Open Journal of Statistics, 2023, 13, 850-871
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Ultra-High Dimensional Feature Selection and Mean Estimation under Missing at Random

Wanhui Li[1], Guangming Deng[2], Dong Pan[1*]

[1]Department of Basic Sciences, Guilin University of Technology at Nanning, Chongzuo, China
[2]College of Science, Guilin University of Technology, Guilin, China
Email: *923417479@qq.com

How to cite this paper: Li, W.H., Deng, G.M. and Pan, D. (2023) Ultra-High Dimensional Feature Selection and Mean Estimation under Missing at Random. *Open Journal of Statistics*, **13**, 850-871.
https://doi.org/10.4236/ojs.2023.136043

Received: October 21, 2023
Accepted: December 15, 2023
Published: December 18, 2023

Abstract

Next Generation Sequencing (NGS) provides an effective basis for estimating the survival time of cancer patients, but it also poses the problem of high data dimensionality, in addition to the fact that some patients drop out of the study, making the data missing, so a method for estimating the mean of the response variable with missing values for the ultra-high dimensional datasets is needed. In this paper, we propose a two-stage ultra-high dimensional variable screening method, RF-SIS, based on random forest regression, which effectively solves the problem of estimating missing values due to excessive data dimension. After the dimension reduction process by applying RF-SIS, mean interpolation is executed on the missing responses. The results of the simulated data show that compared with the estimation method of directly deleting missing observations, the estimation results of RF-SIS-MI have significant advantages in terms of the proportion of intervals covered, the average length of intervals, and the average absolute deviation.

Keywords

Ultrahigh-Dimensional Data, Missing Data, Sure Independent Screening, Mean Estimation

1. Introduction

With the increasingly powerful performance of computers and storage devices, people can easily access high-dimensional or even ultra-high-dimensional data, and more and more research fields urgently need ultra-high-dimensional data analysis methods, so the statistical analysis of ultra-high-dimensional data has become a focus of attention in recent years. The dimensional of ultra-high dimensional data is exponentially related to the sample size, and the dimension of the variable is much larger than the sample size, *i.e.*, $\log p = O\left(n^{\alpha}\right)$. Ultra-high-dimensional data to bring more information to the people at the same time also caused great difficulties in data organization and analysis, that is, often referred to as the "dimension of the disaster", which exists so that the traditional statistical analysis methods cannot be directly applied to the ultra-large data sets.

Data integrity is the most basic requirement of classical statistical methods for data, if the data are missing it is impossible to carry out statistical modeling. There are many uncertainties in the process of collecting and organizing data, which can lead to incomplete data, and the existence of missing data will not only increase the difficulty and complexity of statistical analysis, but also lead to the loss of validity of the results of statistical analysis. Barzi *et al.* [1] pointed out that, when the rate of missing data is very high, such as reaching more than 60%, any data interpolation method cannot restore the data effects of missing data. When the missing data rate is not high, interpolation methods can be used to improve the results of statistical analysis. The "dimension catastrophe" makes existing statistical inference methods applied to missing data impossible to apply because of too many co-variable and too few samples, so data dimension reduction is the first problem to be solved.

In high-dimensional statistical modeling, the feature screening method is more popular in practice. Tibshirani [2] proposed Least Absolute Shrinkage and Selection Operator(LASSO), an algorithm which, by adding L1 paradigm as a penalty term in the optimization process of minimizing the sum of squares of the residuals, makes it possible to solve the optimization variables with small absolute values of parameter estimates during the problem will be compressed to 0, thus obtaining a sparse regression model, but LASSO was difficult to solve until Bradley Efron *et al.* proposed Least Angle Regression(LAR), which made LASSO popular.

Penalizing the least squares method in ultra-high dimensional datasets encounters problems such as slow computation, algorithmic instability, and loss of statistical certainty, and often fails to yield better results. In the case of high dimension of co-variable, it is generally believed that the number of variables that have a significant effect on the response variable is generally small. Fan Jianqing *et al.* [3] proposed Sure Independence Screening (SIS), an ultra-high dimensional feature screening method. the SIS screening method uses the marginal correlation coefficient as the correlation measure of the variables, and sets a hard threshold for ranking and eliminates irrelevant co-variable at the bottom of the rankings, which is able to rapidly reduce the dimension of the data from ultra-high dimension to general high dimensionality, which makes it possible for the high dimensional statistical analysis methods to be used in a more efficient way. It makes the application of high-dimensional statistical analysis methods to ultra-high-dimensional data possible.

The SIS, proposed based on the linear model, measures its correlation using marginal correlation coefficients. However, these coefficients can only assess linear correlation between variables. As a result, the SIS is unable to screen out variables that have complex relationships with the response variable. Furthermore, strong correlations between co-variable can make it more difficult for SIS to identify important variables. It is inconceivable to set a specific model category for an ultra-high-dimensional model, which may lead to modeling errors. When the sample size is small, it is impossible to guarantee filterability. Later, Fan *et al.* [4] proposed ISIS (Iterative Deterministic Independence Screening) based on SIS, which solved this problem to some extent. However, the risk of model setting error still exists, and the problem of insufficient identification of correlation indicators based on marginal correlation coefficients remains unresolved. Recent studies have shown that model-free dimensionality reduction methods perform well in identifying nonlinear features and are more effective in identifying real variables in complex models than parametric methods. In this paper, we propose the Radom Forests Sure Independent Screening (RF-SIS), an ultra-high-dimensional variable screening method for random forests that is model-free. RF-SIS replaces marginal correlation coefficients with indicators of variable importance in random forest regression, allowing SIS to screen important variables with complex relationships with co-variable [5] [6] [7].

In cancer patient survival analysis, it is of utmost importance to address the challenges of missing data regarding patient survival time and the high dimensionality of the data. Firstly, the absence of survival time data can significantly impact the accuracy and completeness of the analysis results. Since survival time is a crucial variable in predicting patients' outcomes and assessing treatment effectiveness, its absence can lead to erroneous conclusions that may negatively impact the doctor's treatment plan.

Secondly, with the advancement of medical technology, the amount of collected data is increasing, resulting in higher dimensionality. However, filtering out irrelevant features and identifying key factors in these high-dimensional datasets is a daunting task. Feature selection becomes even more challenging to perform effectively, while avoiding overfitting, to extract the variables that truly impact patients' survival time.

Therefore, it is crucial to address these two issues to enhance the accuracy and reliability of cancer patient survival analysis.

2. Methods

2.1. Sure Independent Screening

In ultra-high dimensional datasets, we generally assume that the important variables are sparse. This means that only a few co-variables have a significant effect on the response variable. Based on the above assumptions, we compute the marginal correlations of all co-variable on the response variable, and then pre-screen them according to their ordering to achieve fast dimension reduction. The following linear regression model is considered:

$$Y = X\beta + \varepsilon . \tag{1}$$

Let $X = (x_1, x_2, \cdots, x_p)^{\mathrm{T}}$ be the p-dimensionally normalized covariate and y be a response variable, where $p \gg n$. let $\mathcal{A}^* = \{1 \le i \le p : \beta_i \ne 0\}$ denote the subscript of the corresponding active variable in the true model, *i.e.*, the indicator set of the true sparse model, and then consider the marginal correlation of the covariate to the response variable ω:

$$\omega = (\omega_1, \omega_2, \cdots, \omega_p)^{\mathrm{T}} = X^{\mathrm{T}} y, \tag{2}$$

ω_i denotes the marginal correlation coefficient of the co-variable on the response variable. For $\delta \in [0,1]$, we have an estimate $\mathcal{A}_\delta^*$ of the true sparse set of variables $\mathcal{A}_*$

$$\mathcal{A}_\delta^* = \{1 \le i \le p : |\omega_i|, \text{the first } [\delta n] \text{ in } |\omega|\}, \tag{3}$$

where $[\delta n] < n$.

The SIS efficiently filters variables weakly correlated with the response variable by reducing dimension from $p = \exp(O(n^\alpha))$ to $p' = [\delta n]$ through the computation of subsample correlation coefficients. SIS is a two-stage feature screening algorithm. SIS is a two-stage feature screening algorithm. The size of the candidate set, $\mathcal{M}_\gamma$, has not been thoroughly researched by scholars. We typically use the empirical value $d = [n/\log n]$ or $n-1$, as proposed by Fan [3]. In situations where the sample is sufficient, $|\mathcal{M}_\gamma|$ is usually set to $[n/\log n]$. Following pre-screening, we can reduce dimension using methods such as Lasso, elastic net Zou H *et al.* [8], Adaptive Lasso Zou H *et al.* [9] or SCAD Fan *et al.* [10] method (**Figure 1**).

2.2. Random Forest Regression

Random Forest is a machine learning algorithm proposed by Breiman [11] that integrates the advantages of Bagging and Random Subspace with the CART method. It avoids overfitting by using multiple predictive models and has excellent extrapolation prediction ability. The algorithm is effective in determining the significance of co-variable on response variables. Random forests are centered on the randomness of the sample data and the randomness of the variables (**Figure 2**).

As shown in **Figure 2**, random forest regression generates different regression trees in such a way that the variables considered in each tree are only a subset of all variables. In addition, the samples used to construct these regression trees are also sets obtained by resampling through the Bootstrap method. Therefore,

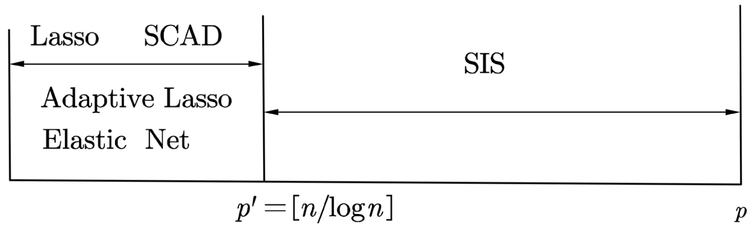


Figure 1. SIS ultra-high-dimensional model selection.

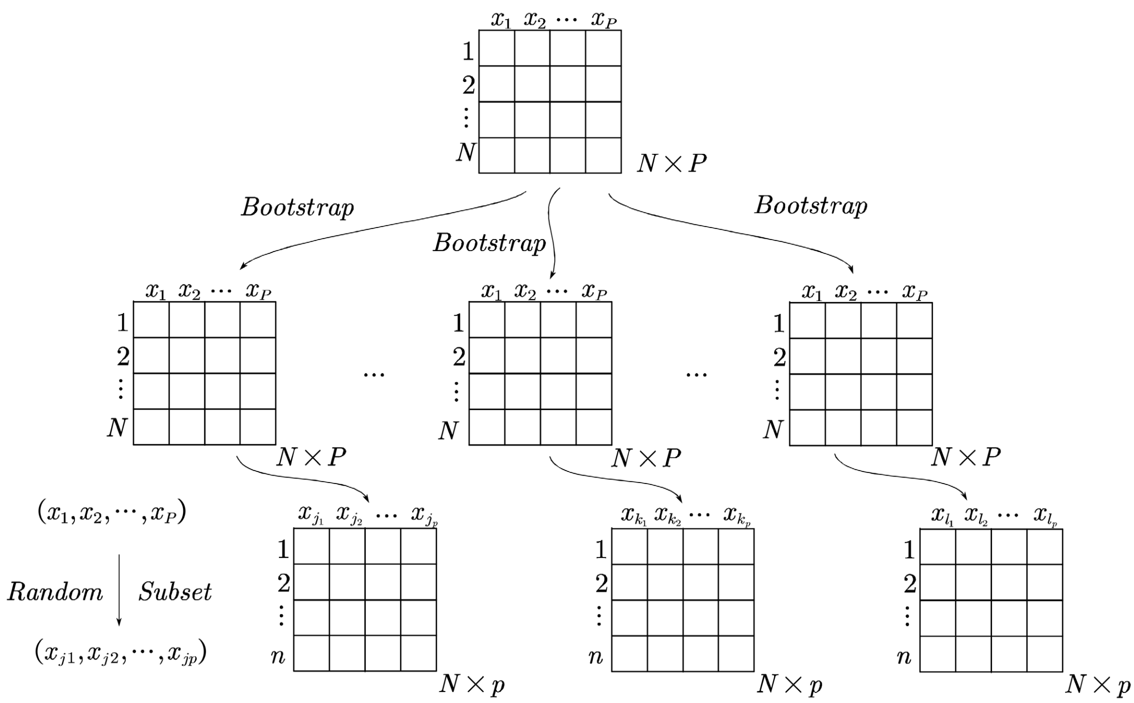


Figure 2. Variable randomization and sample randomization.

when forming a random forest, we can form diversified regression trees. This diversified construction method can improve the prediction accuracy and stability of the model.

Due to the presence of both variable and sample randomness, Random Forest is capable of simultaneously creating a multitude of distinct trees, while also maintaining their mutual independence and promoting prediction result diversity. This functionality not only resolves overfitting issues but also guarantees model precision significantly. Random Forest is inherently fit for parallel computing, which optimizes computer resources, and the program's run time remains manageable even with an increase in CART tree numbers.

2.3. Variable Importance Measures

Measurement of variable importance are complex problems. The problems of multicollinearity, nonlinear correlation, and combinatorial correlation of variables make the results of traditional statistical tests of variable parameters do not reflect the variable importance of the co-variable to the response variable well. In this paper, we use Permutation Importance Measure (PIM), which is a measure of variable importance based on the rearrangement mechanism of the change in the reduction of the mean square error.

Assuming that we have t out-of-bag sample sets, after the random forest modeling is completed, we obtain t out-of-bag $MSE_1, MSE_2, \cdots, MSE_t$, respectively, in the out-of-bag sample set OBB_t for all co-variable x_i sequentially and randomly rearranged to calculate MSE_{it}, to obtain the importance measure pimit of covariate X_i on this out-of-bag dataset (Figure 3).

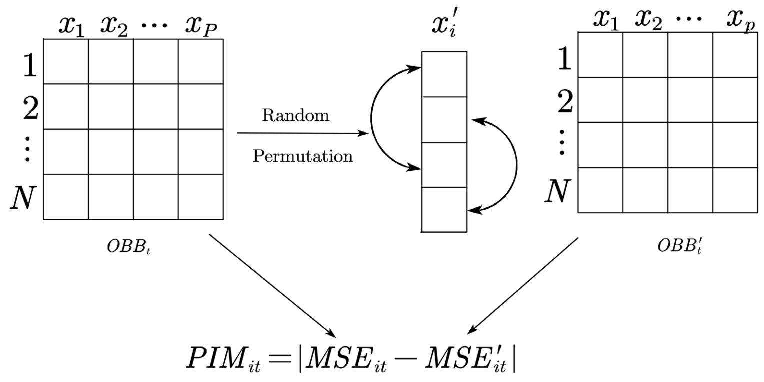


Figure 3. Permutation importance measure.

The matrix is obtained as follows:

$$\begin{bmatrix} PIM_{11} & PIM_{12} & \cdots & PIM_{1t} \\ PIM_{21} & PIM_{22} & \cdots & PIM_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ PIM_{p1} & PIM_{p2} & \cdots & PIM_{pt} \end{bmatrix}, \tag{4}$$

where each row PIM_{ij} represents the amount of change in the jth out-of-bag mean square error after rearranging for the ith variable to obtain the ith variable importance measure PIM_i:

$$PIM_i = \frac{1}{t}\sum_{j=1}^{t} PIM_{it} \tag{5}$$

If the covariate has a negligible effect on the response variable, then a change in the value of that variable will not result in a significant change in the MSE when assessing the model error. However, once a covariate has a significant effect on the response variable, then a change in the value of the covariate can have a significant effect on the MSE when it is calculated. The PIM is based on this principle and the order of the covariate values is rearranged when calculating the PIM. If a change in the value of a covariate results in a significant change in the mean square error, then that variable has a significant effect on the response variable, *i.e.*, the larger the PIM the more important the covariate x_i is.

2.4. Random Forest Sure Independent Screening

The correlation coefficients employed by SIS as an initial pre-screening measure of variable significance fail to accurately depict the effect of the variable on the response variable. Furthermore, SIS cannot guarantee screening certainty if the model is incorrectly established. For solving this issue, RF-SIS substitutes the marginal correlation coefficients of SIS with PIM, which is determined through the out-of-bag mean-square error variation based on the rearrangement mechanism.

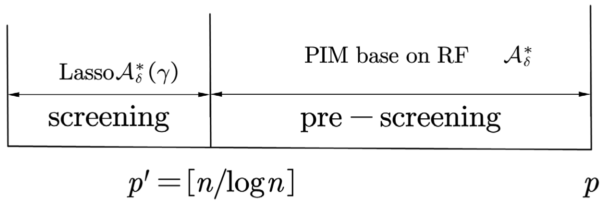
The algorithm follows the steps outlined below:

Step 1: Initialize y and X;

Step 2: Random forest regression modeling to calculate PIM_i;

Step 3: PIM_i descending order obtained to $\mathcal{A}_{\delta}^{*}$;

Step 4: Establishing a Lasso regression on (y, a) and obtaining $\mathcal{A}_{\delta}^{*}(\gamma)$ based on the BIC criteria.



2.5. Feature Screening Simulation

To assess the ability of random forests to screen variables, we perform the following numerical simulations using the SIS proposed by Fan (2010) as a benchmark method.

We considered a sample set for ultra-high dimensional simulation consisting of 100 and 200 samples and dimensions of 1000 and 2000, where the co-variable covariance matrix was set to $\Sigma = cov(X_i, X_j) = \rho^{|i-j|}$. The simulation involved different covariate autocorrelations categorized into three datasets for simulation, *i.e.*, low, medium, and high, which were repeated 500 times. To fully showcase the model-free screening effects of RF-SIS, we conducted five groups of experiments and compared them to Fan's ISIS method. The experimental model settings are listed below:

Model 1: $y = x_1 + x_2 + x_3 + \varepsilon, \varepsilon \sim N(0,1)$.

Model 2: $y = 2x_1 + x_2^2 + x_3^3 + \varepsilon, \varepsilon \sim N(0,1)$.

Model 3: $y = 2x_1 + x_2^2 + 3\sin(x_3) + \varepsilon, \varepsilon \sim N(0,1)$.

Model 4: $y = \cos(x_1) + x_2^2 + \sin(x_3) + \varepsilon, \varepsilon \sim N(0,1)$.

Model 5: $y = 3\cos(x_1) + 2x_2 - 2|x_3| + \varepsilon, \varepsilon \sim N(0,1)$.

The model evaluation indicators are as follows:

$p(x_i)$: Proportion of identifying variables x_i in 500 simulations.

$p(x_{all})$: Proportion of all variables identified in 500 simulations.

Model 1 is a linear model, and as can be seen from **Table 1**, Fan's SIS method is able to identify all the true co-variable at the same time when the correlation between the co-variable is low, *i.e.*, $\rho = 0.3$. When $\rho = 0.6$, the SIS method is able to identify 99% of x_2 in a small sample set, and as the sample increases to 200, SIS is able to identify all the true co-variable at the same time. And the RF-SIS method proposed in this paper still achieves more than 95% identification rate even though it cannot identify the true co-variable perfectly in all experimental conditions. Therefore, when the model is a linear model, the screening effect of RF-SIS is excellent based on its screening effect even if it cannot reach the height of SIS.

The Simulation 2 model is a non-linear polynomial model. **Table 2** shows that the accuracy of the marginal correlation-based SIS method reduces substantially when the model is non-linear. Nonetheless, it is evident that SIS in Simulation 2 cannot choose all variables simultaneously with a high percentage, regardless of

Table 1. Model 1: $y = x_1 + x_2 + x_3 + \varepsilon, \varepsilon \sim N(0,1)$.

	ρ	Method	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_{all})$
$n = 100$ $p = 1000$	0.3	SIS	1	1	1	1
		RF-SIS	1	0.970	0.978	0.950
	0.6	SIS	1	0.990	1	0.990
		RF-SIS	1	0.970	0.984	0.958
	0.8	SIS	0.996	0.882	0.962	0.840
		RF-SIS	1	0.992	0.990	0.984
$n = 200$ $p = 2000$	0.3	SIS	1	1	1	1
		RF-SIS	1	0.952	0.976	0.94
	0.6	SIS	1	1	1	1
		RF-SIS	1	0.966	0.968	0.948
	0.8	SIS	1	0.966	1	0.966
		RF-SIS	1	0.964	0.962	0.954

Table 2. Model 2: $y = 2x_1 + x_2^2 + x_3^3 + \varepsilon, \varepsilon \sim N(0,1)$.

	ρ	Method	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_{all})$
$n = 100$ $p = 1000$	0.3	SIS	0.542	0.092	0.384	0.026
		RF-SIS	0.9	0.798	0.766	0.512
	0.6	SIS	0.494	0.236	0.346	0.096
		RF-SIS	0.808	0.742	0.580	0.334
	0.8	SIS	0.462	0.342	0.402	0.204
		RF-SIS	0.646	0.646	0.484	0.196
$n = 200$ $p = 2000$	0.3	SIS	0.678	0.158	0.384	0.068
		RF-SIS	0.796	0.910	0.672	0.454
	0.6	SIS	0.670	0.310	0.404	0.184
		RF-SIS	0.754	0.890	0.472	0.332
	0.8	SIS	0.586	0.378	0.448	0.218
		RF-SIS	0.622	0.786	0.532	0.272

the scenario. The SIS is unable to select the true variables when the co-variable are nonlinearly correlated with the response variables, as $p(x_{all})$ is less than 10%, and the identification ratio of squared and quadratic terms is less than 50% for $n < 100$, $\rho < 0.8$. However, the RF-SIS enhances the screening effectiveness in all scenarios, including univariate and all-variable screening effects, compared to the SIS. Under the conditions of small sample size and low correlation, RF-SIS improves the proportion of all-variable screening from 0.026% to

51.2%. Additionally, both nonlinearly correlated variables x_2 and x_3 exceed 50% in terms of screening proportion.

Model 3 enhances the model's complexity by replacing the quadratic term with a sinusoidal function from Model 2. **Table 3** exhibits that SIS has a high sensitivity to the linear term x_1.

SIS identifies the linear term in all instances at 100%, and RF-SIS identifies it effectively as well. Regarding the square term x_2, RF-SIS outperforms SIS by identifying over 90% of it. Surprisingly, the SIS method demonstrates a strong ability to recognize the sinusoidal term, while RF-SIS exhibits a slightly weaker performance with recognition rates exceeding 50%. Furthermore, RF-SIS outperforms SIS in identifying full variables, with a significantly higher proportion. Although the identification ability of SIS is comparable to that of RF-SIS when covariate correlation is weaker and the sample size is larger, there is still a 10% difference between the two methods.

To further test the comparison between RF-SIS and SIS under a complex model, Simulation 4 substitutes the linear function of the linear term x_1 in Simulation 3 with the cosine function. The model now consists of an additive combination of cosine, sine, and quadratic functions. The results in **Table 4** reveal that the RF-SIS screening effect is stronger than SIS in both the univariate screening effect and all-variable screening ability. Specifically, in Model 4, RF-SIS demonstrates a robust capability to capture the square term, along with better ability to capture the sine-cosine function compared to SIS. Furthermore, while SIS lacks the ability to select all variables, with its selection percentage being less than 5%, RF-SIS is capable of selecting all variables with a percentage of approximately 50%.

Table 3. Model 3: $y = 2x_1 + x_2^2 + 3\sin(x_3) + \varepsilon, \varepsilon \sim N(0,1)$.

	ρ	Method	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_{all})$
$n = 100$ $p = 1000$	0.3	SIS	1	0.45	1	0.45
		RF-SIS	1	0.966	0.752	0.73
	0.6	SIS	1	0.380	1	0.380
		RF-SIS	0.986	0.904	0.652	0.558
	0.8	SIS	0.998	0.248	0.982	0.236
		RF-SIS	0.956	0.886	0.764	0.588
$n = 200$ $p = 2000$	0.3	SIS	1	0.500	1	0.500
		RF-SIS	1	0.984	0.616	0.604
	0.6	SIS	1	0.376	1	0.376
		RF-SIS	0.998	0.962	0.624	0.582
	0.8	SIS	1	0.272	0.998	0.272
		RF-SIS	0.982	0.904	0.660	0.554

According to **Table 5**, RF-SIS exhibits a high proficiency in identifying the cosine function under simulation condition five. Its selection ratio exceeds 90%. Additionally, it can screen absolute value terms with a screening ratio mostly above 50%. Conversely, SIS performs decently in identifying linear terms, but struggles to identify cosine and absolute value functions, resulting in a low overall identification rate for all variables. In comparison, RF-SIS exhibits better

Table 4. Model 4: $y = \cos(x_1) + x_2^2 + \sin(x_3) + \varepsilon, \varepsilon \sim N(0,1)$.

	ρ	Method	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_{all})$
$n = 100$ $p = 1000$	0.3	SIS	0.252	1	0.232	0.006
		RF-SIS	0.798	0.9	0.932	0.644
	0.6	SIS	0.082	0.186	0.710	0.010
		RF-SIS	0.718	0.890	0.908	0.548
	0.8	SIS	0.114	0.196	0.584	0.002
		RF-SIS	0.674	0.818	0.782	0.386
$n = 200$ $p = 2000$	0.3	SIS	0.064	0.172	0.952	0.016
		RF-SIS	0.696	0.988	0.920	0.634
	0.6	SIS	0.150	0.262	0.896	0.026
		RF-SIS	0.640	0.960	0.844	0.516
	0.8	SIS	0.206	0.250	0.838	0.030
		RF-SIS	0.520	0.944	0.588	0.278

Table 5. Model 5: $y = 3\cos(x_1) + 2x_2 - 2|x_3| + \varepsilon, \varepsilon \sim N(0,1)$.

	ρ	Method	$p(x_1)$	$p(x_2)$	$p(x_3)$	$p(x_{all})$
$n = 100$ $p = 1000$	0.3	SIS	0.252	1	0.232	0.06
		RF-SIS	0.964	0.714	0.758	0.532
	0.6	SIS	0.372	0.998	0.334	0.154
		RF-SIS	0.960	0.628	0.512	0.310
	0.8	SIS	0.406	0.994	0.370	0.166
		RF-SIS	0.902	0.606	0.500	0.274
$p = 2000$	0.3	SIS	0.398	1	0.328	0.136
		RF-SIS	0.964	0.686	0.568	0.370
	0.6	SIS	0.454	1	0.424	0.226
		RF-SIS	0.948	0.576	0.498	0.286
	0.8	SIS	0.364	1	0.318	0.152
		RF-SIS	0.896	0.628	0.424	0.234

abilities in identifying changes in all variables across all cases.

2.6. Missing at Radom

Little *et al.* [12] classify missing mechanisms into three main categories, Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR or Nonignorable)

Let Y represent an $n \times p$ matrix of complete data, where Y_{obs} indicates the observed data and Y_{miss} indicates the missing data that has not been observed. Let δ be an $n \times p$ matrix consisting of indicator variables, where the element δ_{ij} within the matrix indicates whether the data is missing or not. When $\delta_{ij} = 1$, it signifies that Y_{ij} is not missing, and when $\delta_{ij} = 0$, Y_{ij} is missing. The missing mechanism denotes the likelihood of the data being missing with respect to the missing mechanism, the missing data, and the observations. The missing probability, noted as $P(\delta \mid Y, \xi)$, and ξ, a parameter in the missing mechanism, are explained.

The missing at random mechanism means that the probability of missing data depends on Y_{obs}. In the Survey of Income Levels of the Population, the fixed income of respondents varies with age. For example, if the respondent is younger and has no fixed income, it will result in a missing fixed income. *i.e.*

$$P(\delta = 1 \mid Y_{obs}, Y_{miss}, \xi) = P(\delta \mid Y_{obs}, \xi). \tag{6}$$

The conditions for satisfying a random deletion are not strict, and when $P(\delta = 1)$ is related only to Y_{obs} and not to Y_{miss}, the deletion mechanism can be identified as MAR, which is more common in clinical medical research.

2.7. Response Variable Mean Estimation with MAR

The response variable for ultra-high dimensional data frequently has missing values. Consequently, it is necessary to utilize appropriate methods to address this issue. Given that existing data interpolation methods are inadequate for ultra-high dimensional data, this paper uses RF-SIS to screen its features. Afterwards, it employs established processing methods, including regression interpolation and inverse probability weighting, to handle the missing data. This is done to obtain a mean estimation of the response variable.

The response variable missing processing method under the random missing mechanism MAR cannot be applied to ultra-high dimensional data, and needs to be combined with the RF-SIS proposed in this paper to be able to use it.

For the Inverse Probability Weighting (IPW):

$$\hat{\mu}_{ipw} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i y_i}{P(\delta = 1 \mid X)}, \tag{7}$$

By the law of large numbers, it follows that

$$\hat{\mu} \xrightarrow{P} E\left[\frac{\delta y}{P(\delta = 1 \mid X)}\right], \tag{8}$$

$$
\begin{aligned}
E\left[E\left\{\frac{\delta y}{P(\delta=1\mid X)}\mid X\right\}\right] &= E\left[\frac{\delta y}{P(\delta=1\mid X)}E\{\delta\mid X\}\right] \\
&= E\left\{\frac{y}{P(\delta=1\mid X)}P(\delta=1\mid X)\right\} \\
&= E\{y\} = \mu.
\end{aligned} \tag{9}
$$

where $P = P(\delta=1\mid X)$ is Propensity. The propensity score is frequently estimated utilizing a logistic regression model. However, the variables acquired from RF-SIS pose a challenge for $P = P(\delta=1\mid X)$.

RF-SIS is able to reduce the variable dimension $\log p = O(n^{\alpha})$ to $d = [n/\log n]$. Assuming that the Lasso step fails, the dimension d is still high for regression modeling in regression interpolation and estimation of missing probabilities in the inverse probability weighting method. In order to be able to extract further effective information about the covariates on the response variable, SIR [13], the dimensionality of the estimated model is further reduced to obtain more effective modeling results.

The SIR method is shown below:

Step 1: Standardized X

Step 2: Divide y into H intervals, each containing roughly the same number of observations.

Step 3: Calculate the conditional expectation for all observations in each interval $E[X\mid y_H]$.

Step 4: Calculate their covariance matrix $E\{X_H - E[X\mid y_H]\}$.

Step 5: Solve for the eigenvalues and eigenvectors of this covariance matrix.

Step 6: Based on the magnitude of the eigenvalues, select the eigenvectors $\hat{B}$ corresponding to the largest d eigenvalues

Using SIR we can further compress the $\{X_{\mathcal{A}_\delta^*}\}$ into the $\hat{S} = \hat{B}X_{\mathcal{A}_\delta^*}$. Thereafter, $\tilde{\mu}_{ipw}$ is estimated using such $\hat{S}$, followed by IPW estimation of $\hat{\mu}_{ipw}$, based on the $\hat{P} = P(\delta=1\mid \hat{S})$

$$\tilde{\mu}_{ipw} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i y_i}{\hat{P}}. \tag{10}$$

For the regression interpolation technique, this paper utilizes full samples without any missing data and constructs a linear regression model using $\hat{S}$.

$$\hat{Y} = \hat{\beta}\hat{S}, \tag{11}$$

Interpolation of the response variable values using $\hat{Y}$ yields

$$\hat{y}_i = \delta_i y_i + (1-\delta_i)\hat{S}_i\hat{\beta}, \tag{12}$$

Use the sample means as the estimate for the response variable's mean

$$\tilde{\mu}_{mi} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i. \tag{13}$$

2.8. Mean Estimation Simulation

Consider the following linear model:

$$y_k = X_k^{\mathrm{T}} \beta + \varepsilon_k \tag{14}$$

where X is a random variable with a multivariate normal distribution and a covariance matrix $Cov(X_{ik}, X_{jk}) = \rho^{|i-j|}$, $k, i, j = 1, \cdots, n$. $\boldsymbol{\mu} = (1, 1, \cdots, 1)^{\mathrm{T}}$, $\varepsilon_k \sim N(0,1)$, $\beta = (5, 0, \cdots, 5, 5)^{\mathrm{T}} \in R^p$. Let the propensity score function be

$$P(\delta = 1 \mid X, \theta) = \frac{e}{1+e}, e = \exp(X\theta) \tag{15}$$

In order to increase the complexity of the model, X_2, X_3 are replaced by $U(-2,0)$ and $U(1,2)$, respectively. In order to verify the validity of this paper's method, all the samples are used as benchmarks, and the inverse probability weighting and mean interpolation methods are used to estimate the mean value of the complete samples after the missing samples are eliminated, respectively, and the effects are compared. In this paper, the parameters of sample size and number of covariates in the simulation process are set to $(n, p) = (100, 1000)$ and $(n, p) = (100, 1000)$ and the covariance matrix coefficient is set to $\rho = 0.3, \rho = 0.6, \rho = 0.8$, θ controls the missing rate of the datasets, and when $\theta_1 = (1, 0, 0, \cdots, 2, 2)^{\mathrm{T}}$, the average missing proportion is 20%, and when $\theta_2 = (-3, 0, 0, \cdots, 2, 2)^{\mathrm{T}}$, and $\theta_3 = (-6, 0, 0, \cdots, 2, 1)^{\mathrm{T}}$, the average missing proportion of the response variable of the simulated dataset is 40% and 60%, respectively, and simulation is carried out under the above parameter settings for 1000 times.

To measure the effectiveness of the proposed methodology, the following indicators are defined:

1) **CP**: Coverage Probability, Probability that $\hat{\mu}$ is included in the true 95% confidence interval.

2) **AB**: Average Bias, $n^{-1}|\mu - \hat{\mu}|$.

3) **AL**: Average Length, average length of confidence intervals.

Tables 6-8 present the estimation results of the simulated dataset, consisting of 100 samples, 1000 covariate dimensions, and varying degrees of correlation between the variables, ranging from weak to highly correlated. Our analysis shows that the estimation results remain largely unaffected by the strength of correlation among the covariates across the three sets of simulation results. In terms of estimating coverage probability (CP), the inverse probability weighting (IPW) and regression mean interpolation (MI) produce accurate mean estimates when the response variable's missing proportions are 20% and 40%. These methods achieve a true coverage proportion around 95%, with IPW estimators showing higher coverage probability than MI estimates overall. In contrast, the direct deletion (CC) method yields a mean estimate with CP lower than 0.95, and its maximum value is only 0.92. When the proportion of missing data reaches 60%, the estimation accuracy of both the IPW and MI methods are significantly affected. Specifically, the coverage probability of the IPW estimation decreases from over 95% to approximately 25%, whereas the coverage probability of the MI estimation decreases from 95% to around 70%. Although the MI method also experiences estimation inaccuracies, the degree of its decrease is

Table 6. $n = 100$, $p = 1000$, $\rho = 0.3$.

Method	Miss Rate	CP	AB	AL
Full	20%	0.940	0.750	3.713
CC		0.734	1.711	4.789
IPW		0.972	1.180	8.211
MI		0.932	0.803	3.675
Full	40%	0.934	0.736	3.719
CC		0.744	1.681	4.7968
IPW		0.98	1.152	8.41
MI		0.944	0.775	3.685
Full	60%	0.960	0.792	3.714
CC		0.810	2.021	6.473
IPW		0.253	7.581	12.896
MI		0.690	1.874	3.532

Table 7. $n = 100$, $p = 1000$, $\rho = 0.6$.

Method	Miss Rate	CP	AB	AL
Full	20%	0.941	0.845	4.033
CC		0.482	1.932	3.759
IPW		0.949	1.254	8.055
MI		0.944	0.845	4.018
Full	40%	0.942	0.842	4.033
CC		0.570	2.386	5.124
IPW		0.976	1.297	8.173
MI		0.930	0.869	4.005
Full	60%	0.945	0.847	4.032
CC		0.885	1.815	7.069
IPW		0.245	7.966	14.136
MI		0.715	1.779	3.908

Table 8. $n = 100$, $p = 1000$, $\rho = 0.8$.

Method	Miss Rate	CP	AB	AL
Full	20%	0.934	0.736	3.719
CC		0.744	1.681	4.796
IPW		0.98	1.152	8.415
MI		0.944	0.775	3.685

Continued

Full	40%	0.934	0.736	3.719
CC		0.744	1.681	4.796
IPW		0.98	1.152	8.415
MI		0.944	0.775	3.685
Full	60%	0.946	0.853	4.203
CC		0.920	1.598	7.286
IPW		0.271	7.716	15.844
MI		0.716	1.909	4.148

lower, indicating a certain level of robustness.

On the absolute mean bias AB, the MI method's absolute bias is nearly identical to the bias of the mean estimate acquired from complete data, which is on par. The sample mean gathered from complete data is an unbiased estimation, leading to the belief that the MI method's estimation is also unbiased. Meanwhile, the IPW's mean bias is excessively high, indicating that the estimation obtained by IPW is biased. When the proportion of missing data reaches 60%, the absolute mean bias of estimates obtained by the multiple imputation (MI) method is comparable to the absolute bias obtained by the direct deletion method, suggesting that MI is not unbiased. Additionally, the mean absolute bias of inverse probability weighting (IPW) estimates increases drastically, reaching up to seven times the conventional level, indicating that the IPW method is highly sensitive to missing proportions and lacks robustness.

In terms of the average length of confidence intervals (AL), when the missing proportion is 20% and 60%, both estimation methods show similar changes to the average absolute deviation. The confidence intervals estimated by multiple imputation (MI) are shorter in length and at the same level as the estimation effect obtained from complete data. It is noteworthy that the MI method's average confidence interval length remains insensitive to the missing proportion. Specifically, when the proportion of missing data reaches 60%, the MI method displays almost no change in the confidence interval length. This suggests that the MI method is robust with regard to this indicator. When the proportion of missing data reaches 60%, the interpolation method is less effective and estimates may not be as accurate as those generated by CC. These results align with Barzi's [1] description of the limited impact of the interpolation method under those circumstances.

Tables 9-11 present the simulation data for three groups with a sample size of 200 and a covariate dimension of 2000. The degree of correlation between the variables ranges from 0.3 to 0.8. The results reflect consistent variation in correlation coefficients among covariates as found in **Tables 6-8**, and the degree of correlation shows limited influence on the estimation results. Concerning estimated value coverage probability, the results align with **Tables 6-8** estimation

Table 9. $n = 200$, $p = 2000$, $\rho = 0.3$.

Method	Miss Rate	CP	AB	AL
Full	20%	0.950	0.523	2.639
CC		0.322	1.539	3.874
IPW		0.956	0.541	5.364
MI		0.954	0.521	2.629
CC	40%	0.958	0.537	2.637
IPW		0.542	1.574	3.409
MI		0.968	0.685	5.192
CC		0.962	0.550	2.618
Full	60%	0.958	0.540	2.641
CC		0.694	1.763	4.559
IPW		0.818	2.648	11.836
MI		0.876	0.705	2.594

Table 10. $n = 200$, $p = 2000$, $\rho = 0.6$.

Method	Miss Rate	CP	AB	AL
Full	20%	0.938	0.592	2.849
CC		0.169	1.933	2.650
IPW		0.927	0.649	3.043
MI		0.939	0.590	2.841
Full	40%	0.980	0.472	2.639
CC		0.560	1.563	3.447
IPW		0.980	0.671	5.212
MI		0.920	0.523	2.615
Full	60%	0.936	0.609	2.846
CC		0.880	1.289	4.923
IPW		0.828	2.545	11.910
MI		0.860	0.756	2.812

Table 11. $n = 200$, $p = 2000$, $\rho = 0.8$.

Method	Miss Rate	CP	AB	AL
Full	20%	0.949	0.600	2.973
CC		0.105	2.202	2.749
IPW		0.936	0.652	3.073
MI		0.950	0.601	2.966

Continued

Full	40%	0.949	0.600	2.973
CC		0.105	2.202	2.749
IPW		0.936	0.652	3.073
MI		0.950	0.601	2.966
Full	60%	0.950	0.597	2.977
CC		0.916	1.200	5.154
IPW		0.828	2.876	12.768
MI		0.848	0.807	2.951

performance, where the IPW and MI methods provide superior estimations with low missing proportion. When the sample size is increased to 200, the coverage probability of the IPW method and MI method significantly improves under a missing proportion of 60%. Specifically, the coverage probability of the IPW method increases from 25% to 80%, while the coverage probability of the MI method increases from 70% to 85%. These findings suggest that increasing the sample size can improve the accuracy of estimation values when facing a missing proportion of 60%.

As the sample size increases from 100 to 200, the MI estimate's absolute mean bias significantly decreases when the missing proportion is at 60%, indicating that increasing the sample size permits the MI method estimate to regain unbiasedness. Additionally, it appears that the problem of excessive bias in the IPW estimate is addressed.

When the sample size is increased by 200, the average length of intervals decreases for the CC and IPW methods with a missing proportion of 60%. Specifically, the length decreases from approximately 7 to 5 for the CC method and from approximately 15 to 12 for the IPW method. This shows that increasing the sample size can slightly alleviate the problem of excessively long estimation confidence intervals. Additionally, the MI method remains highly robust for estimating the confidence interval length.

2.9. Real Data Example: Ovarian Cancer

Next-Generation Sequencing (NGS), also known as high-throughput sequencing, refers to a number of different modern sequencing technologies that allow us to sequence DNA and RNA much faster and cheaper than the previously used Sanger sequencing. These NGS technologies can generate thousands or millions of sequences concurrently, enabling a wide variety of applications and opening new avenues of genomic research.

NGS is a powerful tool for analyzing the genetic sequence of cancerous tissues. When there is a mutation in the DNA, normal cells have the potential to transform into cancerous cells. Using NGS, we can determine not only the specific type of cancer, but also the stage of the cancer. In addition, NGS can quantify the expression levels of cancer genes. This data reflects the current status of a patient's cancer lesions and can be combined with existing clinical records to create a predictive model of patient survival time in relation to gene expression levels. Healthcare practitioners can then use the model to make an estimate of a patient's survival time, guiding them to develop an appropriate and effective treatment plan.

The genetic data analyzed in this paper was sourced from biopsies of lesion sites taken from 562 ovarian cancer patients through the TCGA program. The dataset includes 16,383 gene expression levels, which is considerably larger than typical clinical medical data. Consequently, it is characterized by high dimensionality, small samples, and high noise. The datasets include information on Status, Days to death, and Days to last follow-up. In this paper, we used the duration until death as the response variable with gene expression level serving as a covariate to estimate the mean (**Figure 4**).

The TCGA report indicated that TP53 gene is one of the important oncogenes. In this section, the proposed RF-SIS ultra-high dimensional feature screening method will be used to perform feature screening with the TP53 gene expression level as the response variable and the rest of the gene expression as the independent variable, and a comparison will be made using the SIS proposed by Fan, QC-SIS by Ma X. *et al.* [14], and BC-SIS by Pan *et al.* [15] to examine the different method The similarities and differences of the screened genes. The overlapping results of RF-SIS ultra-high dimensional feature screening with SIS, QC-SIS, and BC-SIS are shown in the following results (**Table 12**, **Table 13**).

Only two genes were selected for SIS, numbered 4590 and 10197, and it is worth noting that genes 4590 and 10197 also existed in the variable sets of RF-SIS, QC-SIS, and BC-SIS. 8 genes intersected with RF-SIS, QC-SIS, and BC-SIS were 4590, 10197, 11389, 4872, 11487, 4812, 4973, 2691, 1713, and 3991; 18 genes intersected with BC-SIS and 19 genes intersected with QC-SIS; and 19 genes intersected with BC-SIS. 11487, 4812, 4973, 2691, 1713, 3991; the total number of genes intersecting with BC-SIS is 18, while the total number of genes having intersection with QC-SIS is 19. The above results show that the screening results of RF-SIS proposed in this paper can achieve the same screening effect as the genes that can be screened by the methods proposed by the previous authors. In the actual screening, RF-SIS screened a total of 63 genes, and on this basis provided medical researchers with genes that may be associated with key oncogenic TP53.

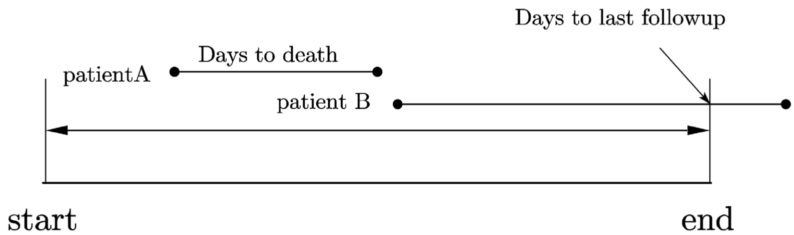


Figure 4. Patient survival over the course of the project.

Table 12. Comparative results of four different feature screenings.

SIS	RF-SIS	BC-SIS	QC-SIS	RF-SIS	BC-SIS	QC-SIS
4590	4590	4590	4590	15370		15,370
10,197	10,197	10,197	10,197	1213		1213
	11,389	11,389	11,389	11,556	11,556	
	4827	4827	4827	9666		9666
	11,487	11,487	11,487	7420	7420	
	4812	4812	4812	8241	8241	
	4973	4973	4973	7764	7764	7764
	2691	2691	2691	11,453	11,453	11,453
	1713	1713	1713	771		771
	3911	3911	3911	12,973		12,973
	3521		3521	1154	1154	
	14,285		14,285	6458	6458	6458
	15,630	15,630		15,484	15,484	
	10,486	10,486		15,018		15,018
	2866	2866		1764	1764	

Table 13. Genetic intersection.

Gene number	Name
4590	DULLARD
10,197	NFKB2
11,389	PFN1
4872	EMB
11,487	PI4K2A
4812	EIF4A1
4973	EPS15L1
2691	CCDC107
1713	C15orf5
3911	CUEDC2

In the datasets used in this paper, Days to Death has a total of 277 pieces of missing data, with a missing proportion of 50.71%. When the sample size is higher than 200 and the missing proportion of the response variable is not higher than 60%, the mean estimation method with random missing covariates given in this paper is able to effectively obtain the mean estimation of patient survival time.

The estimation results are as follows:

Table 14. Estimation of patient survival time.

Method	Mean	lower	upper	Interval Length
CC	1116.54	1028.319	1204.761	176.442
MI	1101.715	1063.852	1141.675	77.823
IPW	1097.433	974.3215	1220.545	246.223

Table 14 gives the estimation of the mean value of the survival time of the patients under missing randomization. From the results, the mean survival time of the 562 ovarian cancer patients who participated in the trial was 1116.54, 1101.715, and 1097.433 days, respectively, *i.e.*, the mean survival time of the patients was about 3 years, which shows that ovarian cancer is a malignant tumor that poses a serious threat to the life and health of women. Taking the direct deletion method as the standard, it can be found that the estimates obtained by the direct deletion method are not much different from the MI estimation and the IPW estimation, but the length of the confidence interval of the MI estimation method is significantly smaller than that of the CC and the IPW, which indicates that the estimates obtained by the MI method are more accurate, and it can effectively solve the problem of large standard deviation of the estimation that occurs in the direct deletion method. In addition, IPW estimation performs poorly in the actual data, with a confidence interval length of 246.223, which is significantly higher than that of CC estimation, indicating that the scope of application of IPW estimation has limitations, and that only by correctly setting the form of the propensity score function and correctly estimating the probability of missing can we get better results.

3. Discussion

This study examines how to cope with the "dimensional catastrophe" and missing data problems in ultra-high dimensional large-scale datasets. The article proposes an improved SIS method, the RF-SIS method, which uses random forest regression without model setup and utilizes the change in the mean squared error of out-of-bag data as a variable importance measure, thus effectively identifying real variables in complex models. When dealing with missing data, the IPW and MI estimates were obtained by weighting the missing response data using the logistic model estimation propensity score function and by completing the missing data using regression interpolation, respectively. Both methods yielded better estimates when missingness was not higher than 60%. Finally, the validity of these methods was verified by applying them to the estimation of survival time in ovarian cancer patients.

RF-SIS-MI is an advanced method for dealing with missing values of response variables in high dimensional data. This method incorporates techniques such as Random Forest Regression, sure independent screening, and Mean Imputation to obtain important variables through the RF-SIS step and utilize them to fill in the missing data, which improves the data completeness and accuracy of the predictive model. As a result, physicians can utilize more accurate predictions to develop more effective treatment plans.

RF-SIS excels in identifying variables that have complex correlations with response variables, which helps researchers better understand the mechanisms of cancer development. Compared with other interpolation methods, RF-SIS-MI is simpler and more robust, which reduces the bias introduced by interpolation errors and improves the credibility of research results. In addition, RF-SIS-MI can help physicians identify which patients are likely to have a longer or shorter survival period, so that medical resources can be allocated in a targeted manner with the aim of improving treatment outcomes.

4. Conclusion

In summary, the RF-SIS-MI method has an important impact on the survival analysis of cancer patients as well as clinical decision-making or research, which not only improves the prediction accuracy and discovers important features, but also enhances the robustness of the study results and optimizes the allocation of resources.

Acknowledgements

This paper is funded by the 2020 Guangxi Vocational Education Teaching Reform Research Project, Project No. GXGZJG2020B153.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Barzi, F. and Woodward, M. (2004) Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies. *American Journal of Epidemiology*, **160**, 34-45. https://doi.org/10.1093/aje/kwh175

[2] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[3] Fan, J.Q. and Lv, J.C. (2008) Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **70**, 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

[4] Fan, J. and Song, R. (2010) Sure Independence Screening in Generalized Linear Models with NP-Dimensionality. *The Annals of Statistics*, **38**, 3567-3604. https://doi.org/10.1214/10-AOS798

[5] Li, K., Wang, F., Yang, L. and Liu, R. (2023) Deep Feature Screening: Feature Selection for Ultra-High-Dimensional Data via Deep Neural Networks. *Neurocomputing*, **538**, Article ID: 126186. https://doi.org/10.1016/j.neucom.2023.03.047

[6] Zhou, L. and Wang, H. (2022) A Combined Feature Screening Approach of Random Forest and Filterbased Methods for Ultra-High Dimensional Data. *Current Bioinformatics*, **17**, 344-357. https://doi.org/10.2174/1574893617666220221120618

[7] Cheng, X. and Wang, H. (2022) A Generic Model-Free Feature Screening Procedure for Ultra-High Dimensional Data with Categorical Response. *Computer Methods and Programs in Biomedicine*, **229**, Article ID: 107269. https://doi.org/10.1016/j.cmpb.2022.107269

[8] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[9] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[10] 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. https://doi.org/10.1198/016214501753382273

[11] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[12] Little, R.J. and Rubin, D.B. (2019) Statistical Analysis with Missing Data. John Wiley and Sons, Hoboken, NJ. https://doi.org/10.1002/9781119482260

[13] Li, K. (1991) Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, **86**, 316-327. https://doi.org/10.1080/01621459.1991.10475035

[14] Ma, X. and Zhang, J. (2016) Robust Model-Free Feature Screening via Quantile Correlation. *Journal of Multivariate Analysis*, **143**, 472-480. https://doi.org/10.1016/j.jmva.2015.10.010

[15] Pan, W., Wang, X., Xiao, W., *et al.* (2019) A Generic Sure Independence Screening Procedure. *Journal of the American Statistical Association*, **114**, 928-937. https://doi.org/10.1080/01621459.2018.1462709