Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 33(8): 4816–4836.
DOI: 10.3934/era.2025217
Received: 22 June 2025
Revised: 03 August 2025
Accepted: 07 August 2025
Published: 19 August 2025

Research article

Robust variable selection for ultrahigh-dimensional linear models with nonignorable missing response

Yanting Xiao* **and Yifan Shi**

Department of Applied Mathematics, Xi'an University of Technology, Xi'an 710048, China

* **Correspondence:** Email: xiaoyanting03@163.com.

Abstract: We have proposed a robust and efficient variable selection method for ultrahigh-dimensional linear models with nonrandomly missing responses, leveraging modal regression. The propensity score function was specified by a semiparametric model and we introduced a two-step estimation procedure. In the first feature screening stage, the Pearson chi-square (PC) test statistic identifies significant predictors in the sparse propensity score model. The generalized method of moment (GMM) estimates parameters to obtain consistent estimation for the propensity score in the second stage. With the estimated propensity score, we suggested a feature screening and variable selection procedure based on the inverse probability weighting (IPW). A modified sure independence screening (SIS) method first reduces the model dimensionality, followed by a penalized modal regression approach to select significant covariates. The proposed procedure can deal with the ultrahigh-dimensional data with nonignorable nonresponse, and this modal-based procedure is robust against outliers and heavy-tailed errors. Additionally, we established the asymptotic properties of the estimators under mild regularity conditions. Simulation studies and real data applications confirm the method's effectiveness in finite samples and practical settings.

Keywords: variable selection; ultrahigh-dimensional data; nonignorable missing response; modal regression; inverse probability weighting

1. Introduction

With rapid advancements in data collection capabilities and explosive growth in data storage capacity, ultrahigh-dimensional data have become prevalent across various scientific fields, including financial markets, image processing, gene expression, and others. Under the setting of ultrahigh-dimensional data, when the covariate dimension p increases exponentially with the sample size n, traditional statistical methods not only consume a large amount of computational cost, but also suffer from reduced statistical inference accuracy and algorithm instability. This leads to some existing penalized variable

selection methods, such as LASSO [1], the smoothly clipped absolute deviation (SCAD) [2], adaptive LASSO [3], elastic net [4], and minimax concave penalty (MCP) [5], behaving poorly in handing ultrahigh-dimensional data. To this end, numerous feature screening procedures have been proposed primarily to largely reduce the dimensionality of covariates to a moderate scale, enabling the application of traditional variable selection procedures to the simplified models and ultimately achieving the determined model. The seminal work by Fan and Lv [6] introduced a sure independent screening (SIS) procedure based on the marginal Pearson correlation for linear models. Subsequent studies have extended SIS to various models, including generalized linear models [7], additive models [8], varying coefficient models [9], and partially linear models [10]. Parallel developments have produced model-free feature screening approaches, such as sure independence rank screening (SIRS) [11], SIS based on distance correlation (DC) [12], a robust rank correlation screening using Kendall τ correlation [13], and others.

Missing data frequently occurs in various fields such as biomedicine, economics, sociology, survey sampling, and others, due to various reasons such as unwillingness to answer sensitive questions in market surveys, uncontrollable factors leading to information loss, or the accidental death or disappearance of individuals in medical tracking. Little and Rubin [14] classified missing data mechanisms into three main types: missing completely at random (MCAR), missing at random (MAR), and missing nonignorable at random (MNAR). There has been considerable work on variable selection for various semiparametric models with response or covariates missing at random. For instance, Zhao and Xue [15] introduced a variable selection procedure by combining basis function approximations with penalized estimating equations for varying-coefficient partially linear models with missing response at random. Sherwood [16] considered variable selection for additive linear quantile regression models using inverse probability weighting. Wang and Song [17] developed variable selection via penalized quasi-maximum likelihood for spatial autoregressive model with missing response. Meanwhile, researchers have recognized the importance of developing feature screening techniques for ultrahigh-dimensional data to reduce the dimensionality of predictor variables to a moderate scale, particularly in the presence of missing data. Lai et al. [18] proposed a model-free feature screening procedure based on inverse probability weighting, where the Kolmogorov filter method was used to screen important features. Tang et al. [19] investigated ultrahigh-dimensional partially linear models under longitudinal data using profile marginal kernel-assisted estimating equations and imputation techniques. Li et al. [20] developed a novel nonparametric feature screening procedure based on conditionally imputed marginal Spearman rank correlation.

The aforementioned works focus on the MAR mechanism. However, nonresponse may depend on the value of the unobserved response itself in practice, a scenario referred to as the MNAR mechanism. Compared to the MAR assumption, MNAR is a great challenge to handle, because certain parameters are not identifiable without additional restrictions on the propensity score model. In recent years, research on MNAR data analysis has attracted widespread attention. Kim and Yu [21] first proposed an exponential tilting model for the propensity score. Wang et al. [22] introduced an instrumental variable that is associated with the study variable but independent of the propensity. Shao and Wang [23] constructed instrumental estimating equations and established the identifiability of unknown parameters. Tang and Ju [24] provided a comprehensive review of statistical inference for nonignorable missing data problems, including estimation, influence analysis, and model selection. Therefore, how to achieve robust variable selection in ultrahigh-dimensional MNAR data while ensuring computational efficiency

and statistical interpretability has become an urgent problem to be solved.

In this paper, we propose a robust variable selection procedure for ultrahigh-dimensional linear models with nonignorable missing response. Our approach first estimates the propensity score model through a two-step procedure. In the initial screening step, we use a Pearson chi-square-based feature screening procedure [25] to discretize continuous features and identify potentially important covariates. In the model estimation step, generalized moment estimation (GMM) is used to fit a sparse propensity score model, leveraging excluded variables from the initial screening as instrumental variables to address identifiability under MNAR mechanisms. Based on the estimated propensity score, we then develop a feature screening and variable selection procedure utilizing the inverse probability weighting technique. A modified sure independent screening (SIS) procedure is used to reduce the dimension of the linear model to a moderate scale. Then, a penalized objective function based on the modal regression, introduced by Yao et al. [26], and the SCAD penalized function are constructed to identify and estimate the parameter of active predictors simultaneously. Theoretically, the proposed variable selection procedure enjoys both consistency and the oracle property. At last, some simulations and a real example are conducted to assess the performances of the proposed procedures in finite samples.

The key contributions of this work are twofold. First, we develop a two-step propensity score estimation strategy that efficiently handles ultrahigh-dimensional covariates under MNAR, achieving screening consistency through discretization-based feature selection while resolving identifiability issues via GMM with instrumental variables. Second, we construct a robust variable selection procedure that integrates weighted SIS for scalable dimension reduction and modal regression with SCAD penalization, offering resistance to outliers, achieving asymptotic efficiency under normal errors, and computational tractability. Thereby, we provide a complete solution for robust variable selection in ultrahigh-dimensional nonignorable missing data.

The rest of this paper is organized as follows. Section 2 introduces a robust variable selection procedure based on modal regression for ultrahigh-dimensional data with nonignorable missing response. Section 3 establishes the asymptotic properties of the proposed method, including the sure screening property, estimator consistency, and oracle property. In Section 4, we discuss the selection of regularization parameters and the details of the estimation algorithm. Simulation studies and a real example are conducted to evaluate the performances of the proposed estimation procedures in Sections 5 and 6. We make our concluding remarks in Section 7, while technical proofs are provided in the Appendix.

2. Variable selection procedure

2.1. Penalized modal regression with nonignorable nonresponse

Suppose that $\{(Y_i, \boldsymbol{x}_i) : i = 1, \cdots, n\}$ is an independent and identically distributed sample from the linear model

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, i = 1, ..., n, \tag{2.1}$$

where $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ represents the p-dimensional covariate vector, Y_i is the response variable, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ denotes the unknown parameter vector, and ε_i is the model error with mean zero and variance σ^2. In ultrahigh-dimensional settings, we assume that $\boldsymbol{\beta}$ exhibits sparsity, meaning only a small number of covariates in model (2.1) are significantly associated with the response. For the convenience of expression, let $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)^T$ be the n-dimensional response vector, and

$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^T = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_p)$ denote the $n \times p$ design matrix, where $\boldsymbol{X}_j = (x_{1j}, x_{2j} \cdots, x_{nj})^T$ represents the jth column of $\boldsymbol{X}$ corresponding to the jth covariate. In addition, we define $A_0 = \{1 \leq j \leq p | \beta_j \neq 0\}$ as the index set of the nonzero coefficients, indicating the important (active) covariates, and A_0^c as its complement.

We begin by assuming that the observations $\{(Y_i, \boldsymbol{x}_i) : i = 1, \cdots, n\}$ from model (2.1) are completely observable. To reduce dimensionality efficiently, we first apply a screen feature procedure to preliminarily select a subset of covariates as candidates for important predictors. This step rapidly reduces the model (2.1) to a moderate scale. However, since the initially selected set may still include some irrelevant variables, we further refine the candidate set using a penalized variable selection method. Additionally, we standardize each predictor and the response to have mean zero and standard deviation one. The Pearson correlation coefficient between the jth predictor X_j and the response Y is given by

$$\rho_{0j}(X_j, Y) = \frac{\text{cov}(X_j, Y)}{\sigma_{X_j}\sigma_Y} = E(X_j Y), j = 1, \cdots, p, \tag{2.2}$$

which measures the strength of association between X_j and Y. The sample estimator of $\rho_{0j}(X_j, Y)$ is

$$\hat{\rho}_{0j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} Y_i, j = 1, \cdots, p. \tag{2.3}$$

Following the approach of Fan and Lv [6], we define the estimated set of active predictors as

$$\hat{A}_0 = \{j : |\hat{\rho}_{0j}| \text{ is among the top } d \text{ largest of all}\}.$$

Let $\boldsymbol{x}^*$ denote the d-dimensional vector of selected active covariates, and let $\boldsymbol{\beta}^*$ be the corresponding coefficient vector.

Thus, the original p-dimensional model (2.1) is reduced to a d-dimensional model, where $d < p$. Applying variable selection methods to this reduced-dimensional model allows for more effective identification of active covariates. Motivated by the advantages of modal regression, we propose a variable selection procedure to obtain a consistent estimator of $\boldsymbol{\beta}^*$ by maximizing the following objective function:

$$Q(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} \phi_h(Y_i - \boldsymbol{x}_i^{*T}\boldsymbol{\beta}^*) - n\sum_{k=1}^{d} p_{\lambda_k}(|\beta_k^*|), \tag{2.4}$$

where $\phi_h(\cdot) = \phi(\cdot/h)/h$ is a scaled kernel density function with bandwidth h, and $\phi(\cdot)$ denotes the Gaussian kernel density function, as recommended by Yao et al. [26]. For the penalty function, we adopt the smoothly clipped absolute deviation (SCAD) penalty, whose first derivative is given by

$$p'_\lambda(t) = \lambda\Big\{I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\Big\},$$

where $a = 3.7$ and λ is a positive tuning parameter.

However, in the presence of missing data, Eqs (2.3) and (2.4) cannot be directly applied since the response Y_i cannot be observed completely. Let δ_i be a binary indicator variable for Y_i, where $\delta_i = 1$ if Y_i is observed and $\delta_i = 0$ otherwise. We define the propensity score function as $\pi(\boldsymbol{x}_i, Y_i) =$

$P(\delta_i = 1|\boldsymbol{x}_i, Y_i), i = 1, \cdots, n$. Under the MNAR (missing nonignorable at random) assumption, the propensity score $\pi(\boldsymbol{x}_i, Y_i)$ depends not only on the observed covariates $\boldsymbol{x}_i$, but also on the Y_i itself, regardless of whether Y_i is observed or missing. When the respondent probability $\pi(\boldsymbol{x}_i, Y_i)$ is known, we have the following moment condition: $E(\delta_i \pi^{-1}(\boldsymbol{x}_i, Y_i) - 1|\boldsymbol{x}_i, Y_i) = 0$. This motivates an inverse probability weighted version of (2.3), that is,

$$\hat{\rho}^*_{Ij} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\pi(\boldsymbol{x}_i, Y_i)} x_{ij} Y_i, j = 1, \cdots, p. \tag{2.5}$$

Similarly, we can modify the objective function in (2.4) through inverse probability weighting:

$$Q_I(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} \frac{\delta_i}{\pi(\boldsymbol{x}_i, Y_i)} \phi_h(Y_i - \boldsymbol{x}_i^{*T}\boldsymbol{\beta}^*) - n\sum_{k=1}^{d} p_{\lambda_k}(|\beta_k|). \tag{2.6}$$

However, in practice, the propensity score function $\pi(\boldsymbol{x}_i, Y_i)$ is typically unknown and is often replaced by its estimator. The detailed estimation procedure will be presented in the following section.

2.2. Two-step estimation procedure for the propensity score model

In this subsection, we present a two-step estimation procedure for the propensity score model. Under nonignorable missing data assumptions, it is usually assumed that only a few predictor variables are related to the propensity score model. The first stage identifies important features in the sparse propensity model, while the second stage obtains consistent estimation using the reduced model. Following Shao and Wang [23], we assume that the propensity score function has the following semiparametric structure:

$$\pi(\boldsymbol{x}_i, Y_i) = \frac{1}{1 + \psi(\boldsymbol{x}_i) q(Y_i, \boldsymbol{\zeta})}, \tag{2.7}$$

where $q(Y_i, \boldsymbol{\zeta})$ is a known function with unknown parameter $\boldsymbol{\zeta}$, and $\psi(\cdot)$ is an unknown function. We further assume that only a small number of predictors contribute to $\pi(\boldsymbol{x}_i, Y_i)$. Specifically, let $\pi(\boldsymbol{x}_i, Y_i) = P(\delta_i = 1|\boldsymbol{x}_{i(\mathcal{R})}, Y_i) = \pi(\boldsymbol{x}_{i(\mathcal{R})}, Y_i)$ where $\mathcal{R} = \{j : P(\delta = 1|\boldsymbol{x}, Y)$ dependent on $\boldsymbol{x}_j\}$ is the index set of relevant features, and $\boldsymbol{x}_{i(\mathcal{R})}$ is the corresponding subvector of $\boldsymbol{x}_i$. Thus, the above defined propensity score in (2.7) is sparse and identifiable, which is simplified by

$$\pi(\boldsymbol{x}_i, Y_i) = \pi(\boldsymbol{x}_{i(\mathcal{R})}, Y_i) = \frac{1}{1 + \psi(\boldsymbol{x}_{i(\mathcal{R})}) q(Y_i, \boldsymbol{\zeta})}. \tag{2.8}$$

In practice, the active index set $\mathcal{R}$ is unknown and we give its consistent estimation in the first stage. To identify the potential active set $\mathcal{R}$, we employ the Pearson chi-square (PC)-based feature screening procedure which was proposed by Huang et al. [25] for categorical response with ultrahigh-dimensional categorical covariates. Assume the continuous variable X_j is discretized into K equal categories. For a fixed integer $K \geq 2$, let $q_{j,(k)}$ be the k/K-th percentile of $X_j, k = 1, 2, \cdots, K-1, q_{(0)} = -\infty, q_{(K)} = +\infty$, and $X_j = k$ if $X_j \in (q_{j,(k-1)}, q_{j,k}]$. Define $P(\delta_i = r) = P_r, P(X_{ij} = k) = P_{jk}$, and $P(\delta_i = r, X_{ij} = k) = P_{j,rk}$ for $r = 0, 1$. The PC test statistic measures the correlation between the indicator variable δ and the covariate X_j:

$$\Theta_j = \sum_{k=1}^{K}\sum_{r=0}^{1} \frac{(P_r P_{jk} - P_{j,rk})^2}{P_r P_{jk}}.$$

This statistic can be estimated by $\hat{\Theta}_j = \sum_{k=1}^{K}\sum_{r=0}^{1}(\hat{P}_r\hat{P}_{jk} - \hat{P}_{j,rk})^2/\hat{P}_r\hat{P}_{jk}$ with $\hat{P}_r = n^{-1}\sum_{i=1}^{n} I(\delta_i = r)$, $\hat{P}_{jk} = n^{-1}\sum_{i=1}^{n} I(X_{ij} = k)$, and $\hat{P}_{j,rk} = n^{-1}\sum_{i=1}^{n} I(\delta_i = r)I(X_{ij} = k)$. We then select the model

$$\hat{\mathcal{R}} = \{j : \hat{\Theta}_j \text{ is among the top } d^* \text{ largest of all}\},$$

where d^* is a threshold determined by a data-driven method. Following arguments similar to those in Huang et al. [25], we establish the strong screening consistency $P(\hat{\mathcal{R}} = \mathcal{R}) \to 1$.

To obtain a consistent estimator of the propensity score in the second stage, we can express Eq (2.8) as

$$\pi(\boldsymbol{x}_{i(\hat{\mathcal{R}})}, Y_i) = \frac{1}{1 + \psi(\boldsymbol{x}_{i(\hat{\mathcal{R}})})q(Y_i, \boldsymbol{\zeta})}. \tag{2.9}$$

Following the approach of Wang et al. [22], for a given $\boldsymbol{\zeta}$, a nonparametric kernel estimator of $\psi_{\zeta}(\boldsymbol{x}_{(\hat{\mathcal{R}})})$ is given by

$$\hat{\psi}_{\boldsymbol{\zeta}}(\boldsymbol{x}_{(\hat{\mathcal{R}})}) = \frac{\sum_{i=1}^{n}(1-\delta_i)L_b(\boldsymbol{x}_{(\hat{\mathcal{R}})} - \boldsymbol{x}_{i(\hat{\mathcal{R}})})}{\sum_{i=1}^{n}\delta_i q(Y_i, \boldsymbol{\zeta})L_b(\boldsymbol{x}_{(\hat{\mathcal{R}})} - \boldsymbol{x}_{i(\hat{\mathcal{R}})})}, \tag{2.10}$$

where $L_b(\cdot) = b^{-1}L(\cdot/b)$, $L(\cdot)$ is a kernel function and b is the bandwidth. While $\hat{\psi}_{\boldsymbol{\zeta}}(\boldsymbol{x}_{(\hat{\mathcal{R}})})$ is not directly applicable due to the unknown parameter $\boldsymbol{\zeta}$, it serves as a crucial component in the subsequent estimating equations for $\boldsymbol{\zeta}$. We define

$$f(\boldsymbol{x}_{(\hat{\mathcal{R}})}, Y, \delta, \psi_{\zeta}, \boldsymbol{\zeta}) = \Big\{\frac{\delta}{\pi(\boldsymbol{x}_{(\hat{\mathcal{R}})}, Y, \psi_{\zeta}, \boldsymbol{\zeta})} - 1\Big\}h(S), \tag{2.11}$$

where $\pi(\boldsymbol{x}_{(\hat{\mathcal{R}})}, Y, \psi_{\zeta}, \boldsymbol{\zeta}) = [1 + \psi(\boldsymbol{x}_{(\hat{\mathcal{R}})})q(Y, \boldsymbol{\zeta})]^{-1}$ and $h(S) = (1, \boldsymbol{x}_{(\hat{\mathcal{R}}^c)})^T$, $\boldsymbol{x}_{(\hat{\mathcal{R}}^c)}$ represent elements of $\boldsymbol{x}$ corresponding to the index set $\hat{\mathcal{R}}^c$. Here, $\boldsymbol{x}_{(\hat{\mathcal{R}}^c)}$ is excluded from the nonresponse propensity, which is used to create more estimation equations for estimating the propensity and ensures that the propensity is identifiable. Such $\boldsymbol{x}_{(\hat{\mathcal{R}}^c)}$ is referred as the nonresponse instrument variable (Wang et al. [22]).

We apply the generalized method of moments (GMM) approach to get the estimator of parameter $\boldsymbol{\zeta}$. Let $F_n(\hat{\psi}_{\zeta}, \boldsymbol{\zeta}) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_{i(\hat{\mathcal{R}})}, Y_i, \delta_i, \hat{\psi}_{\zeta}, \boldsymbol{\zeta})$ and $\hat{\boldsymbol{\zeta}}^{(1)} = \arg\min_{\boldsymbol{\zeta}} F_n(\hat{\psi}_{\zeta}, \boldsymbol{\zeta})^T F_n(\hat{\psi}_{\zeta}, \boldsymbol{\zeta})$. The efficient GMM estimator of $\boldsymbol{\zeta}$ is

$$\hat{\boldsymbol{\zeta}} = \arg\min_{\boldsymbol{\zeta}} {F_n}^T(\hat{\psi}_{\zeta}, \boldsymbol{\zeta})\hat{W}_n^{-1}F_n(\hat{\psi}_{\zeta}, \boldsymbol{\zeta}), \tag{2.12}$$

where $\hat{W}_n = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_{i(\hat{\mathcal{R}})}, Y_i, \delta_i, \hat{\psi}_{\hat{\zeta}^{(1)}}, \hat{\boldsymbol{\zeta}}^{(1)})f^T(\boldsymbol{x}_{i(\hat{\mathcal{R}})}, Y_i, \delta_i, \hat{\psi}_{\hat{\zeta}^{(1)}}, \hat{\boldsymbol{\zeta}}^{(1)})$.

The propensity score model specified in Eq (2.9) can be consistently estimated by

$$\hat{\pi}(\boldsymbol{x}_{i(\hat{\mathcal{R}})}, Y_i) = \frac{1}{1 + \hat{\psi}_{\hat{\zeta}}(\boldsymbol{x}_{i(\hat{\mathcal{R}})})q(Y_i, \hat{\boldsymbol{\zeta}})}. \tag{2.13}$$

With this consistent propensity score estimator, we can compute the inverse probability weighted Pearson correlation coefficient defined in (2.5) as

$$\hat{\rho}_{Ij} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}(\boldsymbol{x}_{i(\hat{\mathcal{R}})}, Y_i)}x_{ij}Y_i,\ j = 1, \cdots, p. \tag{2.14}$$

The index set of active predictors is $\hat{A} = \{j : |\hat{\rho}_{Ij}| \text{ is among the top } d \text{ largest of all}\}$. Following the suggestion of Fan and Lv [6], we set $d = [n/\log(n)]$ in our simulation studies.

Finally, the modal estimator $\hat{\boldsymbol{\beta}}^*$ is obtained by maximizing the penalized objective function

$$\hat{Q}_I(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} \hat{\Delta}_i \phi_h(Y_i - \boldsymbol{x}^{*T}\boldsymbol{\beta}^*) - n\sum_{k=1}^{d} p_{\lambda_k}(|\beta_k^*|), \tag{2.15}$$

where $\hat{\Delta}_i = \delta_i/\hat{\pi}(\boldsymbol{x}_{i(\hat{\mathcal{R}})}, Y_i)$ with $\hat{\pi}(\boldsymbol{x}_{i(\hat{\mathcal{R}})}, Y_i)$ as defined in (2.13).

3. Theoretical properties

In this section, we establish the asymptotic properties of the proposed estimators. For the convenience of expression, let $\pi_i = \pi(\boldsymbol{x}_i, Y_i)$ denote the propensity score, and let $\hat{\pi}_i$ be its corresponding estimator defined in (2.13). Denote the true value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_0$ in model (2.1) with the nonzero element β_{0j}, $j \in A_0$ and $\beta_{0j} = 0$, $j \in A_0^c$. Let $F(\boldsymbol{x}, h) = E(\phi_h''(\varepsilon)|\boldsymbol{x})$ and $G(\boldsymbol{x}, h) = E(\phi_h'(\varepsilon)^2|\boldsymbol{x})$. Denote $a_n = \max_k\{|p'_{\lambda_k}(|\beta_{0k}|)| : \beta_{0k} \neq 0\}$ and $b_n = \max_k\{|p''_{\lambda_k}(|\beta_{0k}|)| : \beta_{0k} \neq 0\}$. The following assumptions are required for establishing the main theoretical results.

A1: There exits a positive constant s_0 such that for all $0 < s < 2s_0$, $\sup_p \max_{1\le k\le p} E\{\exp(sX_k^2)\} < \infty$ and $E\{\exp(sY^2)\} < \infty$.

A2: Assume $\Theta_j = 0$ for any $j \in \mathcal{R}^c$. We further assume that there exits a positive constant $\omega_{\min} > 0$, such that $\min_{j\in\mathcal{R}} \max_{rk}(\omega_j^{rk})^2 > \omega_{\min}$, where $\omega_j^{rk} = \text{cov}\{I(\delta_i = r), I(X_{ij} = k)\}$.

A3: The estimated propensity score function satisfies $\sup_{\mathcal{R}} \|\hat{\pi}_i - \pi_i\| \le O_p(\xi_n)$, where $\xi_n = n^{-\alpha}$ and $\alpha > 0$.

A4: Suppose that $\boldsymbol{\zeta}_0$ is the unique solution to $E\{f(\boldsymbol{x}_{(\hat{\mathcal{R}})}, Y, \delta, \psi_\zeta, \boldsymbol{\zeta})\} = 0$, $\sup_{\boldsymbol{\zeta}} \|E\{f(\boldsymbol{x}_{(\hat{\mathcal{R}})}, Y, \delta, \psi_\zeta, \boldsymbol{\zeta})\}\| < \infty$. $W = E\{f(\boldsymbol{x}_{(\hat{\mathcal{R}})}, Y, \delta, \psi_\zeta, \boldsymbol{\zeta})^{\otimes 2}\}$ is positive definite.

A5: $F(\boldsymbol{x}, h)$ and $G(\boldsymbol{x}, h)$ are continuous with respect to $\boldsymbol{x}$. In addition, $F(\boldsymbol{x}, h) < 0$ for any $h > 0$.

A6: ε satisfies that $E(\phi_h'(\varepsilon)|\boldsymbol{x}) = 0$, and $E(\phi_h''(\varepsilon)^2|\boldsymbol{x})$, $E(\phi_h'(\varepsilon)^3|\boldsymbol{x})$ and $E(\phi_h'''(\varepsilon)|\boldsymbol{x})$ are continuous with respect to $\boldsymbol{x}$.

A7: $\liminf_{n\to\infty} \liminf_{\beta_k\to 0^+} p'_{\lambda_k}(|\beta_k|)/\lambda_k > 0, k \in A_0^c$.

The aforementioned assumptions are relatively mild and standard in the literature. Specifically, Assumption A1 is commonly adopted in ultrahigh-dimensional data analysis to enable theoretical derivations, as seen in Fan and Lv [6]. Assumption A2 ensures the screening consistency property, $P(\hat{\mathcal{R}} = \mathcal{R}) \to 1$ as $n \to \infty$. Assumption A3 gives an upper bound for the estimated propensity score function. Assumption A4 provides the regularity conditions required to establish the asymptotic properties of the GMM estimator $\hat{\boldsymbol{\zeta}}$. Assumptions A5 and A6 represent standard conditions in modal regression analysis, similar to those employed by Yao et al. [26]. Assumption A7 specifies common requirements for the penalty function, following the framework of Fan and Li [2].

Theorem 1. *Suppose that the Assumptions A1–A7 hold. As $n \to \infty$, we have*

$$P(A_0 \subseteq \hat{A}) \to 1.$$

Theorem 1 establishes the sure screening properties of the above proposed feature screening procedure.

Theorem 2. *Suppose that the Assumptions A1–A7 hold. As $n \to \infty$, if $a_n \to 0$ and $b_n \to 0$, then we have*

$$\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0^*\| = O_p(n^{-\frac{r}{2r+1}} + a_n).$$

Theorem 3. *Suppose that the Assumptions A1–A7 hold. Let $\lambda_{max} = \max_k\{\lambda_k\}$ and $\lambda_{min} = \min_k\{\lambda_k\}$. If $\lambda_{max} \to 0$ and $n^{\frac{r}{2r+1}}\lambda_{min} \to \infty$ as $n \to \infty$, then with probability tending to 1, we get*

$$\hat{\beta}_j = 0, \text{ for } j \in A_0^c.$$

Theorems 2 and 3 further demonstrate that our variable selection method possesses not only consistency but also the oracle property. This implies that the estimator attains the same optimal convergence rate as would be achieved if the true sparse model structure were known a priori.

4. Algorithm

This section covers the implementation of a variable selection procedure based on the modal expectation-maximization (MEM) algorithm developed by Li et al. [27], as well as the selection of bandwidth and tuning parameters.

4.1. Selection of optimal bandwidth

To obtain the modal regression estimators of parameter $\boldsymbol{\beta}^*$, the selection of an appropriate bandwidth h is crucial as it directly affects the robustness of the estimators. Following the approach of Yao et al. [26], we define the efficiency ratio of our proposed estimator relative to the least squares estimator as:

$$\hat{R}(h) = \frac{\hat{G}(h)}{\hat{F}(h)^2\hat{\sigma}^2}, \tag{4.1}$$

where $\hat{F}(h) = \frac{1}{n}\sum_{i=1}^{n}\phi_h''(\hat{\varepsilon}_i)$, $\hat{G}(h) = \frac{1}{n}\sum_{i=1}^{n}\phi_h'(\hat{\varepsilon}_i)^2$, $\hat{\varepsilon}_i = \hat{\Delta}_i[Y_i - \boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}}^{\text{int}}]$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$, and $\hat{\boldsymbol{\beta}}^{\text{int}}$ represents the initial least squares estimate of $\boldsymbol{\beta}^*$ via the least squares estimation method.

The optimal bandwidth h is calculated by minimizing $R(h)$ defined in (4.1), and therefore,

$$h_{opt} = \text{argmin}_h\hat{R}(h). \tag{4.2}$$

Following the methodology of Yao et al. [26], we employ a grid search approach to select the optimal bandwidth h. The candidate grid points are constructed as $h = 0.5\hat{\sigma} \times 1.02^j$, $j = 0, 1, \ldots, 100$.

4.2. Selection of tuning parameters

The choice of tuning parameters is indispensable to implement the variable selection in practice. Bayesian information criterion (BIC) is a common strategy which is easy to implement. To facilitate the calculation, define

$$\lambda_k = \frac{\lambda}{|\hat{\beta}_k^{(0)}|}, k = 1, \ldots, d, \tag{4.3}$$

where $\hat{\beta}_k^{(0)}$ denotes the initial non-penalized estimator for the k-th coefficient.

Specifically, we choose λ by minimizing the following BIC objective function,

$$BIC(\lambda) = \sum_{i=1}^{n} \hat{\Delta}_i \phi_h[Y_i - \boldsymbol{x}^{*T}\boldsymbol{\beta}^*(\lambda)] - df(\boldsymbol{\beta}^*(\lambda))n\log(n), \tag{4.4}$$

where $df(\boldsymbol{\beta}^*(\lambda))$ is the degree of freedom of $\boldsymbol{\beta}^*(\lambda)$, defined as the number of nonzero components in $\boldsymbol{\beta}^*(\lambda)$.

4.3. The MEM algorithm for variable selection

To maximize the objective function (2.15), we employ the MEM algorithm. The optimization presents two key challenges: the SCAD penalty function is singular at the origin and the objective function $\hat{Q}_I(\boldsymbol{\beta}^*)$ is nonconvex, making direct application of the Newton-Raphson algorithm infeasible. Following Fan and Li [2], we address these challenges through a local quadratic approximation approach. For initial estimator $\beta_k^{(0)}, k = 1, \ldots, d$, we set $\hat{\beta}_k = 0$ if $\hat{\beta}_k$ is close to 0. Otherwise, we approximate the SCAD penalty function in their neighborhood:

$$p_{\lambda_k}(|\beta_k|) \approx p_{\lambda_k}(|\beta_k^{(0)}|) + \frac{1}{2}\frac{p'_{\lambda_k}(|\beta_k^{(0)}|)}{|\beta_k^{(0)}|}\Big(|\beta_k|^2 - |\beta_k^{(0)}|^2\Big).$$

Define the penalty matrix at iteration $m = 0$ as:

$$\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(m)}) = \text{diag}\Big\{\frac{p'_{\lambda_1}(|\beta_1^{(m)}|)}{|\beta_1^{(m)}|}, \frac{p'_{\lambda_2}(|\beta_2^{(m)}|)}{|\beta_2^{(m)}|}, \ldots, \frac{p'_{\lambda_d}(|\beta_d^{(m)}|)}{|\beta_d^{(m)}|}\Big\}.$$

The MEM algorithm proceeds as follows:

Step 1 (**E step**): We calculate weights $\pi(i\,|\boldsymbol{\beta}^{(m)})$ by

$$\pi(i|\boldsymbol{\beta}^{(m)}) = \frac{\hat{\Delta}_i \phi_h(Y_i - \boldsymbol{x}^{*T}\boldsymbol{\beta}^{(m)})}{\sum_{i=1}^{n} \hat{\Delta}_i \phi_h(Y_i - \boldsymbol{x}^{*T}\boldsymbol{\beta}^{(m)})}, i = 1, \ldots, n.$$

Step 2 (**M step**): Update the parameter estimates:

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \Big\{\pi(i|\boldsymbol{\beta}^{(m)})\hat{\Delta}_i \log \phi_h[Y_i - \boldsymbol{x}^{*T}\boldsymbol{\beta}]\Big\} - \frac{n}{2}\boldsymbol{\beta}^T\Sigma_\lambda(\boldsymbol{\beta}^{(m)})\boldsymbol{\beta}$$

with the closed-form solution:

$$\hat{\boldsymbol{\beta}}^{(m+1)} = (\check{\boldsymbol{X}}^T\boldsymbol{\omega}\check{\boldsymbol{X}} + n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(m)}))^{-1}\check{\boldsymbol{X}}^T\boldsymbol{\omega}\check{\boldsymbol{Y}}$$

where $\boldsymbol{\omega}$ is an $n \times n$ diagonal matrix and $\boldsymbol{\omega} = \text{diag}(\pi(1|\boldsymbol{\beta}^{(m)}), \pi(2|\boldsymbol{\beta}^{(m)}), \ldots, \pi(n|\boldsymbol{\beta}^{(m)}))$, $\check{\boldsymbol{X}} = \boldsymbol{\Delta}\boldsymbol{X}, \check{\boldsymbol{Y}} = \Delta\boldsymbol{Y}, \boldsymbol{\Delta} = \text{diag}(\hat{\Delta}_1, \hat{\Delta}_2, \ldots, \hat{\Delta}_n)$.

Step 3: Iterate between E and M steps until convergence. In practice, we stop the iteration if $\|\hat{\boldsymbol{\beta}}^{(m+1)} - \hat{\boldsymbol{\beta}}^{(m)}\| < 10^{-3}$. The final estimate is denoted by $\hat{\boldsymbol{\beta}}^*$.

5. Numerical studies

To demonstrate the performance of the proposed variable selection procedure, we conduct the following simulation studies. We generate data from the model

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, i = 1, \ldots, n, \tag{5.1}$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ follows a multivariate normal distribution with mean zero and covariance matrix $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$. The p-dimensional parameter vector is set to $\boldsymbol{\beta} = (5, 4, 0, \cdots, 0, 5, 5)^T$. To assess the robustness of the proposed method, we consider three different error distributions: (1) normal distribution: $\varepsilon \sim N(0, 1)$; (2) t distribution: $\varepsilon \sim t(3)$; (3) mixed normal distribution (MN): $\varepsilon \sim 0.9N(0, 1) + 0.1N(0, 9^2)$. Additionally, we generate missing indicators δ_i from a Bernoulli distribution with probability $\pi(Y_i, \boldsymbol{x}_i) = \{1 + \psi(\boldsymbol{x}_i) \exp(\zeta Y_i)\}^{-1}$ where $\psi(\cdot)$ and ζ are specified in four different cases:

Case 1: $\psi(\boldsymbol{x}_i) = \exp(1.3 + 0.8x_{i2} + 0.4x_{i3})$ and $\zeta = 0$;

Case 2: $\psi(\boldsymbol{x}_i) = \exp(1.3 + 0.8x_{i2} + 0.4x_{i3})$ and $\zeta = -0.2$;

Case 3: $\psi(\boldsymbol{x}_i) = \exp(1.8 - 0.6\sin(x_{i2}) - 0.4\sin(x_{i3}))$ and $\zeta = -0.1$;

Case 4: $\psi(\boldsymbol{x}_i) = \exp(0.7 + 0.5\exp(x_{i2}) + 0.3\exp(x_{i3}))$ and $\zeta = -0.3$.

Here, Case 1 corresponds to a random missing mechanism, while Cases 2–4 represent different non-ignorable missing mechanisms. The coefficients are chosen such that the average missing proportion falls between 30% and 40%. For each missing mechanism, we set the sample size $n = 100$ and consider dimensions $p = 500$ and 1000. The simulation is repeated 200 times for each configuration.

To examine the effect of varying slice numbers in the PC statistics on feature selection for the propensity score model, we evaluate the screening performance using the criteria listed in Table 1. Here, "PC-N" denotes the number of slices (N), which we set to 6, 8, and 10, while the true dimension is fixed at $d^* = 2$. The evaluation criteria are defined as follows: (1) CF (correct feature selection rate): The proportion of simulation runs where all truly significant features in the propensity score model were correctly identified. (2) UF (under-fitting rate): The proportion of runs where at least one relevant feature was missed. (3) OF (over-fitting rate): The proportion of runs where irrelevant features were incorrectly selected as significant.

Therefore, the following deductions can be made: (1) The proposed PC method demonstrates excellent performance, achieving nearly 100% correct feature selection (CF) while maintaining consistently low under-fitting (UF) and over-fitting (OF) rates across all scenarios. (2) The feature screening performance of the PC method remains robust, showing no significant variations when applied to different error distributions. (3) The method maintains its effectiveness even in high-dimensional settings ($p =$ 500, 1000) and under MAR missing mechanisms (Case 1). Our results indicate that the number of slices has minimal impact on the feature screening performance. Based on these findings, we employ PC-10 in subsequent simulation studies.

Upon identifying the important features $\boldsymbol{x}_{i(\hat{\mathcal{R}})}$, we proceed to estimate the unknown parameters in the propensity score model using the GMM. Following Shao and Wang [23], we employ a Gaussian kernel function $L(\cdot) = \frac{1}{\sqrt{2\pi}}\exp(-(\cdot)^2/2)$ for the nonparametric kernel estimation with the product kernel $L(u_1, u_2) = L(u_1)L(u_2)$. Adopting the same strategy as in Shao and Wang [23], the bandwidth parameters are selected as $b_k = 1.5\hat{\sigma}_{u_k} n^{-1/3}, \quad k = 1, 2$, where $\hat{\sigma}_{u_1}$ and $\hat{\sigma}_{u_2}$ represent the sample standard deviations of u_{ij} $(i = 1, ..., n;\ j = 1, 2)$, with (u_{i1}, u_{i2}) corresponding to the selected important variables $\boldsymbol{x}_{i(\hat{\mathcal{R}})}$. For

Table 1. Feature selection results of PC methods with different numbers of slices.

ε	Method		p = 500			p = 1000		
			CF	OF	UF	CF	OF	UF
		Case 1	0.90	0.05	0.05	0.89	0.05	0.06
	PC-6	Case 2	0.89	0.06	0.05	0.88	0.06	0.06
		Case 3	0.89	0.05	0.06	0.88	0.05	0.07
		Case 4	0.89	0.04	0.07	0.89	0.06	0.05
		Case 1	0.90	0.03	0.07	0.89	0.05	0.06
$N(0,1)$	PC-8	Case 2	0.90	0.05	0.05	0.89	0.04	0.07
		Case 3	0.91	0.02	0.07	0.88	0.04	0.08
		Case 4	0.90	0.03	0.07	0.89	0.05	0.06
		Case 1	0.95	0.00	0.05	0.90	0.04	0.06
	PC-10	Case 2	0.96	0.01	0.01	0.90	0.04	0.06
		Case 3	0.93	0.00	0.07	0.91	0.03	0.06
		Case 4	0.92	0.00	0.08	0.91	0.02	0.07
		Case 1	0.90	0.03	0.07	0.89	0.05	0.06
	PC-6	Case 2	0.90	0.04	0.06	0.88	0.06	0.06
		Case 3	0.89	0.05	0.06	0.89	0.06	0.05
		Case 4	0.89	0.04	0.07	0.89	0.04	0.07
		Case 1	0.90	0.02	0.08	0.89	0.05	0.06
$t(3)$	PC-8	Case 2	0.91	0.02	0.07	0.90	0.05	0.05
		Case 3	0.92	0.04	0.04	0.90	0.04	0.06
		Case 4	0.92	0.02	0.06	0.89	0.05	0.06
		Case 1	0.95	0.00	0.05	0.91	0.03	0.06
	PC-10	Case 2	0.94	0.01	0.05	0.92	0.02	0.06
		Case 3	0.95	0.01	0.04	0.93	0.02	0.05
		Case 4	0.94	0.00	0.06	0.90	0.03	0.06
		Case 1	0.90	0.02	0.08	0.89	0.04	0.07
	PC-6	Case 2	0.92	0.03	0.05	0.88	0.05	0.07
		Case 3	0.93	0.03	0.04	0.87	0.04	0.09
		Case 4	0.93	0.04	0.03	0.89	0.05	0.06
		Case 1	0.93	0.01	0.06	0.89	0.04	0.07
MN	PC-8	Case 2	0.95	0.00	0.05	0.90	0.05	0.05
		Case 3	0.94	0.01	0.05	0.89	0.04	0.07
		Case 4	0.94	0.00	0.05	0.90	0.04	0.06
		Case 1	0.95	0.01	0.04	0.93	0.02	0.05
	PC-10	Case 2	0.96	0.00	0.04	0.91	0.03	0.06
		Case 3	0.94	0.00	0.06	0.91	0.04	0.05
		Case 4	0.95	0.01	0.04	0.92	0.02	0.06

instrumental variable selection, we rank the estimates $\hat{\Theta}_j$ and pick the last $[n/\log(n)/3]$ variables as instrumental variables $\boldsymbol{x}(\hat{\mathcal{R}}^c)$ in Eq (2.11). These variables are chosen specifically for their minimal influence on the propensity score function.

To assess the performance of our proposed variable selection method, we conduct comprehensive comparative studies with the following estimators. Our proposed estimator, denoted as MNAR-SMR, is obtained by maximizing (2.15) using SCAD-penalized modal regression under MNAR conditions. A counterpart estimator, denoted as MAR-SMR, employs SCAD-penalized modal regression under MAR (missing at random) conditions, with the propensity score function estimated via maximum likelihood. Meanwhile, MNAR-SPLS and MAR-SPLS are compared where modal regression is replaced by ordinary least squares under both missing data mechanisms. The simulation results, presented in Tables 2–4, are organized by different error distributions. We evaluate performance using three key metrics. "Size" denotes the average number of selected non-zero coefficients across 200 simulations. "CIP" (correctly identified proportion) indicates the proportion of truly important variables (X_1, X_2, X_{p-1}, X_p) that are correctly identified by the method. "Bias" is the mean absolute deviation of all coefficient

estimators, which is defined as

$$\text{Bias} = \sum_{j=1}^{p} \left| 200^{-1} \sum_{m=1}^{200} (\hat{\beta}_j^m - \beta_{0j}) \right|,$$

where $\hat{\beta}_j^m$ represents the estimator for β_j in the mth simulation, and β_{0j} denotes the true parameter value.

Table 2. The simulation results of four methods with model error $\varepsilon \sim N(0, 1)$.

Method		p = 500				p = 1000			
		Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
MNAR-SMR	Size	4.32	4.49	4.56	4.45	4.59	4.87	5.11	4.95
	CIP	98%	98%	100%	100%	96%	97%	97%	98%
	Bias	0.1691	0.2135	0.2562	0.2284	0.2479	0.2992	0.3452	0.3146
MAR-SMR	Size	4.33	4.94	5.12	5.08	4.44	4.98	5.23	5.18
	CIP	97%	93%	90%	92.5%	94%	92%	90%	89%
	Bias	0.1628	0.3257	0.3916	0.3885	0.2351	0.3578	0.3961	0.3938
MNAR-SPLS	Size	4.30	4.45	4.52	4.51	4.52	4.89	5.05	4.99
	CIP	98%	99%	100%	100%	96%	96.5%	97%	98%
	Bias	0.1699	0.2130	0.2603	0.2081	0.2388	0.2981	0.3328	0.3098
MAR-SPLS	Size	4.30	4.88	5.05	5.10	4.50	5.01	5.25	5.15
	CIP	97%	94%	90%	92%	95%	93%	91%	90%
	Bias	0.1774	0.3197	0.3828	0.3803	0.2205	0.3482	0.3881	0.3819

Table 3. The simulation results of four methods with model error $\varepsilon \sim t(3)$.

Method		p = 500				p = 1000			
		Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
MNAR-SMR	Size	4.27	4.35	4.58	4.25	5.05	5.21	4.98	4.86
	CIP	100%	99%	99%	100%	96%	97%	96%	97%
	Bias	0.2492	0.2376	0.2432	0.1393	0.3483	0.3629	0.2949	0.3554
MAR-SMR	Size	4.25	5.02	5.21	4.97	5.03	5.56	5.45	5.38
	CIP	98%	92%	91%	92%	95%	90%	89%	90.5%
	Bias	0.2452	0.3370	0.4062	0.3378	0.3789	0.4040	0.4081	0.3995
MNAR-SPLS	Size	4.36	4.45	4.65	4.76	4.99	5.28	5.05	5.01
	CIP	91%	90.5%	93%	92%	90%	90%	92%	89.5%
	Bias	0.3198	0.3521	0.3487	0.3696	0.3829	0.3978	0.3806	0.3893
MAR-SPLS	Size	4.35	4.50	4.71	4.75	4.97	5.09	5.15	4.99
	CIP	91%	90%	89%	90.5%	89.5%	89%	90%	88%
	Bias	0.3501	0.4105	0.3987	0.3952	0.4099	0.4008	0.3902	0.3909

Tables 2–4 allow for the following conclusions to be made.

(1) Under most simulated scenarios, including Case 1, the proposed MNAR-SMR method achieves higher correctly identified proportion (CIP) in terms of select active variables and lower bias in terms of parameter estimation, compared to alternative approaches. Additionally, it correctly identifies all significant variables with an accuracy exceeding 95%, demonstrating its effectiveness. As expected, the MAR-SMR method performs well under Case 1 (MAR mechanism), but its estimates exhibit substantial deviation under nonignorable missingness (Cases 2–4), confirming the appropriateness of the propensity score model specification.

Table 4. The simulation results of four methods with model error $\varepsilon \sim MN$.

Method		$p = 500$				$p = 1000$			
		Case 1	Case 2	Case 3	Case 4	Case 1	Case 2	Case 3	Case 4
MNAR-SMR	Size	4.32	4.28	4.41	4.50	4.48	5.34	5.27	4.75
	CIP	98%	100%	100%	99%	96%	97%	96%	96%
	Bias	0.2642	0.1982	0.2431	0.2393	0.2482	0.3418	0.3494	0.3242
MAR-SMR	Size	4.34	5.02	5.24	5.17	4.45	5.29	5.25	5.22
	CIP	96%	90%	89%	91%	94.5%	89%	88%	89.5%
	Bias	0.2552	0.3555	0.4062	0.3780	0.2497	0.3740	0.4281	0.4095
MNAR-SPLS	Size	4.45	4.31	4.50	4.59	4.76	5.25	5.09	4.98
	CIP	90%	91%	89%	90%	89%	89.5%	88%	89%
	Bias	0.3208	0.3805	0.4120	0.3988	0.3088	0.3998	0.4027	0.4103
MAR-SPLS	Size	4.31	4.43	4.67	4.58	4.50	5.08	5.13	5.19
	CIP	94	88%	89.5%	89%	89%	88%	87%	89%
	Bias	0.2991	0.3989	0.4003	0.4015	0.3021	0.4051	0.4431	0.4238

(2) When comparing the modal regression-based (MR) approach with the classical least squares (LS) method, we observe that under normal errors, the LS approach yields slightly higher CIP values and better variable selection accuracy. However, when the error follows t distribution or mixed normal distribution, the MR approach provides CIP values much closer to 100%, significantly outperforming LS. This highlights the robustness of the MR method in handling heavy-tailed errors or outliers.

(3) As the covariate dimension increases, estimation bias tends to rise under the same missingness ratio. Nevertheless, across most simulation scenarios, the proposed MNAR-SMR method remains reliable, delivering consistent and accurate results even in higher-dimensional settings.

6. A real data analysis

In this section, we demonstrate the application of our proposed method using a Diffuse Large B-cell Lymphoma (DLBCL) dataset obtained from The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/). In this dataset, the response variable is the survival time (log-transformed, following Zhou and Zhu [28]) of 50 patients, which includes 38 deaths during the follow-ups. The covariates are high-dimensional gene expression profiles (5299 genes) measured via microarray technology. To demonstrate our proposed method using this dataset, we consider the censored responses as missing observations, with approximately 24% of the response data being unobserved. Given that the unobserved survival times of the patients seem to be related to the patients themselves, which means that the censoring probability appears to be associated with the latent true survival times, the missingness mechanism satisfies the conditions for nonignorable missing data. This aligns with our methodological focus on MNAR scenarios.

For the DLBCL dataset, we selected a final set of $d = [n/\log(n)] = 12$ genes through our variable selection procedure. To evaluate our proposed MNAR-SMR method, we conducted comparative analysis with three alternative approaches: the MAR-SMR, MNAR-SPLS, and MAR-SPLS. To ensure robust estimation of the propensity score function, we partitioned the covariates into 10 distinct slices. Table 5 presents the number of genes selected by each of the four variable selection procedures.

To assess the prognostic significance of the 12 selected genes, we adopted an analytical approach

Table 5. Results of 12 selected gene numbers in DLBCL data with four methods.

Order	MNAR-SMR	MAR-SMR	MNAR-SPLS	MAR-SPLS
1	1716	4829	4042	3490
2	508	2997	4777	3669
3	3086	5686	2448	2403
4	4777	5213	3911	4007
5	1819	1872	2057	1191
6	4279	228	3083	5221
7	1610	1003	4279	604
8	5256	5221	5256	4341
9	1273	2711	852	450
10	3722	3669	3490	1010
11	2448	3922	3797	2448
12	5297	3146	2650	771

similar to Zhu et al. [11]. The DLBCL dataset is randomly divided into a training set ($n_1 = 35$) and a test set ($n_2 = 15$). We fitted a Cox proportional risk model using the training set and calculate risk scores for both the training and test set. Then, the patients in the test set were classified into low-risk and high-risk groups using the median risk score from the training set as the cutoff threshold. Figure 1 presents the estimated Kaplan-Meier survival curves by the four methods (MNAR-SMR, MAR-SMR, MNAR-SPLS, and MAR-SPLS) for the two risk groups of patients in the test set.

Observation of the results in Figure 1 shows that the best separation of the two curves is obtained by the MNAR-SMR approach, and the log-rank test yields a p-value of 0.05, indicating that the fitted model has a better predictive effect. While the corresponding p-values of the MAR-SMR, MNAR-SPLS, and MAR-SPLS methods are 0.56, 0.19, and 0.69. This indicates that the fitted models obtained from NMAR assumption have better prediction results than the MAR assumption.

To comprehensively assess the predictive performance of each method, we employed the average check loss prediction error (ACLPE), defined as:

$$\text{ACLPE} = \frac{1}{n_2} \sum_{i \in \mathcal{R}} \delta_i (Y_i - \hat{Y}_i), \tag{6.1}$$

where $\mathcal{R}$ represents the set of test sample indicators, δ_i is the indicator variable denoting whether the response Y_i is observed ($\delta_i = 1$) or censored ($\delta_i = 0$), $\hat{Y}_i$ is the predicted value of the response variable, and n_2 is the size of the test set. We conducted 200 simulation replicates. Figure 2 displays boxplots of the average check loss prediction error with different methods.

As shown in Figure 2, the modal regression approach yields superior prediction accuracy compared to the least squares method, as evidenced by its consistently lower prediction errors. This empirical evidence strongly supports the validity of the MNAR assumption for our dataset, suggesting that accounting for nonignorable missingness leads to more reliable predictions in this clinical context.

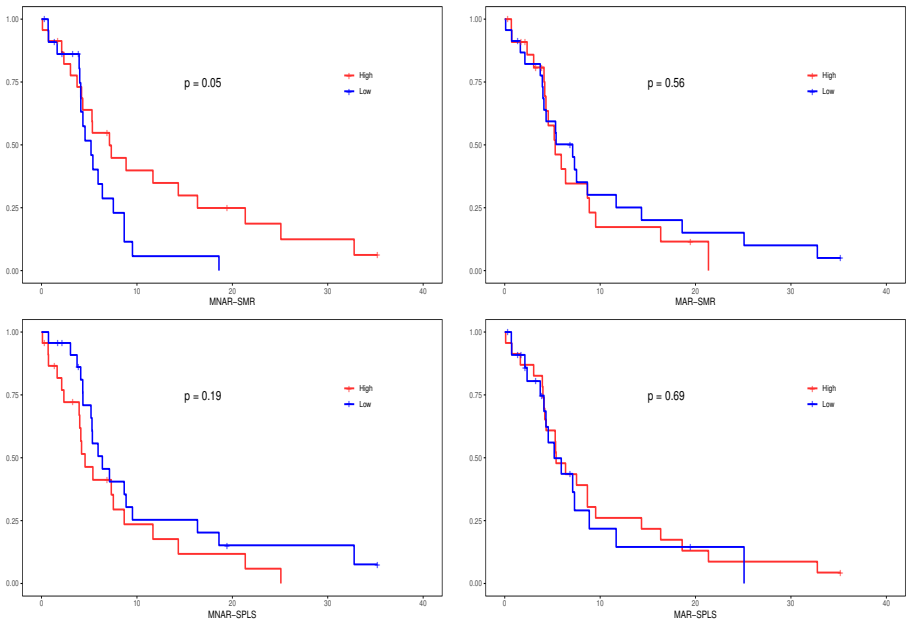

Figure 1. Survival curves for the two risk groups in the test set.

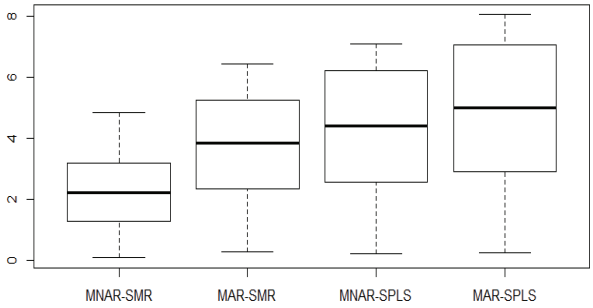

Figure 2. The boxplots of average check loss prediction error under four methods.

7. Conclusions

This paper proposes a robust variable selection method for ultrahigh-dimensional linear models with nonignorable missing responses, employing modal regression to enhance robustness. To address the identifiability issue in the semiparametric propensity score function, we develop a two-step estimation procedure. The PC statistic identifies important features in the sparse propensity score in the first step and the GMM method estimates unknown parameters for the downscaled propensity score in the second step. To overcome the challenges encountered with ultrahigh-dimensionality, we first reduce the model dimensionality using a modified sure independence screening (SIS) procedure with inverse probability weighting (IPW), and then the variable selection procedure can be effectively applied to the reduced model. The IPW technique is incorporated to mitigate bias induced by missing data. Subsequently, we perform the variable selection procedure by applying the SCAD penalty to the modal regression-

based objective function. Our approach alleviates the curse of dimensionality, ensuring scalability for high-dimensional data and maintaining robustness against heavy-tailed errors and outliers. Under mild regularity conditions, we establish the asymptotic properties of the proposed estimator. Simulation studies and an analysis of the DLBCL dataset demonstrate the method's effectiveness in finite samples, confirming its practical utility.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the Shaanxi Fundamental Science Research Project for Mathematics and Physics (No.23JSY046).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. B*, **58** (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
2. J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Stat. Assoc.*, **96** (2001), 1348–1360. https://doi.org/10.1198/016214501753382273
3. H. Zou, The adaptive lasso and its oracle properties, *J. Am. Stat. Assoc.*, **101** (2006), 1418–1429. https://doi.org/10.1198/016214506000000735
4. H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. B*, **67** (2005), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
5. C. Zhang, Nearly unbiased variable selection under minimax concave penalty, *Ann. Statist.*, **38** (2010), 894–942. https://doi.org/10.1214/09-AOS729
6. J. Fan, J. Lv, Sure independence screening for ultra-high dimensional feature space, *J. R. Stat. Soc. B*, **70** (2008), 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x
7. J. Fan, R. Song, Sure independence screening in generalized linear models with NP-dimensionality, *Ann. Statist.*, **38** (2010), 3567–3604. https://doi.org/10.1214/10-AOS798
8. J. Fan, Y. Feng, R. Song, Nonparametric independence screening in sparse ultra-high-dimensional additive models, *J. Am. Stat. Assoc.*, **106** (2011), 544–557. https://doi.org/10.1198/jasa.2011.tm09779
9. J. Fan, Y. Ma, W. Dai, Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models, *J. Am. Stat. Assoc.*, **109** (2014), 1270–1284. https://doi.org/10.1080/01621459.2013.879828

10. H. Liang, H. Wang, C. Tsai, Profiled forward regression for ultrahigh dimensional variable screening in semiparametric partially linear models, *Stat. Sin.,* **22** (2012), 531–554. https://doi.org/10.2139/ssrn.1746315
11. L. Zhu, L. Li, R. Li, L. Zhu, Model-free feature screening for ultrahigh-dimensional data, *J. Am. Stat. Assoc.*, **106** (2011), 1464–1475. https://doi.org/10.1198/jasa.2011.tm10563
12. R. Li, W. Zhong, L. Zhu, Feature screening via distance correlation learning, *J. R. Stat. Soc. B*, **107** (2012), 1129–1139. https://doi.org/10.1080/01621459.2012.695654
13. G. Li, H. Peng, J. Zhang, L. Zhu, Robust rank correlation based screening, *Ann. Statist.*, **40** (2012), 1846–1877. https://doi.org/10.1214/12-AOS1024
14. R. Little, D Rubin, *Statistical Analysis with Missing Data, 2nd* edition, New York, 2002. https://doi.org/10.1002/9781119013563
15. P. Zhao, L. Xue, Variable selection for semiparametric varying-coefficient partially linear models with missing response at random, *Acta Math. Sin. Engl. Ser.*, **27** (2011), 2205–2216. https://doi.org/10.1007/s10114-011-9200-1
16. B. Sherwood, Variable selection for additive partial linear quantile regression with missing covariates, *J. Multivar. Anal.*, **152** (2016), 206–223. https://doi.org/10.1016/j.jmva.2016.08.009
17. Y. Wang, Y. Song, Variable selection via penalized quasi-maximum likelihood method for spatial autoregressive model with missing response, *Spat. Stat.*, **59** (2024), 100809. https://doi.org/10.1016/j.spasta.2023.100809
18. P. Lai, Y. Liu, Z. Liu, Y. Wan, Model free feature screening for ultrahigh dimensional data with responses missing at random, *Comput. Stat. Data Anal.*, **105** (2017), 201–216. https://doi.org/10.1016/j.csda.2016.08.008
19. N. Tang, L. Xia, X. Yan, Feature screening in ultrahigh-dimensional partially linear models with missing responses at random, *Comput. Stat. Data Anal.*, **133** (2019), 208–227. https://doi.org/10.1016/j.csda.2018.10.003
20. X. Li, N. Tang, J. Xie, X. Yan, A nonparametric feature screening method for ultrahigh-dimensional missing response, *Comput. Stat. Data Anal.*, **142** (2020), 106828. https://doi.org/10.1016/j.csda.2019.106828
21. J. Kim, C. Yu, A semiparametric estimation of mean functionals with nonignorable missing data, *J. Am. Stat. Assoc.*, **106** (2011), 157–165. https://doi.org/10.1198/jasa.2011.tm10104
22. S. Wang, J. Shao, J. Kim, An instrumental variable approach for identification and estimation with nonignorable nonresponse, *Stat. Sin.*, **24** (2014), 1097–1116. https://doi.org/10.5705/ss.2012.074
23. J. Shao, L. Wang, Semiparametric inverse propensity weighting for nonignorable missing data, *Biometrika*, **103** (2016), 175–187. https://doi.org/10.1093/biomet/asv071
24. N. Tang, Y. Ju, Statistical inference for nonignorable missing-data problems: a selective review, *Stat. Theory Relat. Fields*, **2** (2018), 105–133. https://doi.org/10.1080/24754269.2018.1522481
25. D. Huang, R. Li, H. Wang, Feature screening for ultrahigh dimensional categorical data with applications, *J. Bus. Econ. Stat.*, **32** (2014), 237–244. https://doi.org/10.1080/07350015.2013.863158
26. W. Yao, B. Lindsay, R. Li, Local modal regression, *J. Nonparametr. Stat.*, **24** (2012), 647–663. https://doi.org/10.1080/10485252.2012.678848

27. J. Li, S. Ray, B. G. Lindsay, A nonparametric statistical approach to clustering via mode identification, *J. Mach. Learn. Res.*, **8** (2007), 1687–1723. https://doi.org/10.5555/1314498.1314541

28. T. Zhou, L. Zhu, Model-free feature screening for ultrahigh dimensional censored regression, *Stat. Comput.*, **27** (2017), 947–961. https://doi.org/10.1007/s11222-016-9664-z

Appendix: Proofs of theorems

Proof of Theorem 1. According to the absolute value inequality, for any given $\varepsilon > 0$, we have

$$P(||\hat{\rho}_{Ij}| - |\rho_{0j}|| \geq \varepsilon) \leq P(|\hat{\rho}_{Ij} - \rho_{0j}| \geq \varepsilon). \tag{A.1}$$

We consider $P(|\hat{\rho}_{Ij} - \rho_{0j}| \geq \varepsilon)$ first. To facilitate the proof, we first give the following notations. Define $\rho_{0j} = E(X_jY)$, $\hat{\rho}_{Ij} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}x_{ij}Y_i$, and $\hat{\rho}^*_{Ij} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}x_{ij}Y_i$. Therefore, when n is large enough, we have

$$P(|\hat{\rho}_{Ij} - \rho_{0j}| \geq \varepsilon) \leq P(|\hat{\rho}_{Ij} - \hat{\rho}^*_{Ij}| + |\hat{\rho}^*_{Ij} - \rho_{0j}| \geq \varepsilon) \leq P(|\hat{\rho}^*_{Ij} - \rho_{0j}| \geq \varepsilon/2). \tag{A.2}$$

We use truncation techniques to deal with $\hat{\rho}^*_{Ij}$ and ρ_{0j}, and for any $M > 0$,

$$\begin{aligned}\hat{\rho}^*_{Ij} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}x_{ij}Y_iI\Big(\Big|\frac{\delta_i}{\pi_i}x_{ij}Y_i\Big| \leq M\Big) + \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}x_{ij}Y_iI\Big(\Big|\frac{\delta_i}{\pi_i}x_{ij}Y_i\Big| > M\Big)\\ &= \hat{S}_{j1} + \hat{S}_{j2},\end{aligned}$$

and

$$\rho_{0j} = E(X_jYI(|X_jY| \leq M)) + E(X_jYI(|X_jY| > M)) = S_{j1} + S_{j2}.$$

It is easy to see that $\hat{S}_{j1}$ and $\hat{S}_{j2}$ are the unbiased estimators of S_{j1} and S_{j2}, respectively. Furthermore, define $I_{M_i} = I(|\frac{\delta_i}{\pi_i}x_{ij}Y_i| \leq M)$, $I^c_{M_i} = I(|\frac{\delta_i}{\pi_i}x_{ij}Y_i| > M)$, $I_M = I(|X_jY| \leq M)$, and $I^c_M = I(|X_jY| > M)$. Then, we have

$$\begin{aligned}P(|\hat{\rho}^*_{Ij} - \rho_{0j}| \geq \varepsilon/2) &\leq P(|\hat{S}_{j1} - S_{j1}| \geq \varepsilon/4) + P(|\hat{S}_{j2} - S_{j2}| \geq \varepsilon/4)\\ &= F_1 + F_2\end{aligned} \tag{A.3}$$

For F_1, by the exponential Markov inequality, $\forall t > 0$, let $\varepsilon_1 = \varepsilon/4$, and we can obtain that

$$\begin{aligned}P(\hat{S}_{j1} - S_{j1} \geq \varepsilon_1) &\leq E\{\exp(t(\hat{S}_{j1} - S_{j1}))\}\exp(-t\varepsilon_1)\\ &\leq \exp(-t\varepsilon_1)\exp(-tS_{j1})E\{\exp(t\hat{S}_{j1})\}\end{aligned}$$

It is noted that $E\{\exp(t\hat{S}_{j1})\} = E\{\exp(\frac{t}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}x_{ij}Y_iI_{M_i})\} = E^n\{\exp(\frac{t}{n}\frac{\delta_i}{\pi_i}x_{ij}Y_iI_{M_i})\}$. By Lemma 1 from Li et al. [12], we can get that

$$\begin{aligned}P(\hat{S}_{j1} - S_{j1} \geq \varepsilon_1) &\leq \exp(-t\varepsilon_1)E^n\Big\{\exp\Big[\frac{t}{n}\Big(\frac{\delta_i}{\pi_i}x_{ij}Y_iI_{M_i} - E(X_jYI_M)\Big)\Big]\Big\}\\ &\leq \exp\Big(-t\varepsilon_1 + \frac{t^2M^2}{2n}\Big).\end{aligned}$$

Let $t = \frac{n\varepsilon_1}{M^2}$, and then we have $P(\hat{S}_{j1} - S_{j1} \geq \varepsilon_1) \leq \exp(-\frac{n\varepsilon_1^2}{2M^2})$. Therefore,

$$F_1 = P(|\hat{S}_{j1} - S_{j1}| \geq \varepsilon_1) \leq 2\exp\Big(-\frac{n\varepsilon_1^2}{2M^2}\Big). \tag{A.4}$$

For F_2, by the Cauchy-Schwartz inequality and exponential Markov inequality, $\forall s > 0$, we have

$$\begin{aligned} S_{j2}^2 = E^2(X_jYI(|X_jY| > M)) &\leq E(X_jY)^2P(|X_jY| > M) \\ &\leq E(X_jY)^2E[\exp(s|X_jY|)]\exp(-sM) \\ &\leq E(X_jY)^2\{E[\exp(sX_j^2)]E[\exp(sY^2)]\}^{1/2}\exp(-sM) \end{aligned}$$

By Assumption A1, and for sufficient $M \to \infty$, we have $S_{j2} \leq \varepsilon_1/2$ as $n \to \infty$. Therefore,

$$\begin{aligned} F_2 &= P(|\hat{S}_{j2} - S_{j2}| \geq \varepsilon_1)) \leq P(|\hat{S}_{j2}| \geq \varepsilon_1/2) \\ &\leq nP(X_j^2 + Y^2 > M/2) \leq nP(X_j^2 > M/4) + nP(Y^2 > M/4) \end{aligned}$$

By the exponential Markov inequality and Assumption A1, we have $nP(Y^2 > M/4) \leq nE\{\exp(sY^2)\}\exp(-sM/4) \leq c_1n\exp(-sM/4)$ and $nP(X_j^2 > M/4) \leq c_1n\exp(-sM/4)$. Then, we have

$$F_2 \leq c_1n\exp(-sM/4) \tag{A.5}$$

Combining (A.3)–(A.5), we have

$$P(|\hat{\rho}_{Ij}^* - \rho_{0j}| \geq \varepsilon/2) \leq 2\exp\Big(-\frac{n\varepsilon_1^2}{2M^2}\Big) + c_1n\exp\Big(-\frac{sM}{4}\Big) \tag{A.6}$$

Let $M = c_0n^\gamma$ with $0 < \gamma < 1/2 - \kappa$, and we have

$$P(|\hat{\rho}_{Ij}^* - \rho_{0j}| \geq \varepsilon/2) \leq 2\exp(-c_2\varepsilon^2n^{1-2\gamma}) + c_1n\exp(-c_3n^\gamma) \tag{A.7}$$

where c_0, c_1, c_2, and c_3 are some constants.

Let $\varepsilon = cn^{-\kappa}$, and combining (A.1), (A.2), and (A.7), we have

$$\begin{aligned} P(\max_{1\leq j\leq p} ||\hat{\rho}_{Ij}| - |\rho_{0j}|| \geq cn^{-\kappa}) &\leq pP(||\hat{\rho}_{Ij}| - |\rho_{0j}|| \geq cn^{-\kappa}) \\ &\leq c_1p\exp(-c_2n^{1-2\gamma-2\kappa}) + c_3np\exp(-c_4n^\gamma) \end{aligned} \tag{A.8}$$

From (A.8), for $p = O(\exp(n^b))$ with $0 < b < \min(1 - 2\gamma - 2\kappa, \gamma)$, it is easy to show $\eta_n = \{\max_{j\in A_0} ||\hat{\rho}_{Ij}| - |\rho_{0j}|| < cn^{-\kappa}\} \subseteq \{A_0 \subseteq \hat{A}\}$, and therefore

$$\begin{aligned} P(A_0 \subseteq \hat{A}) &\geq P(\eta_n) = 1 - P(\eta_n^c) = 1 - P(\max_{j\in A_0} ||\hat{\rho}_{Ij}| - |\rho_{0j}|| \geq cn^{-\kappa}) \\ &\geq 1 - sP(||\hat{\rho}_{Ij}| - |\rho_{0j}|| \geq cn^{-\kappa}) \\ &\geq 1 - c_1s\exp(-c_2n^{1-2\gamma-2\kappa}) - c_3ns\exp(-c_4n^\gamma) \longrightarrow 1, \end{aligned} \tag{A.9}$$

where s is the number of A_0. □

Proof of Theorem 2. Let $\delta_n = n^{-r/(2r+1)} + a_n$. Define $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0^* + \delta_n \boldsymbol{v}$ with $\boldsymbol{v} = (v_1, v_2, \cdots, v_d)^T$, and $\boldsymbol{\beta}_0^*$ are the true values of $\boldsymbol{\beta}^*$.

We next show that, for any given $\varepsilon > 0$, there exists a large enough constant C such that

$$P\{\sup_{\|\boldsymbol{v}\| \geq C} \hat{Q}_I(\boldsymbol{\beta}^*) < \hat{Q}_I(\boldsymbol{\beta}_0^*)\} \leq 1 - \varepsilon, \tag{A.10}$$

where $\hat{Q}_I(\boldsymbol{\beta}^*)$ is defined in (2.15).

Let $\tau(\boldsymbol{\beta}^*) = \hat{Q}_I(\boldsymbol{\beta}^*) - \hat{Q}_I(\boldsymbol{\beta}_0^*)$, and then we have

$$\begin{aligned} \tau(\boldsymbol{\beta}^*) &= \sum_{i=1}^{n} \hat{\Delta}_i \phi_h(Y_i - \boldsymbol{x}^{*T}(\boldsymbol{\beta}_0^* + \delta_n \boldsymbol{v})) - \hat{\Delta}_i \phi_h(Y_i - \boldsymbol{x}^{*T}\boldsymbol{\beta}_0^*) - n \sum_{k=1}^{d} \{p_{\lambda_k}(|\beta_{0k}^* + \delta_n v_k|) - p_{\lambda_k}(|\beta_k^*|)\} \\ &= J_1 + J_2 \end{aligned} \tag{A.11}$$

Taylor expanding J_1 around ε_i, we have

$$\begin{aligned} J_1 &= \hat{\Delta}_i \sum_{i=1}^{n} \delta_n \phi_h'(\varepsilon_i) \boldsymbol{x}^{*T}\boldsymbol{v} + \hat{\Delta}_i \sum_{i=1}^{n} \delta_n^2 \phi_h''(\varepsilon_i)(\boldsymbol{x}^{*T}\boldsymbol{v})^2 + \hat{\Delta}_i \sum_{i=1}^{n} \delta_n^3 \phi_h'''(\xi_i)(\boldsymbol{x}^{*T}\boldsymbol{v})^3 \\ &= J_{11} + J_{12} + J_{13} \end{aligned} \tag{A.12}$$

where ξ_i lies in ε_i and $\varepsilon_i - \delta_n \boldsymbol{x}^{*T}\boldsymbol{v}$.

Invoking Assumptions A5 and A6 and simple calculation yields $J_{11} = O_p(n\delta_n^2 \|\boldsymbol{v}\|)$. Similarly, we also have $J_{13} = O_p(n\delta_n^3 \|\boldsymbol{v}\|^3)$. From $J_{12} = \delta_n^2 n F(\boldsymbol{x}, h) \boldsymbol{v}^T E(\boldsymbol{x}^T \boldsymbol{x}) \boldsymbol{v}(1 + Op(1))$, we have $J_{12} = O_p(n\delta_n^2 \|\boldsymbol{v}\|^2)$. Therefore, by choosing a sufficient large C, J_{12} dominates J_{11} and J_{13} uniformly in $\|\boldsymbol{v}\| = C$.

Moreover, invoking $p_{\lambda_k}(0) = 0$ and an argument of Taylor expansion, we get that

$$\begin{aligned} J_2 &\leq n\delta_n \sum_{k=1}^{s} \left\{ p_{\lambda_k}'(|\beta_{0k}^*|)\text{sign}(\beta_{0k}^*) v_k + \frac{1}{2}\delta_n p_{\lambda k}''(|\beta_{0k}^*|) v_k^2 \right\} \\ &\leq n\sqrt{s}\delta a_n \|\boldsymbol{v}\| + n\delta_n^2 b_n \|\boldsymbol{v}\|^2. \end{aligned} \tag{A.13}$$

Then by the condition $b_n \to 0$, J_2 is also dominated by J_{12} in $\|\boldsymbol{v}\| = C$. Hence, by choosing a sufficiently large C, (A.10) holds. There exists local maximizers $\hat{\boldsymbol{\beta}}^*$ such that $\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0^*\| = O_p(\delta_n)$. □

Proof of Theorem 3. It is sufficient to show that, for any $\hat{\boldsymbol{\beta}}^*$ that satisfies $\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0^*\| = O_p(n^{-r/(2r+1)})$ and for some given $\varepsilon = Cn^{-r/(2r+1)}$, when $n \to \infty$ with probability tending to 1, we have

$$\frac{\partial \hat{Q}_I(\boldsymbol{\beta}^*)}{\partial \beta_j} < 0, \quad \text{for } 0 < \beta_j < \varepsilon, j \in A_0^c, \tag{A.14}$$

and

$$\frac{\partial \hat{Q}_I(\boldsymbol{\beta}^*)}{\partial \beta_j} > 0, \quad \text{for } -\varepsilon < \beta_j < 0, j \in A_0^c. \tag{A.15}$$

By a similar proof as in (A.12) in Theorem 2, we can obtain that

$$\frac{\partial \hat{Q}_I(\boldsymbol{\beta}^*)}{\partial \beta_j} = -\sum_{i=1}^{n} \hat{\Delta}_i \phi_h'(Y_i - \boldsymbol{x}^{*T}\boldsymbol{\beta}^*) x_{ij} - n p_{\lambda_j}'(|\beta_j|)\text{sign}(\beta_j)$$

$$
\begin{aligned}
&= -\sum_{i=1}^{n} \hat{\Delta}_i \{\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)\boldsymbol{x}^{*T}\boldsymbol{v} + \phi_h'''(\xi_i)(\boldsymbol{x}^{*T}\boldsymbol{v})^2\} x_{ij} - n p_{\lambda_j}'(|\beta_j|)\mathrm{sign}(\beta_j) \\
&= n\lambda_j \{\frac{p_{\lambda_j}'(|\beta_j|)\mathrm{sign}(\beta_j)}{\lambda_j} + O_p(\frac{n^{-r/(2r+1)}}{\lambda_j})\} \qquad \text{(A.16)}
\end{aligned}
$$

where ξ_i lies in ε_i and $\varepsilon_i - \delta_n \boldsymbol{x}^{*T}\boldsymbol{v}$.

By Assumption A7 and $\lambda_j n^{r/(2r+1)} > \lambda_{min} n^{r/(2r+1)} \longrightarrow \infty$, the sign of derivation is determined by that of β_j. Then (A.14) and (A.15) hold, which imply that $\hat{\beta}_j = 0$, $j \in A_0^c$, with probability tending to 1.

□

AIMS Press