymptotic distribution of $\hat{\omega}_k$, which is challenging because of the complicated analytic expression of $\hat{\omega}_k$. These topics are left to future research.

Supplementary Material

The online Supplementary Material includes technical proofs and tables.

Acknowldgments

The authors are grateful to the editor, associate editor, and two anonymous referees for their constructive suggestions. This work was supported by a grant from the Key Project of the National Natural Science Foundation of China (Grant No. 11731011), and the Growth Project of Young Scientific and Technological Talents in Colleges and Universities of Guizhou Province (Grant No. 2019177).

References

Bouř, P., Kůs, V. and Franc, J. (2017). Kernel and divergence techniques in high energy physics separations. *Journal of Physics: Conference Series* **898**, 072004.

Chang, J., Tang, C. and Wu, Y. (2013). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *The Annals of Statistics* **44**, 515–539.

Cheng, P. E. and Chu, C. K. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statistica Sinica* **6**, 63–78.

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

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.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

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

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

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

Huang, D., Li, R. and Wang, H. (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business and Economic Statistics* **32**, 237–244.

Lai, P., Liu, Y., Liu, Z. and Wan, Y. (2017). Model free feature screening for ultrahigh dimensional data with responses missing at random. *Computational Statistics and Data Analysis* **105**, 201–216.

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

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

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd Edition. Wiley, New York.

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

Ni, L. and Fang, F. (2016). Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification. *Journal of Nonparametric Statistics* **28**, 515–530.

Shao, L., Yu, Y. and Zhou, Y. (2016). Sure feature screening for high-dimensional dichotomous classification. *Science China: Mathematics* **59**, 2527–2542.

Tang, N., Xia, L. and Yan, X. (2019). Feature screening in ultrahigh-dimensional partially linear models with missing responeses at random. *Computational Statistics and Data Analysis* **133**, 208–227.

Tang, N., Zhao, P. and Zhu, H. (2014). Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica* **24**, 723–747.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288.

Wang, L., Liu, J., Li, Y. and Li, R. (2017). Model-free conditional independence feature screening for ultrahigh dimensional data. *Science China: Mathematics* **60**, 551–568.

Wang, Q. and Li, Y. (2018). How to make model-free feature screening approaches for full data applicable to the case of missing response. *Scandinavian Journal of Statistics* **45**, 324–346.

Wu, Y. and Yin, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika* **102**, 65–76.

Yan, X., Tang, N., Xie, J., Ding, X. and Wang, Z. (2018). Fused mean-variance filter for feature screening. *Computational Statistics and Data Analysis* **122**, 18–32.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

Zhong, W., Zhu, L., Li, R. and Cui, H. (2016). Regularized quantile regression and robust feature screening for single index models. *Statistica Sinica* **26**, 69–95.

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Linli Xia

Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming 650500, China.

School of Data Science, Tongren University, Tongren, Guizhou 554300, China.

E-mail: linli_xia1@163.com

Niansheng Tang

Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming 650500, China.

E-mail: nstang@ynu.edu.cn

(Received November 2019; accepted July 2021)