AIMS Mathematics, 10(6): 13795–13824.
DOI: 10.3934/math.2025621
Received: 16 April 2025
Revised: 02 June 2025
Accepted: 06 June 2025
Published: 16 June 2025

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

Research article

An efficient iterative model averaging framework for ultrahigh-dimensional linear regression models with missing data

Xianwen Ding[1], **Tong Su**[2,*] **and Yunqi Zhang**[2]

[1] Department of Statistics, Jiangsu University of Technology, Changzhou, China
[2] Key Laboratory of Statistical Modeling and Data Analysis of Yunnan Province, Yunnan University, Kunming, China

* **Correspondence:** Email: tsu@ynu.edu.cn.

Abstract: This paper addresses the prediction problem in linear regression models with ultrahigh-dimensional covariates and missing response data. Assuming a missing at random mechanism, we introduced a novel nonparametric multiple imputation method to handle missing response values. Based on these imputed responses, we proposed an efficient iterative model averaging method that integrates an iterative screening process within a model averaging framework. The weights for the candidate models were determined using the Bayesian information criterion, ensuring an optimal balance between model fit and complexity. The computational feasibility of the proposed approach stems from its iterative structure, which significantly reduces the computational burden compared to conventional methods. Under certain regularity conditions, we demonstrated that the proposed method effectively mitigates the risk of overfitting and yields consistent estimators for the regression coefficients. Simulation studies and a real-world data application illustrate the practical efficacy of the proposed approach, showing its superior performance in terms of predictive accuracy and flexibility when compared to several competing approaches.

Keywords: Bayesian information criterion; iterative model averaging; missing response data; multiple imputation; ultrahigh-dimensional regression
Mathematics Subject Classification: 62D10

1. Introduction

Missing data are a common challenge across various research fields, including health studies, clinical trials, and longitudinal investigations. Such missingness can arise from multiple factors, such as the high cost of measuring critical response variables or logistical constraints that hinder data collection due to demographic or economic limitations. Missing values not only reduce the effective

sample size, which in turn diminishes estimation efficiency, but also compromise the validity of standard complete-data analysis methods. To mitigate this issue, several statistical techniques have been developed, including multiple imputation, inverse probability weighting, and augmented inverse probability weighting (AIPW), all aimed at improving the robustness of estimation. Particularly when response data are missing at random (MAR), significant research has focused on developing methods to enhance both bias correction and estimation efficiency. References [1] and [2] offer comprehensive reviews of the theoretical advancements and practical applications in this area.

Regression-based prediction methods have become fundamental tools for improving accuracy in various application domains. Related efforts in predictive modeling have demonstrated success using deep regression frameworks for thermal control in photovoltaic systems [3], as well as hybrid associative classifiers incorporating support vector machines to enhance classification accuracy [4]. Competitively in theory with these machine learning approaches, model averaging is a well-established statistical technique designed to further improve prediction performance by combining multiple candidate models. By assigning weights based on model quality, model averaging aims to produce more reliable and robust estimates. In recent years, frequentist model averaging has gained considerable traction, with an expanding body of research exploring a variety of methodologies. Notable contributions include Mallows model averaging (MMA, [5]), jackknife model averaging (JMA, [6]), the heteroskedasticity-robust C_p criterion [7], and Kullback-Leibler loss model averaging [8]. In the context of missing data, model averaging has garnered significant attention. Schomaker et al. [9] introduced two model averaging approaches specifically tailored to handle missing data, while Dardanoni et al. [10] applied model averaging to navigate the bias-precision trade-off in linear regression models with missing covariates. Zhang [11] extended the MMA framework to cases where covariates are entirely missing at random, and Fang et al. [12] proposed a novel model averaging framework for fragmentary data. Liang and Wang [13] further contributed by developing a robust model averaging method for partially linear models with responses missing at random. Liang and Zhou [14] then proposed a new model averaging method based on the weighted generalized method of moments for missing response problems. More recently, Liang et al. [15] addressed optimal model averaging for partially linear models with responses missing at random and measurement error in some covariates.

In the realm of high-dimensional data analysis, significant progress has been made in extending model averaging techniques to accommodate the challenges posed by a growing number of predictors. Lu and Su [16] expanded the JMA criterion of [6] to quantile regression models, allowing the number of predictors to scale with the sample size. Zhang et al. [17] introduced a novel criterion for selecting weights, enabling the development of parsimonious model averaging estimators. In scenarios where the number of predictors increases exponentially with the sample size, Ando and Li [18] introduced a two-step model averaging method that combines marginal screening with JMA for ultrahigh-dimensional linear regression models. This approach was later expanded by [19] to accommodate generalized linear models. Cheng and Hansen [20] further explored model averaging procedures for ultrahigh-dimensional factor-augmented linear regression models using principal component analysis, while Chen et al. [21] investigated semiparametric model averaging methods for nonlinear dynamic time series regression models in similar settings. Lan et al. [22] introduced the sequential screening approach into model averaging for high-dimensional linear regression models. Despite these advancements, much of the research on model averaging for ultrahigh-dimensional data has been limited to complete data settings, leaving the complexities of missing data in high-dimensional contexts largely unaddressed.

This paper introduces a novel iterative model averaging (IMA) method that integrates an iterative screening procedure with model averaging to handle ultrahigh-dimensional regression models when some of the response data are subject to MAR. To address the missing responses, we develop a nonparametric multiple imputation procedure, which offers greater flexibility compared to the parametric assumptions of the MAR mechanism studied in [23]. The imputed response values are iteratively updated based on the residuals from prior iterations. This iterative screening process reduces the dominance of highly weighted predictors in earlier stages, thereby allowing additional relevant predictors to incrementally contribute to parameter estimation. The candidate model weights are determined using the Bayesian information criterion (BIC), enabling larger weights to be assigned to more influential predictors. This process also ensures that the method remains computationally feasible, even in ultrahigh-dimensional settings, by limiting each step to candidate models of size one. Under regularity conditions, we demonstrate that the proposed IMA method produces consistent estimators of regression coefficients and exhibits strong model-fitting performance. Our theoretical results are supported by several numerical studies, confirming the efficacy of the method in practice.

The paper is organized as follows: In Section 2, we outline the model setup and introduce the IMA procedure designed for regression models with missing responses. Section 3 presents the theoretical properties of the IMA procedure. Sections 4 and 5 provide a comprehensive evaluation of the method through extensive simulation studies and its application to a real-world dataset. Finally, Section 6 provides the conclusion of the paper, and Section 7 addresses the limitations and potential directions for future research. All technical proofs can be found in the Appendix.

2. Methodologies

2.1. Missing response imputation using multiple imputation

Let $\{(y_i, \boldsymbol{X}_i)^\top\}_{i=1}^n$ represent a set of n independent and identically distributed (i.i.d.) random samples, where y_i denotes the response variable, and $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ip_n})^\top \in \mathbb{R}^{p_n}$ denotes the ultrahigh-dimensional covariates. Without loss of generality, it is assumed that $\mathbb{E}(x_{ij}) = 0$ and consider the following linear regression relationship:

$$y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^\top$ is a p_n-dimensional vector of regression coefficients, and ε_i are i.i.d. random errors with mean zero and finite variance σ^2. Throughout this paper, the number of covariates p_n is permitted to diverge with the sample size n, satisfying $p_n \gg n$, i.e., $\log(p_n) = o(n^\alpha)$ for some constant $\alpha \in (0, 1)$. In this framework, the response variables y_i may be missing, while the covariates $\boldsymbol{X}_i$ are assumed to be fully observed. Thus, the dataset comprises the observations $\{(y_i, \boldsymbol{X}_i, \delta_i)^\top\}_{i=1}^n$, where $\delta_i = 1$ indicates that y_i is observed, and $\delta_i = 0$ otherwise. Let $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p_n}) \in \mathbb{R}^{n \times p_n}$ denote the design matrix, where $\boldsymbol{x}_j$ represents the j-th column of covariates.

We assume that the missing data mechanism adheres to the MAR assumption, implying that the missingness indicator δ_i is conditionally independent of the response y_i, given the covariates $\boldsymbol{X}_i$. Formally, this can be expressed as $\Pr(\delta_i = 1 \mid \boldsymbol{X}_i, y_i) = \Pr(\delta_i = 1 \mid \boldsymbol{X}_i) \triangleq \pi(\boldsymbol{X}_i)$, where $\pi(\cdot)$ is the selection probability function that models the selection bias underlying the missingness pattern.

To overcome the "curse of dimensionality" frequently encountered in ultrahigh-dimensional data with missing responses, we assume that $y \perp \delta \mid x_j$ for each $j = 1, \ldots, p_n$ (see [24]), where $\perp$ stands for

statistical independence. This assumption enables us to impute the missing values of y_i by leveraging information from individual covariates $\boldsymbol{x}_j$, rather than relying on the full covariate vector $\boldsymbol{X}$. The multiple imputation procedure for estimating missing responses is then defined as follows:

$$\hat{y}_i = \delta_i y_i + (1-\delta_i)\frac{1}{p_n\mathcal{K}}\sum_{j=1}^{p_n}\sum_{k=1}^{\mathcal{K}}\tilde{y}_{ik}^{(j)}, \qquad i = 1,\ldots,n,$$

where $\mathcal{K}$ is the number of multiple imputations, and $\{\tilde{y}_{ik}^{(j)}\}_{k=1}^{\mathcal{K}}$ are $\mathcal{K}$ independent imputations for the missing value of y_i, drawn from the estimated conditional distribution $\hat{F}(\tilde{y}\,|\,x_{ij})$. This conditional distribution $\hat{F}(\tilde{y}\,|\,x_{ij}) = \sum_{l=1}^{n}\kappa_{il}^{(j)}I(y_l \le \tilde{y})$ is a kernel estimator of the true distribution $F(\tilde{y}\,|\,x_{ij})$, where $\kappa_{il}^{(j)} = \delta_l K_h(x_{lj}-x_{ij})/\sum_{k=1}^{n}\delta_k K_h(x_{kj}-x_{ij})$ is the kernel weight. Here, $K_h(u) = K(u/h)$ is the kernel function with bandwidth $h = h_n$, a positive smoothing parameter that tends to zero when n is growing, and $I(\cdot)$ is the indicator function. It is important to note that the imputed values $\tilde{y}_{ik}^{(j)}$ have a discrete distribution, where the selecting probability of y_l (when $\delta_l = 1$) is given by $\kappa_{il}^{(j)}$. This structure ensures that the imputation procedure integrates information across covariates while maintaining computational efficiency in ultrahigh-dimensional settings.

Remark 2.1. *An alternative imputation method commonly used for handling missing response data is the AIPW method. In this approach, the estimated response value for the i-th individual is given by*

$$\hat{y}_i^* = \frac{\delta_i y_i}{\pi(\boldsymbol{X}_i)} + \left(1-\frac{\delta_i}{\pi(\boldsymbol{X}_i)}\right)\frac{1}{p_n\mathcal{K}}\sum_{j=1}^{p_n}\sum_{k=1}^{\mathcal{K}}\tilde{y}_{ik}^{(j)}, \qquad i = 1,\ldots,n. \tag{2.1}$$

As shown in Eq (2.1)*, the AIPW method involves imputing both the missing and the observed responses. Within the empirical likelihood framework, Tang and Qin [25] established the semiparametric efficiency of the AIPW estimator. In the setting of robust quantile regression, Chen et al. [26] further demonstrated that quantile estimators based on both $\hat{y}_i^*$ and $\hat{y}_i$ can achieve the semiparametric efficiency bound, provided that the error distributions are identical. Nevertheless, since the AIPW approach requires repeated imputation even for observed responses, it may lead to increased risk of overfitting, particularly in ultrahigh-dimensional scenarios where $p \gg n$. In contrast, the proposed imputation method is computationally more efficient, as it avoids redundant imputation for observed values. By averaging over the imputed values, it offers a direct and practical solution for handling missing response data. In addition, the proposed multiple imputation method, facilitated by kernel-assisted imputation, is less sensitive to model misspecification and tends to be more robust in high-dimensional contexts.*

2.2. Univariate model averaging with missing response

Let $\mathcal{A}_{\mathcal{F}} = \{1,\ldots,p_n\}$ represent the full predictor index set, and let $\mathcal{A} = \{j_1,\ldots,j_q\} \subseteq \mathcal{A}_{\mathcal{F}}$ be some nonempty subset of cardinality $|\mathcal{A}| = q \ge 1$. The complement of $\mathcal{A}$ is denoted by $\mathcal{A}^c = \mathcal{A}_{\mathcal{F}} \setminus \mathcal{A}$. Define $\mathbf{X}_{\mathcal{A}} = (\boldsymbol{x}_j : j \in \mathcal{A}) \in \mathbb{R}^{n\times|\mathcal{A}|}$ and $\boldsymbol{\beta}_{0\mathcal{A}}$ as the sub-matrix of the design matrix $\mathbf{X}$ and the sub-vector of $\boldsymbol{\beta}_0$ indexed by $\mathcal{A}$, respectively. Similarly, we define $\mathbf{X}_{\mathcal{A}^c}$ and $\boldsymbol{\beta}_{0\mathcal{A}^c}$. The candidate model $\mathcal{A}$ with fully observed response data is given by

$$y_i = \boldsymbol{X}_{i\mathcal{A}}^{\top}\boldsymbol{\beta}_{0\mathcal{A}} + \varepsilon_i, \qquad i = 1,\ldots,n, \tag{2.2}$$

where $\boldsymbol{X}_{i\mathcal{A}} = (x_{ij} : j \in \mathcal{A}) \in \mathbb{R}^{|\mathcal{A}|}$ represents the covariates in $\boldsymbol{X}_i$ corresponding to the predictors in the candidate model $\mathcal{A}$. A model averaging estimator $\hat{\boldsymbol{\beta}}^*_{\mathcal{A}}$ of $\boldsymbol{\beta}_{0\mathcal{A}}$ in model (2.2) can be expressed as

$$\hat{\boldsymbol{\beta}}^*_{\mathcal{A}} = \sum_{\mathcal{A}\subseteq\mathcal{A}_{\mathcal{F}}} \hat{\omega}^*_{\mathcal{A}} \hat{\boldsymbol{\beta}}_{\mathcal{A}},$$

where $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ is an estimator of $\boldsymbol{\beta}_{\mathcal{A}}$, and $\hat{\omega}^*_{\mathcal{A}}$ represents the weight assigned to model $\mathcal{A}$, which can be estimated using the commonly applied BIC criterion.

The classical BIC-based model averaging procedures become computationally prohibitive in high-dimensional settings due to the cardinality of candidate models scaling exponentially with predictor dimensionality. To alleviate this issue, we propose a simplified BIC model averaging procedure that focuses on univariate candidate models ($|\mathcal{A}| = 1$), reducing the number of candidate models to p_n. Without loss of generality, let the design matrix $\mathbf{X}$ be standardized such that for each $j = 1, \ldots, p_n$, the j-th column of $\mathbf{X}$ satisfies $n^{-1}\|\boldsymbol{x}_j\|^2 = n^{-1}\boldsymbol{x}_j^\top \boldsymbol{x}_j = 1$. For the j-th model, the BIC is defined as

$$\mathrm{BIC}_j = n\log(\|\hat{\boldsymbol{Y}} - \boldsymbol{x}_j\hat{\beta}_j\|^2) + \log(n) + 2\log(p_n), \qquad j = 1, \ldots, p_n,$$

where $\hat{\boldsymbol{Y}} = (\hat{y}_1, \ldots, \hat{y}_n)^\top$ and $\hat{\beta}_j = \boldsymbol{x}_j^\top \hat{\boldsymbol{Y}}/n$ are the estimated coefficient for the j-th model. The corresponding BIC-based model averaging estimator is then given by $\hat{\boldsymbol{\beta}}_U = (\hat{\omega}^U_1\hat{\beta}_1, \ldots, \hat{\omega}^U_{p_n}\hat{\beta}_{p_n})^\top$, where the BIC weight for the j-th model is defined as

$$\hat{\omega}^U_j = \frac{\exp(-\mathrm{BIC}_j/2)}{\sum_{j=0}^{p_n}\exp(-\mathrm{BIC}_j/2)}, \qquad j = 0, 1, \ldots, p_n,$$

with $\mathrm{BIC}_0 = n\log(\|\hat{\boldsymbol{Y}}\|^2)$.

2.3. Iterative model averaging with missing response

Univariate model averaging may lead to inaccurate predictions due to its reliance on a limited subset of the information pool, thereby ignoring the majority of potentially explanatory data features. A natural extension to improve prediction accuracy is to increase the candidate model size from 1 to 2, but this approach can become computationally expensive, particularly in high-dimensional settings, where the number of candidate models scales to p_n^2. Motivated by the forward regression approach, we propose an iterative method that reduces the impact of heavily weighted predictors by updating the response at each iteration based on the residuals. This enables enhanced integration of auxiliary predictors into the inferential process, thereby improving the efficiency of parameter estimation. We designate this methodology as IMA. The detailed procedure is outlined below.

Let $\hat{\boldsymbol{Y}}^{(1)} = \hat{\boldsymbol{Y}}$ represent the initial response vector, and $\hat{\boldsymbol{Y}}^{(m)} = (\hat{y}_1^{(m)}, \ldots, \hat{y}_n^{(m)})^\top$ represent the response vector at the m-th iteration. In the m-th iteration, we fit p_n univariate models using the current response $\hat{y}^{(m)}$ and the covariates x_j, $j = 1, \ldots, p_n$. The corresponding estimator is given by $\hat{\beta}_{mj} = \boldsymbol{x}_j^\top \hat{\boldsymbol{Y}}^{(m)}/n$. The BIC for the j-th model is calculated as

$$\mathrm{BIC}_{mj} = n\log(\|\hat{\boldsymbol{Y}}^{(m)} - \boldsymbol{x}_j\hat{\beta}_{mj}\|^2) + \log(n) + 2\log(p_n).$$

Although the weighting coefficient assigned to the null model is less than 1, the covariates can still explain some information about the response. Therefore, the null model is also fitted, resulting in

$\text{BIC}_{m0} = n\log(\|\hat{\boldsymbol{Y}}^{(m)}\|^2)$. The corresponding BIC model averaging weights for each candidate model are denoted as

$$\hat{\omega}_{mj} = \frac{\exp(-\text{BIC}_{mj}/2)}{\sum_{j=0}^{p_n}\exp(-\text{BIC}_{mj}/2)}, \qquad j = 0, 1, \ldots, p_n.$$

Subsequently, the response vector is updated for the $(m+1)$-th iteration as

$$\hat{\boldsymbol{Y}}^{(m+1)} = \hat{\boldsymbol{Y}}^{(m)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(m)},$$

where $\hat{\boldsymbol{\beta}}^{(m)} = (\hat{\omega}_{m1}\hat{\beta}_{m1}, \ldots, \hat{\omega}_{mp_n}\hat{\beta}_{mp_n})^\top \in \mathbb{R}^{p_n}$. Without loss of generality, this iterative process is assumed to cease after M iterations. The final iterative model averaging estimator is then given by

$$\hat{\boldsymbol{\beta}}^M = \sum_{m=1}^{M}\hat{\boldsymbol{\beta}}^{(m)}.$$

The convergence criteria for this algorithm will be discussed later in this work. Let $(y_*, \boldsymbol{X}_*)$ be an independent copy of $(y_i, \boldsymbol{X}_i)$ for some $1 \le i \le n$, and we can predict the value of y_* by $\boldsymbol{X}_*^\top\hat{\boldsymbol{\beta}}^M$.

Remark 2.2. *Commonly used kernel functions include: (1) uniform kernel $K(u) = (1/2)I(|u| \le 1)$, (2) logistic kernel $K(u) = e^{-u}/(1+e^{-u})^2$, (3) Epanechnikov kernel $K(u) = (3/4)(1-u^2)I(|u| \le 1)$, (4) triangular kernel $K(u) = (1-|u|)I(|u| \le 1)$, (5) biweight kernel $K(u) = (15/16)(1-u^2)^2I(|u| \le 1)$, and (6) Gaussian kernel $K(u) = (2\pi)^{-1/2}e^{-u^2/2}$. Through a sensitivity analysis of kernel function choices in the numerical experiments presented in Section 4, we find that the proposed kernel-assisted IMA procedure exhibits robustness to the choice of kernel function. Therefore, for practical applications, we recommend employing the Gaussian kernel for multiple imputation, primarily due to its computational simplicity and widespread applicability.*

2.4. Algorithmic steps of the IMA procedure

The IMA algorithm with missing response is designed to estimate regression coefficients in high-dimensional settings where the response variable may be partially missing. The procedure begins with a training dataset $\{(y_i, \boldsymbol{X}_i, \delta_i)^\top\}_{i=1}^n$, where δ_i indicates whether y_i is observed. For observations with missing responses ($\delta_i = 0$), the method employs a kernel-based nonparametric imputation strategy. Specifically, for each covariate x_{ij}, imputed response values $\{\tilde{y}_{ik}^{(j)}\}_{k=1}^{\mathcal{K}}$ are sampled from an estimated conditional distribution $\hat{F}(\tilde{y} \mid x_{ij})$, constructed using a kernel smoothing technique over observed data.

Next, the IMA procedure then proceeds, starting with the initialization of iteration index $m = 0$ and weight $\hat{\omega}_{00} = 0$. In each iteration, the null model's raw weight is computed based on the norm of the current response vector $\hat{\boldsymbol{Y}}^{(m)}$. For each covariate x_j, a univariate regression is fitted, yielding coefficient estimates $\hat{\beta}_{mj}$, and associated raw weights $\hat{\omega}_{mj}$ are calculated based on the BIC that accounts for model complexity.

All raw weights, including that of the null model, are normalized to ensure they sum to one. The iteration produces a weighted average estimator $\hat{\boldsymbol{\beta}}^{(m)}$, which is then used to update the response vector for the subsequent iteration via residual adjustment: $\hat{\boldsymbol{Y}}^{(m+1)} = \hat{\boldsymbol{Y}}^{(m)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(m)}$. This iterative process continues until the change in the null model's weight across successive iterations falls below a pre-specified threshold ϵ, ensuring convergence.

Algorithm 1: Iterative model averaging with missing response

Input: Training dataset $\{(y_i, \boldsymbol{X}_i, \delta_i)^\top\}_{i=1}^n$.

Output: Iterative model averaging estimator $\hat{\boldsymbol{\beta}}^M$.

for $i = 1$ **to** n **do**

  if $\delta_i = 0$ **then**

    for $j = 1$ **to** p_n **do**

      Generate imputed values $\{\tilde{y}_{ik}^{(j)}\}_{k=1}^{\mathcal{K}}$ by sampling from the conditional distribution

$$\hat{F}(\tilde{y} \mid x_{ij}) = \sum_{l=1}^{n} \frac{\delta_l K_h(x_{lj} - x_{ij})}{\sum_{k=1}^{n} \delta_k K_h(x_{kj} - x_{ij})} I(y_l \le \tilde{y});$$

  Initialize $\hat{y}_i^{(1)}$ by

$$\hat{y}_i^{(1)} = \delta_i y_i + (1 - \delta_i) \frac{1}{p_n \mathcal{K}} \sum_{j=1}^{p_n} \sum_{k=1}^{\mathcal{K}} \tilde{y}_{ik}^{(j)};$$

Initialize $m = 0$ and $\hat{\omega}_{00} = 0$;

repeat

  Set $m \leftarrow m + 1$;

  Calculate the raw weights $\hat{\omega}_{10}$ of the null model by

$$\hat{\omega}_{10} = \exp\Big\{-n \log(\|\hat{\boldsymbol{Y}}^{(m)}\|^2)\Big\};$$

  for $j = 1$ **to** p_n **do**

    Fit the j-th univariate model $\hat{\boldsymbol{Y}}^{(m)} = \boldsymbol{x}_j \hat{\beta}_{mj}$ by

$$\hat{\beta}_{mj} = n^{-1} \boldsymbol{x}_j^\top \hat{\boldsymbol{Y}}^{(m)};$$

    Calculate the raw weights $\hat{\omega}_{mj}$ of the j-th univariate model by

$$\hat{\omega}_{mj} = \exp\Big\{-n \log(\|\hat{\boldsymbol{Y}}^{(m)} - \boldsymbol{x}_j \hat{\beta}_{mj}\|^2) - \log(n) - 2\log(p_n)\Big\};$$

  for $j = 0$ **to** p_n **do**

    Normalize the raw weights $\hat{\omega}_{mj}$ by

$$\hat{\omega}_{mj} \leftarrow \frac{\hat{\omega}_{mj}}{\sum_{j=0}^{p_n} \hat{\omega}_{mj}};$$

  Storage the m-th iterative univariate model averaging estimator of $\hat{\boldsymbol{\beta}}^{(m)}$ by

$$\hat{\boldsymbol{\beta}}^{(m)} = (\hat{\omega}_{m1}\hat{\beta}_{m1}, \ldots, \hat{\omega}_{mp_n}\hat{\beta}_{mp_n})^\top;$$

  Update the response vector of the $(m+1)$-th iteration $\hat{\boldsymbol{Y}}^{(m+1)}$ by

$$\hat{\boldsymbol{Y}}^{(m+1)} = \hat{\boldsymbol{Y}}^{(m)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(m)};$$

until *met termination condition:* $\{\hat{\omega}_{m0} - \hat{\omega}_{(m-1)0}\}/\hat{\omega}_{(m-1)0} < \epsilon$;

return $\hat{\boldsymbol{\beta}}^M = \sum_{m=1}^{M} \hat{\boldsymbol{\beta}}^{(m)}$.

Finally, the cumulative model averaging estimator is obtained as $\hat{\boldsymbol{\beta}}^M = \sum_{m=1}^{M} \hat{\boldsymbol{\beta}}^{(m)}$, where M denotes the total number of iterations. The IMA algorithm effectively integrates imputation with adaptive model averaging to address challenges arising from missing data in high-dimensional regression problems.

The detailed algorithmic steps are outlined in Algorithm 1.

As demonstrated in Algorithm 1, the proposed IMA procedure is fundamentally data-driven and can be effectively implemented using the observed dataset $\{(y_i, \boldsymbol{X}_i, \delta_i)^\top\}_{i=1}^n$. Nevertheless, several hyperparameters, which are recommended based on our empirical findings in Section 4, play a crucial role in enhancing computational efficiency. These are summarized in Table 1.

Table 1. Recommended hyperparameters in the IMA procedure.

Hyperparameter	**Description**	**Value**
$\mathcal{K}$	Number of multiple imputations	30
$K_h(u) = K(u/h)$	Kernel function	Gaussian kernel function
$h = h_n$	Bandwidth of given kernel function $K(\cdot)$	Selected via the data-driven procedure
ϵ	Stopping threshold for the IMA procedure	0.0001
M	Number of iterative steps	20

3. Theoretical properties

To establish the theoretical properties of the proposed IMA procedure, we introduce the following notations and conditions. Let $\beta^2_{(1)} \geq \beta^2_{(2)} \geq \ldots \geq \beta^2_{(p_n)}$ and $\hat{\omega}^U_{(1)} \geq \hat{\omega}^U_{(2)} \geq \ldots \geq \hat{\omega}^U_{(p_n)}$ be the ordered statistics of $\{\beta^2_{0j} : 1 \leq j \leq p_n\}$ and $\{\hat{\omega}^U_j : 1 \leq j \leq p_n\}$, respectively.

Condition 1: The missing propensity function $\pi(\boldsymbol{X}) = \Pr(\delta = 1 \,|\, \boldsymbol{X})$ and the probability density function $f_j(x)$ of x_j $(j = 1, \ldots, p_n)$ both have continuous and bounded second-order derivatives over the support $\mathcal{X}_j$ of x_j. Additionally, there exists a constant $C_0 > 0$ such that $\inf_x \pi(x) \geq C_0$.

Condition 2: The kernel function $K(\cdot)$ is a probability density function satisfying: (i) It is bounded and has compact support; (ii) It is symmetric with $\int t^2 K(t)dt < \infty$; (iii) There exists a constant $C_1 > 0$ such that $K(t) \geq C_1$ in a closed interval centered at zero. Furthermore, the smoothing bandwidth h satisfies $nh \to \infty$ and $\sqrt{n}h^2 \to 0$ as $n \to \infty$.

Condition 3: The random variables x_j, y, and $x_j y$ satisfy a sub-exponential tail probability uniformly in p_n. That is, there exists a constant $u_0 > 0$ such that for all $0 < u \leq 2u_0$,

$$\max_{1 \leq j \leq p_n} \mathbb{E}\{\exp(2ux_j^2)\} < \infty, \quad \max_{1 \leq j \leq p_n} \mathbb{E}\{\exp(2ux_j y)\} < \infty, \quad \mathbb{E}\{\exp(2uy^2)\} < \infty.$$

Condition 4: There exist positive constants C_2 and C_3, independent of n and p_n, such that $\max_j |\beta_{0j}| \leq C_2$, and $\max\{\max_{j_1 \neq j_2} |\sigma_{j_1 j_2}|, \max_j |\mathrm{corr}(x_j, y)|\} \leq C_3 < 1$.

Condition 5: Let $\beta^2_{m(1)} \geq \beta^2_{m(2)} \geq \ldots \geq \beta^2_{m(p_n)}$ be the ordered statistics of $\{\beta^2_{0mj} : 1 \leq j \leq p_n\}$ for any $m = 1, \ldots, M$. There exist positive constants C_5 and C_6, independent of n and p_n, such that $\max_{1 \leq m \leq M} |\beta_{m(1)}| \leq C_5$ and $\max_{1 \leq m \leq M} \max_j |\mathrm{corr}(x_j, y^{(m)})| \leq C_6 < 1$.

Condition 6: There exists a sequence $\mathcal{A}_n$ such that (i) $|\mathcal{A}_n| \to \infty$ and $|\mathcal{A}_n|/n \to 0$ as $n \to \infty$; (ii) $|\mathcal{A}_n^c| \, ||\boldsymbol{\beta}_{0\mathcal{A}^c}||^2 \to 0$ as $M, n \to \infty$; (iii) $|\mathcal{A}_n^c| \max_{m>1} \sum_{j \in \mathcal{A}_n^c} \hat{\omega}^2_{mj} \to 0$ as $n \to \infty$.

Condition 7: There exist positive constants C_8 and C_9 such that the eigenvalues of the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$ satisfy $C_8 \leq \lambda_{\min}[\mathrm{Cov}(\boldsymbol{X})] \leq \lambda_{\max}[\mathrm{Cov}(\boldsymbol{X})] \leq C_9$, where $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ represent the smallest and largest eigenvalues of the matrix $\mathbf{A}$, respectively.

Conditions 1 and 2 impose standard assumptions on the missing data propensity function $\pi(\boldsymbol{X})$ and the probability density functions $f_j(x)$, which are generally satisfied by commonly used distributions.

Condition 3 holds under widely employed distributions, such as the multivariate Gaussian. Similar conditions have been discussed in the work of [27]. Condition 4 places restrictions on the parameters to prevent perfect correlation between any predictor and the response, which aligns with the assumptions made in [28] and [22]. Condition 5 extends these assumptions to each iterative step. Condition 6, adapted from [22], is essential for establishing the asymptotic properties of $\hat{\boldsymbol{\beta}}^M$ as $M, n \to \infty$. Lastly, Condition 7 ensures that the design matrix behaves well, similar to the assumptions made by [29].

Theorem 3.1. *Under Conditions 1–4, suppose that $\beta^2_{(1)} - \beta^2_{(2)} > C_4$ for some positive constant C_4. Then we have $\hat{\omega}^U_{(1)} \to_p 1$ as $n \to \infty$, where $\hat{\omega}^U_{(1)}$ is the weight assigned to the univariate model averaging of the covariate with the largest absolute marginal regression coefficient.*

This theorem establishes that the information used for estimation and prediction is dominantly governed by the largest correlated covariates. Moreover, Theorem 3.1 can be easily extended to assign nearly equal weights to the first two largest correlated predictors with the response, specifically when $\beta^2_{(1)} = \beta^2_{(2)} + o(1) > \beta^2_{(3)}$, while the contributions from other predictors are negligible.

Theorem 3.2. *For $m \geq 1$, we have*

$$\text{(i)} \quad \|\hat{\boldsymbol{Y}}^{(m)}\|^2 - \|\hat{\boldsymbol{Y}}^{(m+1)}\|^2 \geq \sum_{j=1}^{p_n} n\hat{\omega}_{mj}\hat{\beta}^2_{mj},$$

$$\text{(ii)} \quad \|\hat{\boldsymbol{Y}}^{(m)}\|^2 - \|\hat{\boldsymbol{Y}}^{(m+1)}\|^2 \leq 2n(1 - \hat{\omega}_{m0})\hat{\beta}^2_{m(1)}.$$

The first part of Theorem 3.2 ensures that the proposed IMA procedure can progressively approximate the true regression relationship within a limited number of iterations. In particular, the residual sum of squares $\|\hat{\boldsymbol{Y}}^{(m)}\|^2$ decreases monotonically with each iteration, which reflects the method's strong fitting capability. The second part of Theorem 3.2 indicates that the IMA framework effectively mitigates the risk of overfitting. When no informative predictors are selected during the m-th iteration, the weight $\hat{\omega}_{m0}$ approaches 1, implying that the procedure primarily relies on the baseline model and avoids spurious fitting. Conversely, if significant predictors are present, the largest coefficient $\hat{\beta}^2_{m(1)}$ and its corresponding weight $\hat{\omega}_{m(1)}$ become dominant, while $\hat{\omega}_{m0}$ shrinks toward 0. Consequently, the upper bound in part (ii) of Theorem 3.2 demonstrates that the IMA procedure not only maintains fitting accuracy but also suppresses excessive model complexity, thus effectively alleviating overfitting.

Theorem 3.3. *Under Conditions 1–3 and 5, for those values of m satisfying $\beta^2_{m(1)} - \beta^2_{m(2)} > C_7$ with some positive constant C_7, we have $\hat{\omega}_{m(1)} \to_p 1$ as $n \to \infty$, where $\hat{\omega}_{m(1)} \geq \ldots \geq \hat{\omega}_{m(p_n)}$ denote the ordered statistics of $\{\hat{\omega}_j : 1 \leq j \leq p_n\}$. In addition, for those values of m satisfying $\beta^2_{m(1)} = O(n^{-\gamma})$, $\gamma > 0$, we have $\hat{\omega}_{m0} \to_p 1$ as $n \to \infty$.*

The first part of this theorem indicates that the largest weight in each iterative step will be assigned to the most significant coefficient. The second part of this theorem implies that the weight $\hat{\omega}_{m0}$ increases with m. Therefore, the IMA algorithm can be stopped if $\{\hat{\omega}_{(m+1)0} - \hat{\omega}_{m0}\}/\hat{\omega}_{m0} < \epsilon$ for some small $\epsilon > 0$. In our simulations, we set $\epsilon = 0.0001$.

Theorem 3.4. *Under Conditions 1–3 and 6–7, we have*

$$\|\hat{\boldsymbol{\beta}}^M - \boldsymbol{\beta}_0\| \to_p 0, \qquad as \ \min\{M, n\} \to \infty.$$

This theorem demonstrates that the IMA procedure provides a consistent estimator of $\boldsymbol{\beta}_0$. Accordingly, $y_* = \boldsymbol{X}_*^\top \hat{\boldsymbol{\beta}}^M$ is a consistent estimator of $\boldsymbol{X}_*^\top \boldsymbol{\beta}_0$ for the given $\boldsymbol{X}_*$. This implies that $\mathbb{E}\{\|\boldsymbol{X}_*^\top \hat{\boldsymbol{\beta}}^M - \boldsymbol{X}_*^\top \boldsymbol{\beta}_0\|^2\} \to 0$ as $n \to \infty$.

Remark 3.1. *In fields such as genomics and epidemiology, high-dimensional datasets typically consist of thousands of feature variables, while the response variable may be subject to MAR. The proposed iterative model averaging procedure effectively mitigates prediction risks in such ultrahigh-dimensional settings, as demonstrated in Theorem 3.4. This makes the proposed approach highly applicable in the analysis of genomic data and disease prediction.*

4. Simulation studies

In this section, several simulation studies are conducted to evaluate the finite-sample performance of the proposed estimation procedure, referred to hereafter as the IMA method. The simulation is based on 200 independent replications, each employing a sample size of $n = 100$ and two feature dimensions, $p_n = 1000$ and 3000. For each t-th replication, the data are represented as $\{\boldsymbol{Y}_{[t]}, \mathbf{X}_{[t]}, \boldsymbol{\delta}_{[t]}\}$, where $\boldsymbol{Y}_{[t]} \in \mathbb{R}^n$, $\mathbf{X}_{[t]} \in \mathbb{R}^{n \times p_n}$, and $\boldsymbol{\delta}_{[t]} \in \mathbb{R}^n$. Based on this dataset, the corresponding IMA estimator $\hat{\boldsymbol{\beta}}^M_{[t]}$ is computed. To evaluate the estimation accuracy of our proposed method and compare it with several alternative approaches, we employ the following mean squared error (MSE) criterion:

$$\mathrm{MSE}(t) = \frac{1}{100} \sum_{i=1}^{100} \left(\boldsymbol{X}_{*i}^\top \boldsymbol{\beta}_0 - \boldsymbol{X}_{*i}^\top \hat{\boldsymbol{\beta}}^M_{[t]} \right)^2, \tag{4.1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_n}$ denotes the true parameter vector, and $\{\boldsymbol{X}_{*i}\}_{i=1}^{100}$ represents an independent test dataset. The overall performance metric is obtained by calculating the median of replication-specific MSE values:

$$\mathrm{MMSE} = \mathrm{median}\{\mathrm{MSE}(1), \ldots, \mathrm{MSE}(200)\}.$$

This comprehensive evaluation framework ensures a rigorous assessment of estimation accuracy, where smaller MSE values indicate superior performance of the estimation method.

In addition to the proposed IMA method, we consider several alternative methods for comparison. These include the sequential model averaging procedure for the complete-case setting (denoted as the "CC" method, [22]), the sequential model averaging procedure without adjustments for missing data (denoted as the "FULL" method, [22]), weighted model averaging with the cross-validation (CV) method, without the constraint $\sum_{s=1}^S \omega_s = 1$ (denoted as "WMCV", [23]), and the weighted model averaging with CV method, incorporating the constraint $\sum_{s=1}^S \omega_s = 1$ (denoted as "WMCV1", [23]). Here, S denotes the number of candidate models. Additionally, we consider the BIC model averaging procedure applied to complete-case data (denoted as "MBIC").

To implement the last three methods, the predictors are grouped according to their marginal multiple-imputation sure independence screening utility [23], retaining the top 100 predictors. Subsequently, we set $S = 10$, resulting in a set of 10 candidate models, each with 10 predictors. The performance of these methods is then evaluated by replacing $\boldsymbol{X}_{*i}^\top \hat{\boldsymbol{\beta}}^M_{[t]}$ in Eq (4.1) with the weighted sum $\sum_{s=1}^S \hat{\omega}_{s[t]} \boldsymbol{X}_{*i}^\top \hat{\boldsymbol{\beta}}_{s[t]}$, where $\hat{\omega}_{s[t]}$ and $\hat{\boldsymbol{\beta}}_{s[t]}$ denote the estimators for the s-th candidate model in the t-th replication.

Experiment 4.1. This experiment is adapted from [18], where the true number of regressors, d, is set to 50. The considered linear regression model is as follows:

$$y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^\top$ is the vector of true regression coefficients, and $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ip_n})^\top$ represents the p_n vector of predictors. The noise terms, ε_i, are independent of the predictors. The covariate vector $\boldsymbol{X}_i$ is generated from a multivariate normal distribution with mean $\mathbf{0}$, and the covariance between x_{j_1} and x_{j_2} is given by $\mathrm{Cov}(x_{j_1}, x_{j_2}) = \rho^{|j_1 - j_2|}$, $1 \le j_1, j_2 \le p_n$, where ρ represents the correlation parameter and takes values of 0 and 0.5, corresponding to low and moderate correlation scenarios, respectively. The true regressors x_j are spaced evenly across the predictor vector, with $j = p_n(k-1)/d+1$, for $k = 1, \ldots, d$. The nonzero entries of $\boldsymbol{\beta}_0$ are generated from a normal distribution with a mean of 0 and a standard deviation of 0.5. Two error distributions are considered: (i) the standard normal distribution and (ii) the Student's t distribution with 3 degrees of freedom. It is assumed that the predictor values $\boldsymbol{X}_i$ are fully observed, but the response values y_i are subject to missingness. To simulate missing data for y_i, the missingness indicator δ_i is generated from a Bernoulli distribution with a probability $\pi(\boldsymbol{X}_i) = \pi(x_{i1}) = \Pr(\delta_i = 1 \mid x_{i1})$. The following three missingness mechanisms are considered:

- **M1**: $\pi(x_{i1}) = (0.3 + 0.175|x_{i1}|)I(|x_{i1}| < 4) + I(|x_{i1}| \ge 4)$. This mechanism induces a missingness pattern that depends on the absolute value of x_{i1}, with truncation at $|x_{i1}| \ge 4$.
- **M2**: $\pi(x_{i1}) = \Phi(0.5 + 3x_{i1})$, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. This mechanism introduces a monotonic relationship between the probability of missingness and the value of x_{i1}.
- **M3**: $\mathrm{logit}(\pi(x_{i1})) = 0.5 + 3x_{i1}$, where the logit function is applied to model the probability of missingness. This mechanism results in a linear relationship between x_{i1} and the log-odds of missingness.

The average proportions of missing data for the three mechanisms are approximately 56%, 44%, and 44%, respectively. For each of these settings, the number of multiple imputations is set to $\mathcal{K} = 30$ for each missing y_i.

Experiment 4.2. This experiment aims to investigate the impact of the proposed method on different sparsity structures within a regression model, following a framework modified from [30]. Specifically, the covariate vector $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ip_n})^\top$ is generated from a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, where the entries of $\boldsymbol{\Sigma}$ are defined as $\mathrm{Cov}(x_{j_1}, x_{j_2}) = 0.5^{|j_1 - j_2|}$, $1 \le j_1, j_2 \le p_n$. The true regression coefficients, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^\top$, are specified such that $\beta_{0j} = (-1)^j \times 0.5$, $1 \le j \le d$, and $\beta_{0j} = 0$ for $j > d$, where d represents the number of true nonzero coefficients. The error terms, ε_i, are generated from the standard normal distribution with a mean of 0 and a standard deviation of 1, and are independent of the covariates x_{ij} for $j = 1, \ldots, p_n$. In alignment with Experiment 4.1, the covariate vectors $\boldsymbol{X}_i$ are assumed to be fully observed, while the response values y_i are subject to missingness. The missingness indicator δ_i is generated from a Bernoulli distribution with probability $\pi(\boldsymbol{X}_i) = \pi(x_{i1}) = \Pr(\delta_i = 1 \mid x_{i1})$, where $\pi(x_{i1})$ follows the same specifications as in Experiment 4.1. The three considered missingness mechanisms yield average missing proportions of approximately 56%, 44%, and 44%, respectively. The simulation results for $d = 20$ are presented in this experiment.

The Gaussian kernel function is employed in the imputation procedure, and the optimal bandwidth for kernel density estimation is selected via the cross-validation method implemented in the `kedd` package

in R. The results from the two experiments with $p_n = 1000$ and 3000 are presented in Figures 1–4. The black dash-dotted line indicates the median MSE value (MMSE) of the proposed method and is included for direct comparison. Additionally, simulation results for Experiment 4.2 with $p_n = 1000$ using the Epanechnikov and biweight kernels are presented in Figure 5.

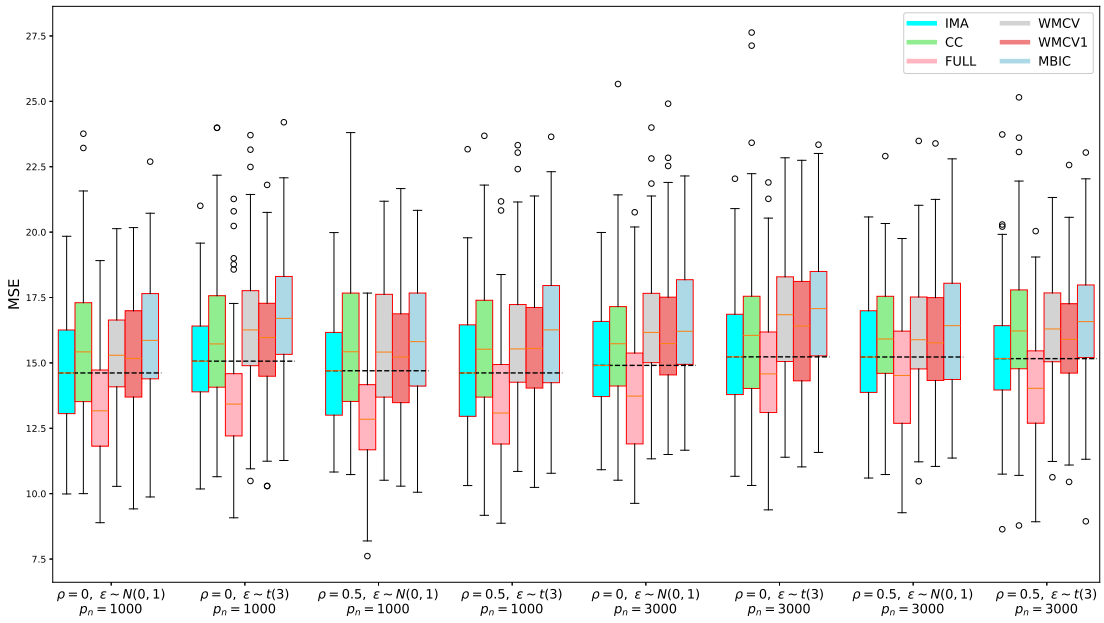

Figure 1. MSE values of six different methods in Experiment 4.1 under M1.

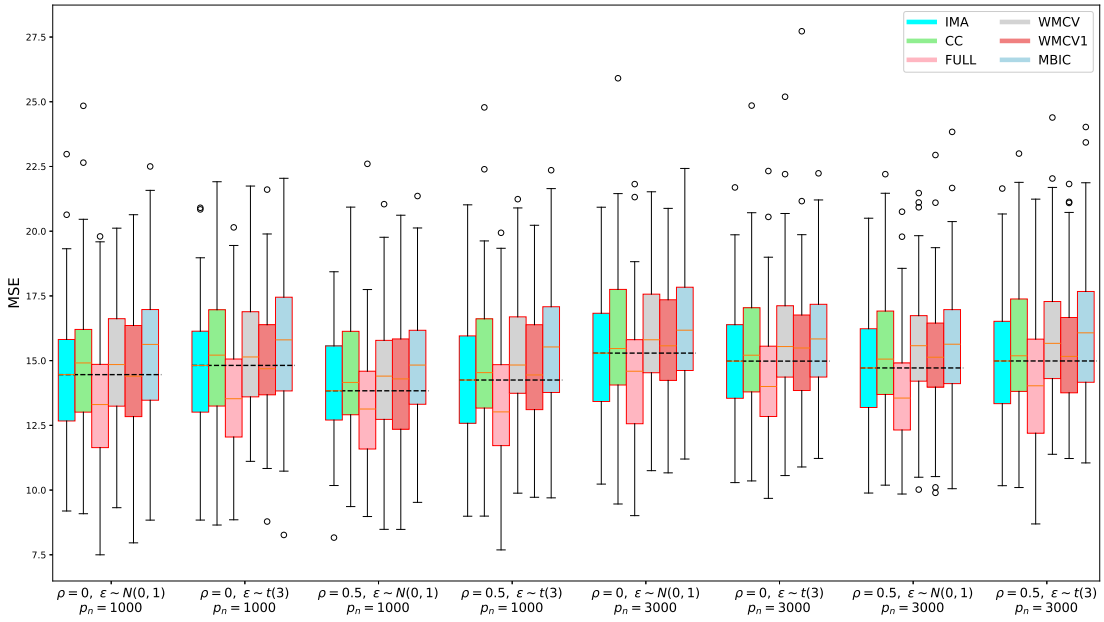

Figure 2. MSE values of six different methods in Experiment 4.1 under M2.

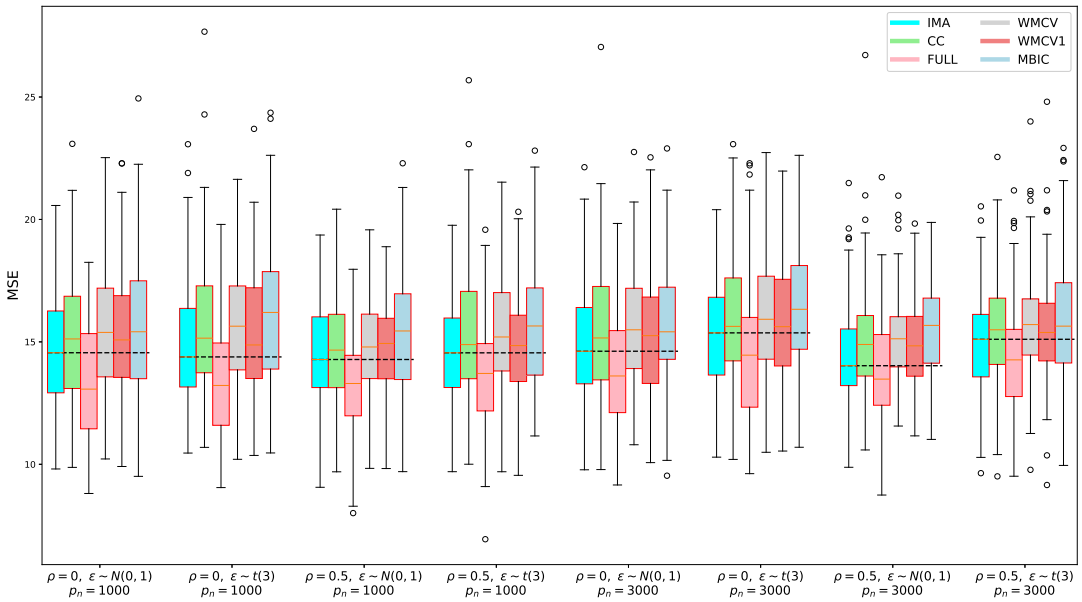

Figure 3. MSE values of six different methods in Experiment 4.1 under M3.

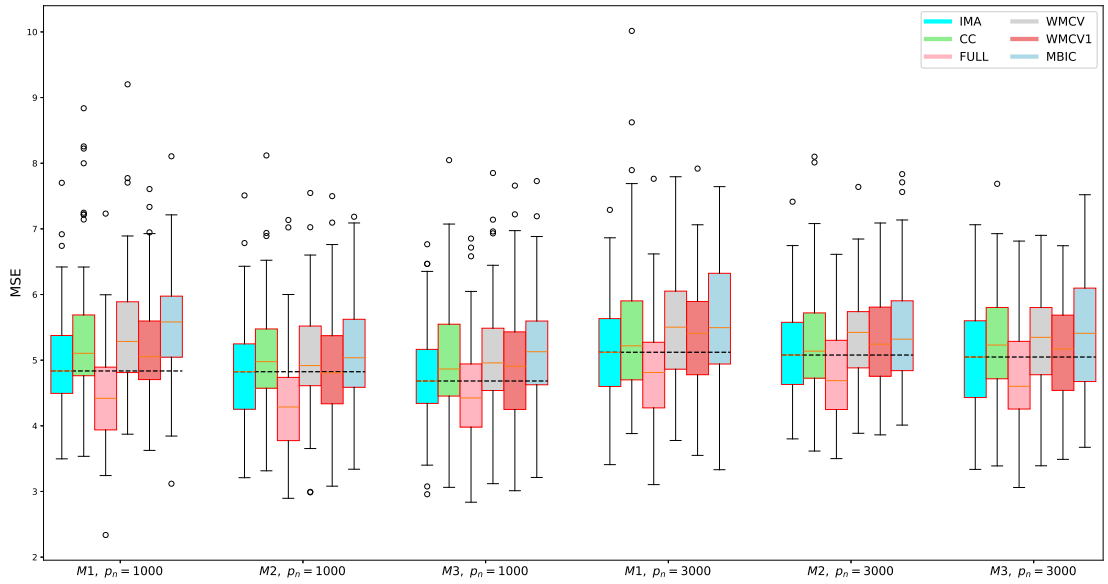

Figure 4. MSE values of six different methods in Experiment 4.2 (Gaussian kernel).

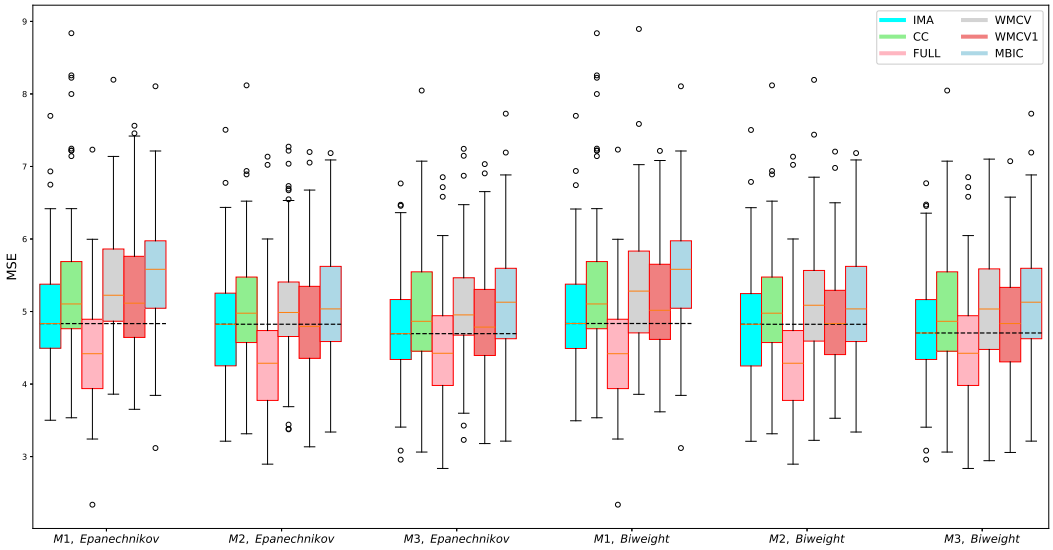

Figure 5. MSE values of six different methods in Experiment 4.2 (Epanechnikov kernel and biweight kernel).

A closer examination of Figures 1–5 has the following observations: (i) The proposed IMA procedure consistently outperforms other model averaging methods, including CC, WMCV, WMCV1, and MBIC, across most settings. This outcome underscores the effectiveness of the proposed multiple imputation techniques. (ii) As anticipated, the sequential model averaging method of "FULL" in [22] performs best across all scenarios, as it relies on completely observed data. (iii) The IMA procedure exhibits robust performance across various missing data mechanisms, indicating its relative insensitivity to changes in the missingness pattern. (iv) The proposed method shows reduced sensitivity to the sparsity of regression model coefficients. (v) The proposed kernel-assisted IMA approach exhibits robustness to the choice of kernel function. In conclusion, the IMA procedure consistently outperforms the competing methods, demonstrating the smallest median distribution of MSE values.

To further assess the impact of the covariate vector distribution on prediction accuracy, we consider the multivariate skew-normal distribution as studied in [31]. Specifically, the covariate vector $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{i,p_n})^\top$ is generated from the multivariate skew-normal distribution $\mathrm{SN}_{p_n}(\boldsymbol{\Sigma}, \boldsymbol{\alpha})$, where the entries of $\boldsymbol{\Sigma}$ are defined by $\mathrm{Cov}(x_{j_1}, x_{j_2}) = 0.5^{|j_1 - j_2|}$, for $1 \le j_1, j_2 \le p_n$, and $\boldsymbol{\alpha}$ is the shape parameter controlling the skewness of the distribution. When $\boldsymbol{\alpha} = \mathbf{0}$, the multivariate skew-normal distribution $\mathrm{SN}_{p_n}(\boldsymbol{\Sigma}, \boldsymbol{\alpha})$ reduces to the multivariate normal distribution $\mathrm{N}_{p_n}(\mathbf{0}, \boldsymbol{\Sigma})$. In contrast, when $\boldsymbol{\alpha} \neq \mathbf{0}$, the distribution remains skew-normal. The simulated MSE values from Experiment 4.2, where $\boldsymbol{\alpha} = (1, -1, 1, -1, \ldots, 1, -1)^\top$ and the Gaussian kernel function is used, are presented in Figure 6. A careful examination of Figure 6 reveals that the proposed method consistently outperforms WMCV, WMCV1, and MBIC across all scenarios. However, for the cases where $p_n = 1000$ and the missingness mechanisms are M2 and M3, the CC method appears to be comparable to the proposed IMA method. Nevertheless, for all other scenarios, the proposed estimation procedure outperforms the CC method. These findings suggest that the performance of the CC method is particularly sensitive to the missingness rate when the skew-normal distribution is considered, as mechanisms M2 and M3 exhibit relatively lower missing rates compared to M1. This further supports the efficiency and robustness of the proposed IMA method based on multiple imputation.

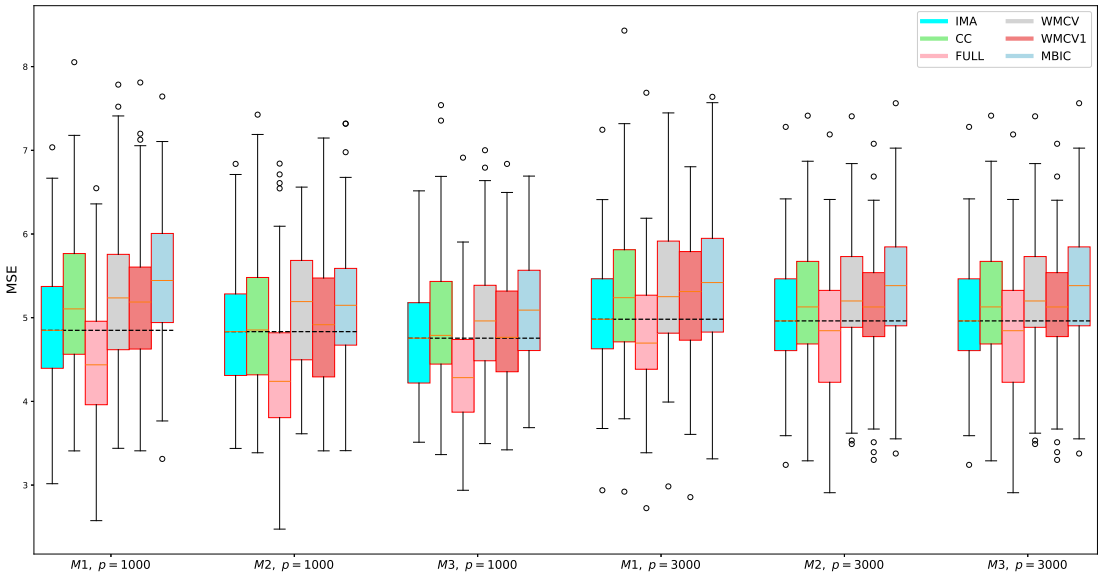

Figure 6. MSE values of six different methods in Experiment 4.2 ($\boldsymbol{X}_i \sim \mathrm{SN}_{p_n}(\boldsymbol{\Sigma}, \boldsymbol{\alpha})$).

All numerical experiments are conducted on a personal computer equipped with an Intel(R) Core(TM) i7-10875H CPU running at 2.30 GHz, featuring 8 physical cores and 16 threads. The machine is configured with 16.0 GB of RAM. The computations are performed using R software, version 4.4.1. This hardware and software configuration provides adequate computational power for implementing and evaluating the proposed algorithms, especially in the context of high-dimensional simulation and iterative model fitting. The average computation times for the six methods in Experiment 4.2 are presented in Table 2.

Table 2. Running times of six methods in Experiment 4.2 (in seconds).

$p_n = 1000$						$p_n = 3000$					
IMA	CC	FULL	WMCV	WMCV1	MBIC	IMA	CC	FULL	WMCV	WMCV1	MBIC
47.04	4.61	8.10	22.75	22.44	27.92	236.40	107.04	126.21	154.47	152.75	154.82

As shown in Table 2, the average computation time for the IMA method is 47.04 seconds for $p_n = 1000$, increasing to 236.40 seconds for $p_n = 3000$. While the computational time increases with p_n, these results highlight the robustness and scalability of the IMA method in high-dimensional modeling scenarios. Notably, when compared with other methods such as CC, WMCV, WMCV1, and MBIC, the IMA method achieves an optimal balance between computational efficiency and model performance, making it particularly suitable for ultrahigh-dimensional settings. Given its robustness, scalability, and competitive performance, the IMA method is recommended for applications that require precise model averaging and prediction in high-dimensional contexts.

Experiment 4.3. The primary aim of this experiment is to evaluate the effectiveness of the proposed stopping rule in ensuring robust predictive performance. For illustration, we focus on the setup from Experiment 4.2, where $n = 100$, $p_n = 1000$, and $d = 20$. Let $M = 100$, and define the stopping criterion at the m-th iteration as $\mathrm{eps}_m = \{\hat{\omega}_{(m+1)0} - \hat{\omega}_{m0}\}/\hat{\omega}_{0m}$, for $m = 1, \ldots, M$. Additionally, we examine the evolution of the coefficients by plotting $\max_j |\hat{\beta}_j^{(m)}|$ as a function of the iteration number, $m = 1, \ldots, M$, to

determine whether the estimated coefficients $\hat{\beta}_j^{(m)}$ become negligible as m increases. The corresponding plots of eps_m and $\max_j |\hat{\beta}_j^{(m)}|$ are presented in Figure 7 for the missingness mechanisms M1–M3.

Examination of Figure 7 reveals the following key observations: (i) The upper three panels demonstrate that the values of eps_m tend to stabilize for $m > 20$ across all three missingness mechanisms. From this, we conservatively conclude that updating the IMA procedure for up to 20 steps is sufficient to achieve reliable predictive performance for all mechanisms (M1–M3) in this experiment. (ii) The lower three panels show that $\max_j |\hat{\beta}_j^{(m)}|$ decreases rapidly toward zero for all missingness mechanisms, indicating that the influence of predictors diminishes significantly after approximately 20 steps.

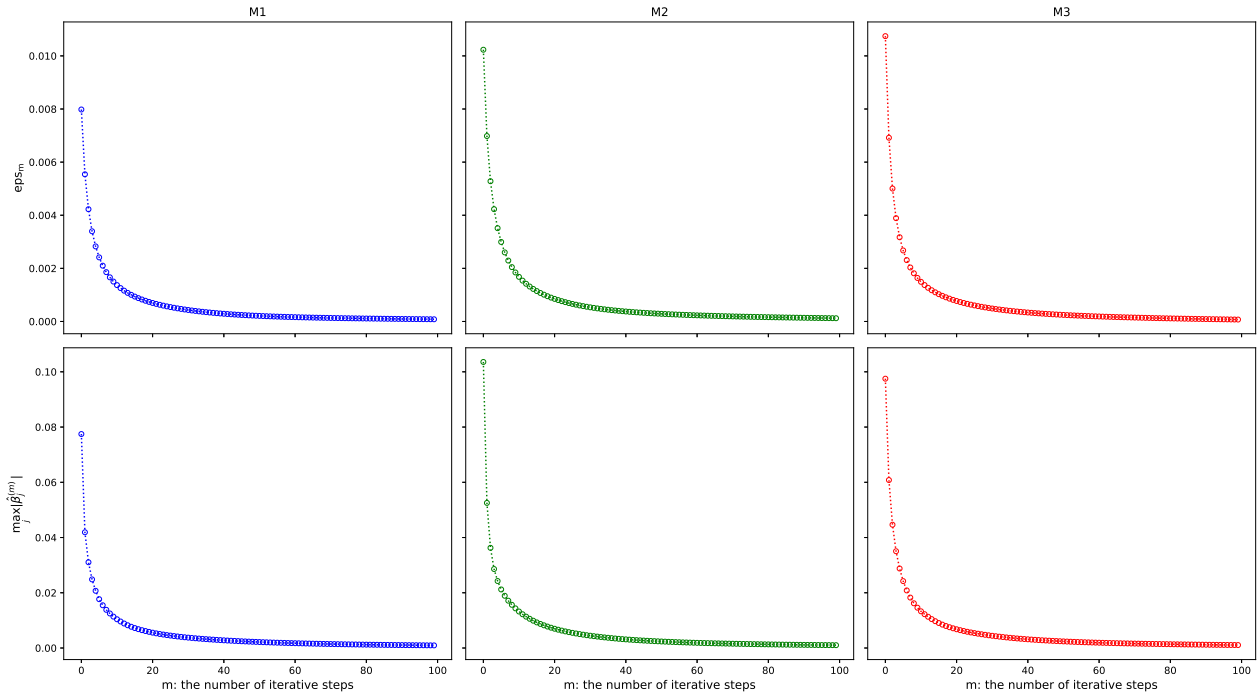

Figure 7. Values of eps_m and $\max_j |\hat{\beta}_j^{(m)}|$ versus the number of iterative numbers m for three settings of missingness mechanisms in Experiment 4.3.

In summary, these findings provide strong evidence that the proposed stopping rule ensures stable and accurate predictions after a relatively small number of iterations, thereby confirming the efficiency of the procedure.

5. Real data analysis

In this section, we illustrate the application of the proposed IMA procedure using a gene expression dataset related to Bardet-Biedl Syndrome [32]. This dataset comprises 120 twelve-week-old male rats, each characterized by 31,042 distinct probe sets, selected for tissue harvesting from the eyes and subsequent microarray analysis. Following the approaches of [32], 18,976 probes were deemed "sufficiently variable" based on expression quantitative trait locus (eQTL) mapping, exhibiting at least two-fold variation. For our analysis, all 18,976 probes were standardized to have zero mean and unit variance. Chiang et al. [33] identified the gene TRIM32 at probe 1389163_at as critical to Bardet-Biedl

Syndrome. In this study, the response variable y corresponds to probe 1389163_at, while the predictors x_j represent the remaining probes. The dataset presents a challenging scenario, as the sample size ($n = 120$) is small compared to the dimensionality ($p_n = 18{,}975$).

The primary goal of this analysis is to evaluate the prediction performance of the proposed IMA procedure in the presence of missing response data. To achieve this, we artificially introduce missingness under a MAR mechanism defined as

$$\text{logit}\{\pi(x_{6329}, \boldsymbol{\gamma})\} = \gamma_0 + \gamma_1 x_{6329},$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)^\top$, and γ_0 is the intercept term. The covariate x_{6329} is selected by ranking all 18,975 probes according to the absolute value of their marginal correlations with the expression of probe 1389163_at, retaining the top-ranked probe as the covariate for the missingness mechanism. The true parameter values are set as $\boldsymbol{\gamma} = (-0.5, 0.8)^\top$, resulting in an approximate missing proportion of 42%.

To assess the predictive performance of various methods, the complete dataset is randomly partitioned into a training set of size 80 and a validation set of size 40, with this process repeated 100 times. For implementing the proposed method, we adopt the Gaussian kernel function $K(u) = \exp(-u^2/2)/(2\pi)^{1/2}$ and perform $\mathcal{K} = 30$ multiple imputations for each missing response. The bandwidth parameter is determined using the `kedd` package in `R`, following the approach taken in Experiment 4.1. The proposed IMA method is then applied to the training set, and the predictive performance of the resulting model is evaluated by calculating the prediction error (PE) on the validation set for each of the 100 replications:

$$\text{PE} = \frac{1}{n_1} \sum_{i \in \mathcal{V}} \delta_i (y_i - \hat{\mu}_i)^2,$$

where $\hat{\mu}_i$ denotes the estimated mean of the response variable y_i, $n_1 = \sum_{i \in \mathcal{V}} \delta_i$, and $\mathcal{V}$ represents the set of indices corresponding to the observations in the validation set. A detailed summary of the dataset characteristics and experimental design is provided in Table 3.

Table 3. Summary of dataset features and experimental design.

Feature	Description
Study objective	Predict expression of TRIM32 (probe 1389163_at)
Organism	12-week-old male rats
Total number of samples	120
Original number of probe sets	31,042
Filtered probes after eQTL selection	18,976
Response variable (y)	Expression of probe 1389163_at
Predictors (x_j)	Remaining 18,975 probe expressions
Data standardization	Mean 0 and variance 1 for all predictors
Missing data mechanism	MAR: $\text{logit}(\pi) = \gamma_0 + \gamma_1 x_{6329}$
Missing rate in y	Approximately 42%
Training set size	80
Validation/Test set size	40
Number of repetitions	100
Number of imputations ($\mathcal{K}$)	30
Kernel function	Gaussian: $K(u) = \exp(-u^2/2)/\sqrt{2\pi}$

In this artificially generated missing-response scenario, we include an evaluation of four alternative methods, namely WMCV, WMCV1, MBIC, and CC, all of which were introduced and analyzed in the previous simulation studies. For the last three methods, genes are first ranked according to the marginal screening utility proposed by [23]. The top 200 genes are retained, and $M = 20$ candidate models are subsequently constructed, each containing 10 genes. The prediction errors on the validation data are illustrated in Figure 8. Examination of Figure 8 clearly demonstrates that the proposed IMA procedure achieves superior predictive efficiency compared to the other four methods, as evidenced by the smallest PE values.

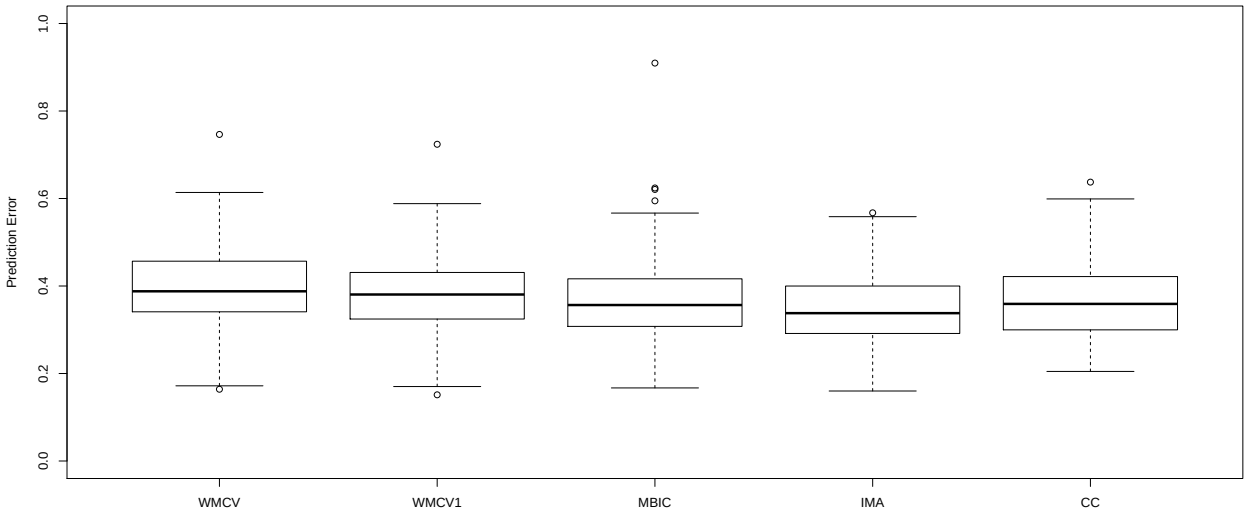

Figure 8. PE values of five different methods in the rat eye dataset.

6. Conclusions

This paper investigated the prediction problem in ultrahigh-dimensional linear regression models with missing responses under the assumption of missing at random. To address the missing response issue, we proposed an effective multiple-imputation procedure for handling the missing data. Additionally, we introduced an IMA procedure that combines iterative screening and model averaging techniques to enhance prediction accuracy. The proposed approach alleviates overfitting and provides consistent estimators for the regression coefficients, ensuring reliable and accurate predictions. Through simulation studies and a real data example, we validated the performance of the proposed IMA procedure. Our results demonstrated that the proposed method outperforms existing approaches, including conventional model averaging techniques, in terms of predictive accuracy and robustness.

7. Discussion and future work

While the current model averaging method assumes that missing responses follow a MAR mechanism, many real-world applications involve missing data mechanisms where the missingness is dependent on the unobserved values themselves, leading to a missing not at random (MNAR) scenario. In practical

applications, testing the assumption of MAR versus MNAR is crucial to ensure the applicability of our method. In ultrahigh-dimensional settings, the Pearson Chi-square test statistic developed by [34] can be employed to identify key features in the missingness data models. Subsequently, the score test techniques proposed by [35] can be utilized to assess the type of missingness mechanism. If the missingness mechanism is identified as MAR, the proposed multiple imputation and iterative model averaging approach can be applied to achieve robust statistical predictions. However, when the missingness mechanism is MNAR, it presents significant challenges for valid inference, including issues related to the identification of population parameters and the development of appropriate multiple imputation methods. Additionally, Condition 3 in Section 3 imposes a sub-exponential tail assumption on covariates to guarantee model stability and prevent overfitting. Relaxing this condition remains an important avenue for future research to enhance the generalizability of the IMA method to data with heavier tails. These considerations suggest potential directions for enhancing the current framework to better accommodate more complex missing data mechanisms and distributional settings.

Author contributions

Xianwen Ding: Writing–original draft, Formal analysis, Investigation, Methodology, Visualization, Software; Tong Su: Conceptualization, Writing–original draft, Formal analysis, Methodology, Project administration; Yunqi Zhang: Conceptualization, Funding acquisition, Validation, Resources, Writing–review & editing. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 12001244, 12426666, 12426668, 12426409), the Major Basic Research Project of the Natural Science Foundation of the Jiangsu Higher Education Institutions (grant No. 19KJB110007), the Yunnan Fundamental Research Projects (grant No. 202401AU070212), the Youth Project of Yunnan Xingdian Talent Program, and the Zhongwu Young Teachers Program for the Innovative Talents of Jiangsu University of Technology.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. R. Little, D. Rubin, *Statistical analysis with missing data*, 3 Eds., New York: John Wiley & Sons, 2019. https://doi.org/10.1002/9781119482260
2. J. K. Kim, J. Shao, *Statistical methods for handling incomplete data*, 2 Eds., New York: Chapman and Hall/CRC, 2021. https://doi.org/10.1201/9780429321740

3. W. M. Elmessery, A. Habib, M. Y. Shams, T. Abd El-Hafeez, T. M. El-Messery, S. Elsayed, et al., Deep regression analysis for enhanced thermal control in photovoltaic energy systems, *Sci. Rep.*, **14** (2024), 30600. https://doi.org/10.1038/s41598-024-81101-x
4. H. M. Farghaly, A. A. Ali, T. Abd El-Hafeez, Building an effective and accurate associative classifier based on support vector machine, *Sylwan*, **164** (2020), 39–56.
5. B. E. Hansen, Least squares model averaging, *Econometrica*, **75** (2007), 1175–1189. https://doi.org/10.1111/j.1468-0262.2007.00785.x
6. B. E. Hansen, J. S. Racine, Jackknife model averaging, *J. Econometrics*, **167** (2012), 38–46. https://doi.org/10.1016/j.jeconom.2011.06.019
7. Q. F. Liu, R. Okui, Heteroscedasticity-robust C_p model averaging, *Econom. J.*, **16** (2013), 463–472. https://doi.org/10.1111/ectj.12009
8. X. Y. Zhang, D. L. Yu, G. H. Zou, H. Liang, Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models, *J. Amer. Statist. Assoc.*, **111** (2016), 1775–1790. https://doi.org/10.1080/01621459.2015.1115762
9. M. Schomaker, A. T. K. Wan, C. Heumann, Frequentist model averaging with missing observations, *Comput. Statist. Data Anal.*, **54** (2010), 3336–3347. https://doi.org/10.1016/j.csda.2009.07.023
10. V. Dardanoni, S. Modica, F. Peracchi, Regression with imputed covariates: a generalized missing-indicator approach, *J. Econometrics*, **162** (2011), 362–368. https://doi.org/10.1016/j.jeconom.2011.02.005
11. X. Y. Zhang, Model averaging with covariates that are missing completely at random, *Econom. Lett.*, **121** (2013), 360–363. https://doi.org/10.1016/j.econlet.2013.09.008
12. F. Fang, W. Lan, J. J. Tong, J. Shao, Model averaging for prediction with fragmentary data, *J. Bus. Econom. Statist.*, **37** (2019), 517–527. https://doi.org/10.1080/07350015.2017.1383263
13. Z. Q. Liang, Q. H. Wang, A robust model averaging approach for partially linear models with responses missing at random, *Scand. J. Statist.*, **50** (2023), 1933–1952. https://doi.org/10.1111/sjos.12659
14. Z. Q. Liang, Y. Q. Zhou, Model averaging based on weighted generalized method of moments with missing responses, *AIMS Math.*, **8** (2023), 21683–21699. https://doi.org/10.3934/math.20231106
15. Z. Q. Liang, S. J. Wang, L. Cai, Optimal model averaging for partially linear models with missing response variables and error-prone covariates, *Stat. Med.*, **43** (2024), 4328–4348. https://doi.org/10.1002/sim.10176
16. X. Lu, L. J. Su, Jackknife model averaging for quantile regressions, *J. Econometrics*, **188** (2015), 40–58. https://doi.org/10.1016/j.jeconom.2014.11.005
17. X. Y. Zhang, G. H. Zou, H. Liang, R. J. Carroll, Parsimonious model averaging with a diverging number of parameters, *J. Amer. Statist. Assoc.*, **115** (2020), 972–984. https://doi.org/10.1080/01621459.2019.1604363
18. T. Ando, K. C. Li, A model-averaging approach for high-dimensional regression, *J. Amer. Statist. Assoc.*, **109** (2014), 254–265. https://doi.org/10.1080/01621459.2013.838168
19. T. Ando, K. C. Li, A weight-relaxed model averaging approach for high-dimensional generalized linear models, *Ann. Statist.*, **45** (2017), 2654–2679. https://doi.org/10.1214/17-AOS1538

20. X. Cheng, B. E. Hansen, Forecasting with factor-augmented regression: a frequentist model averaging approach, *J. Econometrics*, **186** (2015), 280–293. https://doi.org/10.1016/j.jeconom.2015.02.010

21. J. Chen, D. G. Li, O. Linton, Z. D. Lu, Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series, *J. Amer. Statist. Assoc.*, **113** (2018), 919–932. https://doi.org/10.1080/01621459.2017.1302339

22. W. Lan, Y. Y. Ma, J. L. Zhao, H. S. Wang, C. L. Tsai, Sequential model averaging for high dimensional linear regression models, *Statist. Sinica*, **28** (2018), 449–469. https://doi.org/10.5705/ss.202016.0122

23. J. H. Xie, X. D. Yan, N. S. Tang, A model-averaging method for high-dimensional regression with missing responses at random, *Statist. Sinica*, **31** (2021), 1005–1026. https://doi.org/10.5705/ss.202018.0297

24. X. M. He, L. Wang, H. G. Hong, Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, *Ann. Statist.*, **41** (2013), 342–369. https://doi.org/10.1214/13-AOS1087

25. C. Y. Tang, Y. S. Qin, An efficient empirical likelihood approach for estimating equations with missing data, *Biometrika*, **99** (2012), 1001–1007. https://doi.org/10.1093/biomet/ass045

26. X. R. Chen, A. T. K. Wan, Y. Zhou, Efficient quantile regression analysis with missing observations, *J. Amer. Statist. Assoc.*, **110** (2015), 723–741. https://doi.org/10.1080/01621459.2014.928219

27. J. Y. Liu, R. Z. Li, R. L. Wu, Feature selection for varying coefficient models with ultrahigh-dimensional covariates, *J. Amer. Statist. Assoc.*, **109** (2014), 266–274. https://doi.org/10.1080/01621459.2013.850086

28. M. Kalisch, P. Bühlmann, Estimating high-dimensional directed acyclic graphs with the PC-algorithm, *J. Mach. Learn. Res.*, **8** (2007), 613–636. https://doi.org/10.5555/1314498.1314520

29. H. S. Wang, Forward regression for ultra-high dimensional variable screening, *J. Amer. Statist. Assoc.*, **104** (2009), 1512–1524. https://doi.org/10.1198/jasa.2008.tm08516

30. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B Methodol.*, **58** (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

31. A. Azzalini, A. Capitanio, Statistical applications of the multivariate skew normal distribution, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **61** (1999), 579–602. https://doi.org/10.1111/1467-9868.00194

32. T. E. Scheetz, K. Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L. Knudtson, et al., Regulation of gene expression in the mammalian eye and its relevance to eye disease, *Proc. Natl. Acad. Sci.*, **103** (2006), 14429–14434. https://doi.org/10.1073/pnas.0602562103

33. A. P. Chiang, J. S. Beck, H. J. Yen, M. K. Tayeh, T. E. Scheetz, R. E. Swiderski, et al., Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11), *Proc. Natl. Acad. Sci.*, **103** (2006), 6287–6292. https://doi.org/10.1073/pnas.0600158103

34. X. W. Ding, J. D. Chen, X. P. Chen, Regularized quantile regression for ultrahigh-dimensional data with nonignorable missing responses, *Metrika*, **83** (2020), 545–568. https://doi.org/10.1007/s00184-019-00744-3

35. H. R. Wang, Z. P. Lu, Y. K. Liu, Score test for missing at random or not under logistic missingness models, *Biometrics*, **79** (2023), 1268–1279. https://doi.org/10.1111/biom.13666

36. D. Wang, S. X. Chen, Empirical likelihood for estimating equations with missing values, *Ann. Statist.*, **37** (2009), 490–517. https://doi.org/10.1214/07-AOS585

37. B. Y. Jiang, Covariance selection by thresholding the sample correlation matrix, *Statist. Probab. Lett.*, **83** (2013), 2492–2498. https://doi.org/10.1016/j.spl.2013.07.008

38. P. J. Bickel, E. Levina, Covariance regularization by thresholding, *Ann. Statist.*, **36** (2008), 2577–2604. https://doi.org/10.1214/08-AOS600

Appendix

Lemma A.1. *Under Conditions 1–3, for any $\nu \in (0, 1/2)$, constant $c_1 > 0$, and $1 \le j \le p_n$, there exist some positive constants c_2 and c_3 such that*

$$\Pr\left(|\hat{\beta}_j - \beta_{0j}| > c_1 n^{-\nu}\right) \le c_3 n \exp\{-c_2 n^{(1-2\nu)/3}\}.$$

Proof of Lemma A.1. Let $h_j(x) = \mathbb{E}(x_j y | x_j = x)$, and $\hat{h}_j(x) = \sum_{l=1}^n \delta_l K_h(x - x_{lj}) x y_l / \sum_{l=1}^n \delta_l K_h(x - x_{lj})$ for each $j = 1, \ldots, p_n$. Then

$$\begin{aligned}\Pr\left(|\hat{\beta}_j - \beta_{0j}| > c_1 n^{-\nu}\right) &\le \Pr\left\{\left|\frac{1}{n}\sum_{i=1}^n x_{ij}\hat{y}_i - \mathbb{E}(x_j y)\right| \ge c_1 n^{-\nu}\right\} \\ &= \Pr\left(|T_{n1} + T_{n2} + T_{n3} + T_{n4}| \ge c_1 n^{-\nu}\right),\end{aligned}$$

where

$$\begin{aligned}T_{n1} &= \frac{1}{np_n}\sum_{i=1}^n \sum_{l=1}^{p_n}(1-\delta_i)\left\{\frac{1}{\mathcal{K}}\sum_{k=1}^{\mathcal{K}} x_{ij}\tilde{y}_{ik}^{(j)} - \hat{h}_j(x_{ij})\right\}, \\ T_{n2} &= \frac{1}{n}\sum_{i=1}^n (1-\delta_i)\{\hat{h}_j(x_{ij}) - h_j(x_{ij})\}, \\ T_{n3} &= \frac{1}{n}\sum_{i=1}^n \delta_i\{x_{ij}y_i - h_j(x_{ij})\}, \\ T_{n4} &= \frac{1}{n}\sum_{i=1}^n \{h_j(x_{ij}) - \mathbb{E}(x_j y)\}.\end{aligned}$$

Under the assumption $y \perp \delta \,|\, x_j$ for each $j = 1, \ldots, p_n$, it follows from the proof of Lemma A.1 in [36] that

$$T_{n1} = \frac{1}{n}\sum_{i=1}^n (1-\delta_i)\left\{\frac{1}{\mathcal{K}}\sum_{k=1}^{\mathcal{K}} x_{ij}\tilde{y}_{ik}^{(j)} - \hat{h}_j(x_{ij})\right\} = O_p(n^{-1/2}).$$

Also,

$$T_{n2} = \frac{1}{n}\sum_{i=1}^n (1-\delta_i)\{\hat{h}_j(x_{ij}) - h_j(x_{ij})\}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{\{x_{ij}y_i - h_j(x_{ij})\}\{1-\pi(x_{ij})\}}{\pi(x_{ij})} + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{\sum_{l=1}^{n}\delta_l K_h(x_{lj}-x_{ij})\{x_{ij}y_l - h_j(x_{lj})\}/n}{\eta_j(x_{ij})} \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\{x_{ij}y_i - h_j(x_{ij})\}\{1-\pi(x_{ij})\}}{\pi(x_{ij})} + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{\sum_{l=1}^{n}\delta_l K_h(x_{lj}-x_{ij})\{h_j(x_{lj}) - h_j(x_{ij})\}/n}{\eta_j(x_{ij})} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{\{\hat{h}_j(x_{ij}) - h_j(x_{ij})\}\{\eta_j(x_{ij}) - \hat{\eta}_j(x_{ij})\}}{\eta_j(x_{ij})} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\{x_{ij}y_i - h_j(x_{ij})\}\{1-\pi(x_{ij})\}}{\pi(x_{ij})} + O_p(n^{-1/2}) \\
&= \tilde{T}_{n2} + O_p(n^{-1/2}),
\end{aligned}
$$

where $\eta_j(x) = \pi(x)f_j(x)$ and $\hat{\eta}_j(x) = \sum_{l=1}^{n}\delta_l K_h(x_{lj}-x)/n$. Then

$$
\begin{aligned}
&\Pr(|T_{n1} + T_{n2} + T_{n3} + T_{n4}| \geq c_1 n^{-\nu}) \\
&\leq \Pr(|T_{n2} + T_{n3} + T_{n4}| \geq c_1 n^{-\nu}/2) \\
&\leq \Pr(|\tilde{T}_{n2} + T_{n3} + T_{n4}| \geq c_1 n^{-\nu}/4) \\
&\leq \Pr\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(x_{ij})}\{x_{ij}y_i - \mathbb{E}(x_j y)\} \geq c_1 n^{-\nu}/8\right] \\
&\quad + \Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left\{1 - \frac{\delta_i}{\pi(x_{ij})}\right\}[h_j(x_{ij}) - \mathbb{E}\{h_j(x_{ij})\}] \geq c_1 n^{-\nu}/8\right) \\
&= J_1 + J_2.
\end{aligned}
$$

According to Lemma S3 of [27], under Condition 2, we have

$$
\Pr(\max_i |\delta_i x_{ij} y_i| \geq C') \leq n\Pr(|\delta_i x_{ij} y_i| \geq C') \leq nc_4\exp(-c_5 C'),
$$

where $C' > 0$ is any positive constant, and c_4 and c_5 are some positive constants. Then, under Condition 1, by taking $C' = c_6 n^{(1-2\nu)/3}$ for some positive constant c_6, we obtain

$$
\begin{aligned}
J_1 &= \Pr\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(x_{ij})}\{x_{ij}y_i - \mathbb{E}(x_j y)\} \geq c_1 n^{-\nu}/8\right] \\
&\leq \Pr\left[\frac{1}{n}\sum_{i=1}^{n}\delta_i\{x_{ij}y_i - \mathbb{E}(x_j y)\} \geq C_0 c_1 n^{-\nu}/8\right] \\
&\leq \Pr\left[\frac{1}{n}\sum_{i=1}^{n}\delta_i\{x_{ij}y_i - \mathbb{E}(x_j y)\} \geq C_0 c_1 n^{-\nu}/8, \quad \max_i |\delta_i x_{ij} y_i| < C'\right] + \Pr\left(\max_i |\delta_i x_{ij} y_i| \geq C'\right) \\
&\leq 2\exp\left(-c_7 n^{1-2\nu}/C'^2\right) + nc_4\exp\left(-c_5 C'\right) \\
&\leq c_9 n\exp\left\{-c_8 n^{(1-2\nu)/3}\right\},
\end{aligned}
$$

where c_7, c_8, and c_9 are some positive constants, and the last inequality holds due to Hoeffding's inequality. By some similar arguments, it can be shown that J_2 follows a similar bound. Hence, we complete the proof of this lemma. □

Lemma A.2. *Under Conditions 1–4, for any $\nu \in (0, 1/2)$ and constant $c_1' > 0$, there exist positive constants c_2' and c_3' such that*

$$\Pr\left(\max_j |\hat{\beta}_j^2 - \beta_{0j}^2| > c_1' n^{-\nu}\right) \leq c_3' n p_n \exp\left\{-c_2' n^{(1-2\nu)/3}\right\}.$$

Proof of Lemma A.2. For any $\nu \in (0, 1/2)$ and constant c_4', since $\max_j |\beta_{0j}| \leq C_2$, then there exist positive constants c_5' and c_6' such that

$$\begin{aligned}\Pr\left(\max_j |\hat{\beta}_j| \geq C_2 + c_4' n^{-\nu}\right) &\leq p_n \Pr\left\{|\hat{\beta}_j - \beta_{0j}| + |\beta_{0j}| \geq C_2 + c_4' n^{-\nu}\right\} \\ &\leq p_n \Pr\left\{|\hat{\beta}_j - \beta_{0j}| \geq c_4' n^{-\nu}\right\} \\ &\leq c_5' p_n n \exp\left\{-c_6' n^{(1-2\nu)/3}\right\},\end{aligned}$$

where the last inequality holds due to Lemma A.1. This shows that $\max_j |\hat{\beta}_j|$ is bounded in probability. For each $j = 1, \ldots, p_n$, notice that

$$\begin{aligned}\Pr\left(|\hat{\beta}_j^2 - \beta_{0j}^2| \geq c_1' n^{-\nu}\right) &\leq \Pr\left(|\hat{\beta}_j| \cdot |\hat{\beta}_j - \beta_{0j}| + |\beta_{0j}| \cdot |\hat{\beta}_j - \beta_{0j}| \geq c_1' n^{-\nu}\right) \\ &\leq \Pr\left(|\hat{\beta}_j| \cdot |\hat{\beta}_j - \beta_{0j}| \geq c_1' n^{-\nu}/2\right) + \Pr\left(|\beta_{0j}| \cdot |\hat{\beta}_j - \beta_{0j}| \geq c_1' n^{-\nu}/2\right).\end{aligned}$$

For the first term, we have

$$\begin{aligned}\Pr\left(|\hat{\beta}_j| \cdot |\hat{\beta}_j - \beta_{0j}| \geq c_1' n^{-\nu}/2\right) &= \Pr\left(|\hat{\beta}_j| \cdot |\hat{\beta}_j - \beta_{0j}| \geq c_1' n^{-\nu}/2, \quad |\hat{\beta}_j| \geq C_2 + c_4' n^{-\nu}\right) \\ &\quad + \Pr\left(|\hat{\beta}_j| \cdot |\hat{\beta}_j - \beta_{0j}| \geq c_1' n^{-\nu}/2, \quad |\hat{\beta}_j| < C_2 + c_4' n^{-\nu}\right) \\ &\leq \Pr\left(|\hat{\beta}_j| \geq C_2 + c_4' n^{-\nu}\right) + \Pr\left\{(C_2 + c_4' n^{-\nu})|\hat{\beta}_j - \beta_{0j}| > c_1' n^{-\nu}/2\right\} \\ &\leq c_7' n \exp\left\{-c_8 n^{(1-2\nu)/3}\right\},\end{aligned}$$

where c_7' and c_8' are positive constants. We next deal with the second term:

$$\begin{aligned}\Pr\left(|\beta_{0j}| \cdot |\hat{\beta}_j - \beta_{0j}| \geq c_1' n^{-\nu}/2\right) &\leq \Pr\left\{|\hat{\beta}_j - \beta_{0j}| \geq c_1' n^{-\nu}/(2C_2)\right\} \\ &\leq c_9' n \exp\left\{-c_{10}' n^{(1-2\nu)/3}\right\},\end{aligned}$$

where c_9' and c_{10}' are some positive constants. Let $c_3' = c_7' + c_9'$ and $c_2' = \min\{2c_8', 2c_{10}'\}$. Hence,

$$\begin{aligned}\Pr\left(\max_j |\hat{\beta}_j^2 - \beta_{0j}^2| > c_1' n^{-\nu}\right) &\leq p_n \Pr\left\{|\hat{\beta}_j^2 - \beta_{0j}^2| > c_1' n^{-\nu}\right\} \\ &\leq c_3' n \exp\left\{-c_2' n^{(1-2\nu)/3}\right\}.\end{aligned}$$

This completes the proof. □

Proof of Theorem 3.1. By the definition of $\hat{\omega}_j$, we have

$$\hat{\omega}_j = \frac{\exp(-\mathrm{BIC}_j/2)}{\sum_{j=0}^{p_n} \exp(-\mathrm{BIC}_j/2)} = \frac{(\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_j^2)^{-n/2}/\sqrt{n}p_n}{(\|\hat{\boldsymbol{Y}}\|^2)^{-n/2} + (\sum_{j=1}^{p_n} \|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_j^2)^{-n/2}/\sqrt{n}p_n}.$$

It can be shown that $\hat{\omega}_j$ is a monotone increasing function of $\hat{\beta}_j^2$. For the sake of convenience, let $\mathrm{BIC}_{(1)}$ be the BIC score associated with $\hat{\beta}_{(1)}^2$. Then we have

$$\hat{\omega}_{(1)}^U = \frac{\exp(-\mathrm{BIC}_{(1)}/2)}{\sum_{j=0}^{p_n}\exp(-\mathrm{BIC}_j)} = \left[1 + \sum_{j\geq 2}(\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(j)}^2)^{-n/2}(\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(1)}^2)^{n/2} + \sqrt{n}p_n(1 - n\hat{\beta}_{(1)}^2/\|\hat{\boldsymbol{Y}}\|^2)^{n/2}\right]^{-1}.$$

To show $\hat{\omega}_{(1)}^U \to_p 1$, it suffices to show

$$\sum_{j\geq 2}(\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(j)}^2)^{-n/2}(\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(1)}^2)^{n/2} \to_p 0, \tag{A.1}$$

$$\sqrt{n}p_n(1 - n\hat{\beta}_{(1)}^2/\|\hat{\boldsymbol{Y}}\|^2)^{n/2} \to_p 0. \tag{A.2}$$

First, for Eq (A.1), notice that

$$\begin{aligned}\sum_{j\geq 2}(\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(j)}^2)^{-n/2}(\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(1)}^2)^{n/2} &\leq p_n(\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(1)}^2)^{n/2}(\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(2)}^2)^{-n/2}\\ &= \exp\left[\log(p_n) + \frac{n}{2}\log\left\{\frac{\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(1)}^2}{\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(2)}^2}\right\}\right].\end{aligned}$$

Furthermore, by condition $\beta_{(1)}^2 - \beta_{(2)}^2 > C_4$, it is noteworthy that

$$\begin{aligned}\hat{\beta}_{(1)}^2 - \hat{\beta}_{(2)}^2 &\geq \beta_{(1)}^2 - \beta_{(2)}^2 - |\hat{\beta}_{(1)}^2 - \beta_{(1)}^2| - |\hat{\beta}_{(2)}^2 - \beta_{(2)}^2|\\ &\geq C_4 - 2\max_j|\hat{\beta}_j^2 - \beta_{0j}^2|.\end{aligned} \tag{A.3}$$

By Lemma A.2, we know that $\Pr(\max_j|\hat{\beta}_j^2 - \beta_{0j}^2| \geq c_1' n^{-\nu}) \to_p 0$. This, together with Eq (A.3), implies that with probability tending to one, we have $\hat{\beta}_{(1)}^2 - \hat{\beta}_{(2)}^2 > C_4/2$. From the proof of Lemma A.2, we know $\max_j|\hat{\beta}_j|$ is bounded in probability. Under Condition 2 and the MAR assumption, for any $C'' > 0$, there exist positive constants b_1 and b_2 such that

$$\begin{aligned}\Pr(\|\hat{\boldsymbol{Y}}\|^2/n < C'') &= \Pr(\|\boldsymbol{Y}\|^2/n < C'',\ \ \delta = 1) + \Pr(\|\hat{m}(\boldsymbol{X})\|^2/n < C'',\ \ \delta = 0)\\ &\geq 1 - \Pr(\|\boldsymbol{Y}\|^2/n \geq C'',\ \ \delta = 1)\\ &= 1 - \mathbb{E}[\mathbb{E}\{I(\|\boldsymbol{Y}\|^2 \geq C''n, \delta = 1)|\boldsymbol{X}, \boldsymbol{Y}\}]\\ &= 1 - \mathbb{E}\{I(\|\boldsymbol{Y}\|^2 \geq C''n)\Pr(\delta = 1|\boldsymbol{X})\}\\ &\geq 1 - \Pr(\|\boldsymbol{Y}\|^2 \geq C''n)\\ &\geq 1 - b_1\exp(-b_2C''n),\end{aligned} \tag{A.4}$$

where $\hat{m}(\boldsymbol{X}) = \sum_{j=1}^{p_n}\sum_{k=1}^{\mathcal{K}}\tilde{y}_{ik}^{(j)}/(p_n\mathcal{K})$, and the last inequality holds due to the Markov inequality. Then $\|\hat{\boldsymbol{Y}}\|^2/n$ is bounded in probability. Thus, under Condition 4, it follows that $n\hat{\beta}_{(1)}^2/\|\hat{\boldsymbol{Y}}\|^2 < C_3 < 1$. Combining Eq (A.3) and Condition 4, we have

$$\exp\left[\log(p_n) + \frac{n}{2}\log\left\{\frac{\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(1)}^2}{\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(2)}^2}\right\}\right] \leq \exp\left[\log(p_n) + \frac{n}{2}\log\left\{\frac{\|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(1)}^2}{C_4/2 + \|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(1)}^2}\right\}\right]$$

$$= \exp\left[\log(p_n) + \frac{n}{2}\log\left\{1 - \frac{C_4/2}{C_4/2 + \|\hat{\boldsymbol{Y}}\|^2 - n\hat{\beta}_{(1)}^2}\right\}\right]$$
$$\to_p 0.$$

Analogously, we can prove that Eq (A.2) holds. As a result, $\hat{\omega}_{(1)}^U \to_p 1$. □

Proof of Theorem 3.2. Let $\mathbf{H}_j = \boldsymbol{x}_j\boldsymbol{x}_j^\top/n$. Notice that

$$\hat{\boldsymbol{Y}}^{(m+1)} = \hat{\boldsymbol{Y}}^{(m)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{Y}}^{(m)} - \sum_{j=1}^{p_n}\hat{\omega}_{mj}\mathbf{H}_j\hat{\boldsymbol{Y}}^{(m)},$$

where $\hat{\boldsymbol{\beta}}^{(m)} = (\hat{\omega}_{m1}\hat{\beta}_{m1}, \ldots, \hat{\omega}_{mp_n}\hat{\beta}_{mp_n})^\top$. Then we have

$$\begin{aligned}
\|\hat{\boldsymbol{Y}}^{(m+1)}\|^2 &= \|\hat{\boldsymbol{Y}}^{(m)}\|^2 + \|\sum_{j=1}^{p_n}\hat{\omega}_{mj}\mathbf{H}_j\hat{\boldsymbol{Y}}^{(m)}\|^2 - 2\sum_{j=1}^{p_n}\hat{\omega}_{mj}(\hat{\boldsymbol{Y}}^{(m)})^\top\mathbf{H}_j\hat{\boldsymbol{Y}}^{(m)} \\
&= \|\hat{\boldsymbol{Y}}^{(m)}\|^2 + \|\sum_{j=1}^{p_n}\hat{\omega}_{mj}\mathbf{H}_j\hat{\boldsymbol{Y}}^{(m)}\|^2 - 2\|\hat{\boldsymbol{Y}}^{(m)}\|^2\sum_{j=1}^{p_n}n\hat{\omega}_{mj}\hat{\beta}_{mj}^2/\|\hat{\boldsymbol{Y}}^{(m)}\|^2 \\
&= \|\hat{\boldsymbol{Y}}^{(m)}\|^2\left(1 - 2\sum_{j=1}^{p_n}n\hat{\omega}_{mj}\hat{\beta}_{mj}^2/\|\hat{\boldsymbol{Y}}^{(m)}\|^2 + \sum_{1\le j_1,j_2\le p_n}\hat{\omega}_{mj_1}\hat{\omega}_{mj_2}\hat{\beta}_{mj_1}\hat{\beta}_{mj_2}\boldsymbol{x}_{j_1}^\top\boldsymbol{x}_{j_2}/\|\hat{\boldsymbol{Y}}^{(m)}\|^2\right).
\end{aligned}$$

Note that $|\boldsymbol{x}_{j_1}^\top\boldsymbol{x}_{j_2}| \le n$ for any $1 \le j_1, j_2 \le p_n$ due to the fact that $\|\boldsymbol{x}_j\|^2 = n$. This suggests that

$$\begin{aligned}
&\|\hat{\boldsymbol{Y}}^{(m)}\|^2\left(1 - 2\sum_{j=1}^{p_n}n\hat{\omega}_{mj}\hat{\beta}_{mj}^2/\|\hat{\boldsymbol{Y}}^{(m)}\|^2 + \sum_{1\le j_1,j_2\le p_n}\hat{\omega}_{mj_1}\hat{\omega}_{mj_2}\hat{\beta}_{mj_1}\hat{\beta}_{mj_2}\boldsymbol{x}_{j_1}^\top\boldsymbol{x}_{j_2}/\|\hat{\boldsymbol{Y}}^{(m)}\|^2\right) \\
\le\ &\|\hat{\boldsymbol{Y}}^{(m)}\|^2\left\{1 - 2\sum_{j=1}^{p_n}n\hat{\omega}_{mj}\hat{\beta}_{mj}^2/\|\hat{\boldsymbol{Y}}^{(m)}\|^2 + n\left(\sum_{j=1}^{p_n}\hat{\omega}_{mj}|\hat{\beta}_{mj}|\right)^2/\|\hat{\boldsymbol{Y}}^{(m)}\|^2\right\}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\|\hat{\boldsymbol{Y}}^{(m)}\|^2 - \|\hat{\boldsymbol{Y}}^{(m+1)}\|^2 &\ge 2\sum_{j=1}^{p_n}n\hat{\omega}_{mj}\hat{\beta}_{mj}^2 - n\left(\sum_{j=1}^{p_n}\hat{\omega}_{mj}|\hat{\beta}_{mj}|\right)^2 \\
&\ge 2\sum_{j=1}^{p_n}n\hat{\omega}_{mj}\hat{\beta}_{mj}^2 - n\left(\sum_{j=1}^{p_n}\hat{\omega}_{mj}\right)\left(\sum_{j=1}^{p_n}\hat{\omega}_{mj}\hat{\beta}_{mj}^2\right) \\
&\ge \sum_{j=1}^{p_n}n\hat{\omega}_{mj}\hat{\beta}_{mj}^2.
\end{aligned}$$

Now, by the definition of $\hat{\boldsymbol{Y}}^{(m+1)}$, we have

$$\|\hat{\boldsymbol{Y}}^{(m)}\|^2 - \|\hat{\boldsymbol{Y}}^{(m+1)}\|^2 = 2\sum_{j=1}^{p_n}n\hat{\omega}_{mj}\hat{\beta}_{mj}^2 - \left\|\sum_{j=1}^{p_n}\hat{\omega}_{mj}X_j\hat{\beta}_{mj}\right\|^2 \le 2\sum_{j=1}^{p_n}n\hat{\omega}_{mj}\hat{\beta}_{mj}^2.$$

Notice that $\hat{\beta}_{mj}^2 \le \hat{\beta}_{m(1)}^2$ for each $j = 1, \ldots, p_n$. Then

$$\|\hat{\boldsymbol{Y}}^{(m)}\|^2 - \|\hat{\boldsymbol{Y}}^{(m+1)}\|^2 \le 2\sum_{j=1}^{p_n} n\hat{\omega}_{mj}\hat{\beta}_{mj}^2 \le 2\sum_{j=1}^{p_n} n\hat{\omega}_{mj}\hat{\beta}_{m(1)}^2 = 2n(1-\hat{\omega}_{m0})\hat{\beta}_{m(1)}^2. \quad \square$$

Lemma A.3. *Under Conditions 1–4, for every $j \in \{1, 2, \ldots, p_n\}$, we have $\max_j |\hat{\beta}_{mj}^2 - \beta_{mj}^2| \to_p 0$ for any $m = 1, \ldots, M$.*

Proof of Lemma A.3. We can demonstrate this lemma for general m by induction. For the sake of simplicity, we can only show that the result is valid for $m = 2$ by assuming that it holds when $m = 1$. Notice that the result of $m = 1$ can be directly derived from Lemma A.2. Recall that $\boldsymbol{\beta}_0^{(1)} = (\omega_{11}\beta_{11}, \ldots, \omega_{1p_n}\beta_{1p_n})^\top \in \mathbb{R}^{p_n}$ and $\hat{\boldsymbol{\beta}}^{(1)} = (\hat{\omega}_{11}\hat{\beta}_{11}, \ldots, \hat{\omega}_{1p_n}\hat{\beta}_{1p_n})^\top \in \mathbb{R}^{p_n}$. From the definition of $\hat{\beta}_{2j}$, we have

$$\hat{\beta}_{2j} = \frac{1}{n}\boldsymbol{x}_j^\top \hat{\boldsymbol{Y}}^{(2)} = \frac{1}{n}\boldsymbol{x}_j^\top(\hat{\boldsymbol{Y}}^{(1)} - \mathbf{X}\boldsymbol{\beta}_0^{(1)}) + \frac{1}{n}\boldsymbol{x}_j^\top \mathbf{X}(\boldsymbol{\beta}_0^{(1)} - \hat{\boldsymbol{\beta}}^{(1)}).$$

From the proof of Lemma A.1, to prove $\hat{\beta}_{2j} \to_p \beta_{2j}$ for every $j \in \{1, 2, \ldots, p_n\}$, it suffices to show the following result:

$$\max_j \left|\frac{1}{n}\boldsymbol{x}_j^\top \mathbf{X}(\boldsymbol{\beta}_0^{(1)} - \hat{\boldsymbol{\beta}}^{(1)})\right| = o_p(1).$$

By Hölder's inequality, we have

$$\max_j \left|\frac{1}{n}\boldsymbol{x}_j^\top \mathbf{X}(\boldsymbol{\beta}_0^{(1)} - \hat{\boldsymbol{\beta}}^{(1)})\right| \le \max_{j_1, j_2} |\hat{\sigma}_{j_1 j_2}| \cdot |\boldsymbol{\beta}_0^{(1)} - \hat{\boldsymbol{\beta}}^{(1)}|_1,$$

where $\hat{\sigma}_{j_1 j_2} = \boldsymbol{x}_{j_1}^\top \boldsymbol{x}_{j_2}/(\|\boldsymbol{x}_{j_1}\| \, \|\boldsymbol{x}_{j_2}\|)$. By the results of Proposition 1 in [37], under Condition 2, we obtain that $\max_{j_1, j_2} |\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}| \to_p 0$. We have $\max_{j_1, j_2} |\hat{\sigma}_{j_1 j_2}| = O_p(1)$. Let $\omega_{1(1)} \ge \omega_{1(2)} \ge \ldots \ge \omega_{1(p_n)}$ be the ordered statistics of $\{\omega_{1j}, 1 \le j \le p_n\}$ and $\hat{\omega}_{1(j)}$ be the corresponding estimators for $j = 1, 2, \ldots, p_n$. By the result of Theorem 3.1, there exists a positive constant $\xi < 1$ such that

$$\hat{\omega}_{1(1)} \to_p 1, \quad \hat{\omega}_{1(j)} \le \xi^n, \quad \omega_{1(1)} \to_p 1, \quad \text{and} \quad \omega_{1(j)} \le \xi^n.$$

Thus, we have

$$|\boldsymbol{\beta}_0^{(1)} - \hat{\boldsymbol{\beta}}^{(1)}|_1 \le |\beta_{0(1)} - \hat{\beta}_{(1)}| + (p_n - 1)\xi^n.$$

By the result of Lemma A.1, we have $|\boldsymbol{\beta}_0^{(1)} - \hat{\boldsymbol{\beta}}^{(1)}|_1 \to_p 0$. Hence, $\hat{\beta}_{2j} \to_p \beta_{2j}$. The remaining steps are similar to those of Lemma A.2. We omit them here. This completes the proof of Lemma A.3. $\square$

Proof of Theorem 3.3. From the proof of Theorem 3.1, it also can be shown that $\hat{\omega}_{mj}$ is a monotone increasing function of $\hat{\beta}_{mj}^2$. Then

$$\hat{\omega}_{m(1)} = \left[1 + \sum_{j \ge 2}(\|\hat{\boldsymbol{Y}}^{(m)}\|^2 - n\hat{\beta}_{m(j)}^2)^{-n/2}(\|\hat{\boldsymbol{Y}}^{(m)}\|^2 - n\hat{\beta}_{m(1)}^2)^{n/2} + \sqrt{n}p_n\{1 - n\hat{\beta}_{m(1)}^2/\|\hat{\boldsymbol{Y}}^{(m)}\|^2\}\right]^{-1}.$$

Then, to demonstrate $\hat{\omega}_{m(1)} \to_p 1$, it suffices to show

$$\sum_{j\geq 2}(\|\hat{\boldsymbol{Y}}^{(m)}\|^2 - n\hat{\beta}^2_{mj})^{-n/2}(\|\hat{\boldsymbol{Y}}^{(m)}\|^2 - n\hat{\beta}^2_{m(1)})^{n/2} = o_p(1),$$

$$\sqrt{n}p_n(1 - n\hat{\beta}^2_{m(1)}/\|\hat{\boldsymbol{Y}}^{(m)}\|^2) = o_p(1).$$

Following the proof of Theorem 3.1, and under Conditions 1–5, we have completed the proof of the first part of Theorem 3.3.

We here just show the second part of this theorem. Notice that

$$\hat{\omega}_{m0} = \frac{\exp(-\mathrm{BIC}_{m0}/2)}{\sum_{j=0}^{p_n}\exp(-\mathrm{BIC}_{mj}/2)} = \left\{1 + (\sqrt{n}p_n)^{-1}\sum_{j=1}^{p_n}(1 - n\hat{\beta}^2_{mj}/\|\hat{\boldsymbol{Y}}^{(m)}\|^2)^{-n/2}\right\}^{-1}.$$

Similar to the proof of Theorem 3.1, it is straightforward to show $\|\hat{\boldsymbol{Y}}^{(m)}\|^2/n$ is bounded in probability. Thus, by assuming $\beta^2_{m(1)} = O(n^{-\gamma})$, $\gamma > 0$, and using the results of Lemma A.3, we have $(1 - n\hat{\beta}^2_{mj}/\|\hat{\boldsymbol{Y}}^{(m)}\|^2)^{-n/2} = O_p(1)$. As a result, $(\sqrt{n}p_n)^{-1}\sum_{j=1}^{p_n}(1 - n\hat{\beta}^2_{mj}/\|\hat{\boldsymbol{Y}}^{(m)}\|^2)^{-n/2} \to_p 0$, which implies that $\hat{\omega}_{m0} \to_p 1$. Hence, we have completed the proof. □

Proof of Theorem 3.4. Let $\boldsymbol{\Lambda} = \boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^M$, and we have $\|\boldsymbol{\Lambda}\|^2 = \|\boldsymbol{\Lambda}_{\mathcal{A}_n}\|^2 + \|\boldsymbol{\Lambda}_{\mathcal{A}_n^c}\|^2$. To prove $\|\hat{\boldsymbol{\beta}}^M - \boldsymbol{\beta}_0\| \to_p 0$, it suffices to show $\|\boldsymbol{\Lambda}_{\mathcal{A}_n}\| \to_p 0$ and $\|\boldsymbol{\Lambda}_{\mathcal{A}_n^c}\| \to_p 0$. Notice that

$$\begin{aligned}
\lambda_{\max}(\mathbf{X}^\top_{\mathcal{A}_n^c}\mathbf{X}_{\mathcal{A}_n^c})\|\boldsymbol{\Lambda}_{\mathcal{A}_n^c}\|^2 &\leq 2\lambda_{\max}(\mathbf{X}^\top_{\mathcal{A}_n^c}\mathbf{X}_{\mathcal{A}_n^c})\|\hat{\boldsymbol{\beta}}^M_{\mathcal{A}_n^c}\|^2 + 2\lambda_{\max}(\mathbf{X}^\top_{\mathcal{A}_n^c}\mathbf{X}_{\mathcal{A}_n^c})\|\boldsymbol{\beta}_{0\mathcal{A}_n^c}\|^2 \\
&\leq 2\,\mathrm{tr}(\mathbf{X}^\top_{\mathcal{A}_n^c}\mathbf{X}_{\mathcal{A}_n^c})\|\hat{\boldsymbol{\beta}}^M_{\mathcal{A}_n^c}\|^2 + 2\,\mathrm{tr}(\mathbf{X}^\top_{\mathcal{A}_n^c}\mathbf{X}_{\mathcal{A}_n^c})\|\boldsymbol{\beta}_{0\mathcal{A}_n^c}\|^2 \\
&= O(|\mathcal{A}_n^c|\,\|\hat{\boldsymbol{\beta}}^M_{\mathcal{A}_n^c}\|^2 + |\mathcal{A}_n^c|\,\|\boldsymbol{\beta}_{0\mathcal{A}_n^c}\|^2).
\end{aligned}$$

Furthermore, by Condition 6(iii) and the proof of Theorem 3.1, it is noteworthy that

$$\begin{aligned}
|\mathcal{A}_n^c|\,\|\hat{\boldsymbol{\beta}}^M_{\mathcal{A}_n^c}\|^2 &\leq M|\mathcal{A}_n^c|\sum_{j\in\mathcal{A}_n^c}\sum_{m=1}^{M}\hat{\omega}^2_{mj}\hat{\beta}^2_{mj} \\
&\leq M|\mathcal{A}_n^c|\,\|\hat{\boldsymbol{Y}}\|^2/n\sum_{j\in\mathcal{A}_n^c}\sum_{m=1}^{M}\hat{\omega}^2_{mj}\{\hat{\beta}_{mj}/(\sqrt{n}\|\hat{\boldsymbol{Y}}^{(m)}\|)\}^2 \\
&\leq M|\mathcal{A}_n^c|\sum_{j\in\mathcal{A}_n^c}\sum_{m=1}^{M}\hat{\omega}^2_{mj}\|\hat{\boldsymbol{Y}}\|^2/n \\
&\leq C''M|\mathcal{A}_n^c|\sum_{j\in\mathcal{A}_n^c}\sum_{m=1}^{M}\hat{\omega}^2_{mj} \\
&\leq C''M^2|\mathcal{A}_n^c|\sup_{m>1}\sum_{j\in\mathcal{A}_n^c}\hat{\omega}^2_{mj} \to_p 0.
\end{aligned}$$

By Condition 6(ii), together with Eq (A.4), it follows that $\|\boldsymbol{\Lambda}_{\mathcal{A}_n^c}\| \to_p 0$. Next we will show $\|\boldsymbol{\Lambda}_{\mathcal{A}_n}\| \to_p 0$. By Condition 7 and Lemma 3 in [22], we have

$$\|\boldsymbol{\Lambda}_{\mathcal{A}_n}\|^2 \leq \lambda^{-1}_{\min}\boldsymbol{\Lambda}^\top_{\mathcal{A}_n}(n^{-1}\mathbf{X}^\top_{\mathcal{A}_n}\mathbf{X}_{\mathcal{A}_n})\boldsymbol{\Lambda}_{\mathcal{A}_n}$$

$$\leq \lambda_{\min}^{-1}\|\mathbf{\Lambda}_{\mathcal{A}_n}\|_1 \max_j(n^{-1}|\boldsymbol{x}_j^\top \mathbf{X}_{\mathcal{A}_n}\mathbf{\Lambda}_{\mathcal{A}_n}|)$$
$$\leq \lambda_{\min}^{-1}|\mathcal{A}_n|^{1/2}\|\mathbf{\Lambda}_{\mathcal{A}_n}\| \max_j(n^{-1}|\boldsymbol{x}_j^\top \mathbf{X}_{\mathcal{A}_n}\mathbf{\Lambda}_{\mathcal{A}_n}|),$$

where the second inequality holds due to Hölder's inequality. This leads to

$$\|\mathbf{\Lambda}_{\mathcal{A}_n}\| \leq \lambda_{\min}^{-1}|\mathcal{A}_n|^{1/2} \max_j(n^{-1}|\boldsymbol{x}_j^\top \mathbf{X}_{\mathcal{A}_n}\mathbf{\Lambda}_{\mathcal{A}_n}|). \tag{A.5}$$

Notice that $\hat{\boldsymbol{Y}}^{(M+1)} = \hat{\boldsymbol{Y}}^{(M)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(M)} = \hat{\boldsymbol{Y}} - \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{X}\mathbf{\Lambda}$. Then, by the triangle inequality, we have

$$|\mathcal{A}_n|^{1/2} \max_j(n^{-1}|\boldsymbol{x}_j^\top \mathbf{X}\mathbf{\Lambda}|) \leq |\mathcal{A}_n|^{1/2} \max_j\{n^{-1}|\boldsymbol{x}_j^\top(\hat{\boldsymbol{Y}} - \mathbf{X}\boldsymbol{\beta}_0)|\} + |\mathcal{A}_n|^{1/2} \max_j(n^{-1}|\boldsymbol{x}_j^\top \hat{\boldsymbol{Y}}^{(M+1)}|).$$

For the first term of the right-hand side of the above inequality, we have

$$\begin{aligned}
&|\mathcal{A}_n|^{1/2} \max_j\{n^{-1}|\boldsymbol{x}_j^\top(\hat{\boldsymbol{Y}} - \mathbf{X}\boldsymbol{\beta}_0)|\} \\
&\leq |\mathcal{A}_n|^{1/2} \max_j\left(n^{-1}\left|\sum_{i=1}^n x_{ij}\varepsilon_i\right|\right) + |\mathcal{A}_n|^{1/2} \max_j\left[\frac{1}{n}\left|\sum_{i=1}^n x_{ij}(1-\delta_i)\left\{\frac{1}{p_n\mathcal{K}}\sum_{l=1}^{p_n}\sum_{k=1}^{\mathcal{K}}\tilde{y}_{ik}^{(l)} - y_i\right\}\right|\right] \\
&= I_1 + I_2.
\end{aligned}$$

By the Bonferroni inequality and Lemma A.3 in [38], we obtain $I_1 \to_p 0$. From the proof of Lemma A.1, for each $j = 1, \ldots, p_n$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n x_{ij}(1-\delta_i)\left\{\frac{1}{p_n\mathcal{K}}\sum_{l=1}^{p_n}\sum_{k=1}^{\mathcal{K}}\tilde{y}_{ik}^{(l)} - y_i\right\} \\
&= \frac{1}{np_n}\sum_{i=1}^n\sum_{l=1}^{p_n}(1-\delta_i)\left\{\frac{1}{\mathcal{K}}\sum_{k=1}^{\mathcal{K}} x_{ij}\tilde{y}_{ik}^{(l)} - \hat{h}_j(x_{ij})\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^n(1-\delta_i)\{\hat{h}_j(x_{ij}) - h_j(x_{ij})\} + \frac{1}{n}\sum_{i=1}^n(1-\delta_i)\{h_j(x_{ij}) - x_{ij}y_i\} \\
&= O_p(n^{-r}),
\end{aligned}$$

where $0 < r < 1/2$. Then, by Condition 6(i),

$$\begin{aligned}
I_2 &= |\mathcal{A}_n|^{1/2} \max_j\left[\frac{1}{n}\left|\sum_{i=1}^n x_{ij}(1-\delta_i)\left\{\frac{1}{p_n\mathcal{K}}\sum_{l=1}^{p_n}\sum_{k=1}^{\mathcal{K}}\tilde{y}_{ik}^{(l)} - y_i\right\}\right|\right] \\
&= \frac{|\mathcal{A}_n|^{1/2}}{n^{1-r}} \max_j\left[\frac{1}{n^r}\left|\sum_{i=1}^n x_{ij}(1-\delta_i)\left\{\frac{1}{p_n\mathcal{K}}\sum_{l=1}^{p_n}\sum_{k=1}^{\mathcal{K}}\tilde{y}_{ik}^{(l)} - y_i\right\}\right|\right] \\
&\to_p 0.
\end{aligned}$$

As a result, the term $|\mathcal{A}_n|^{1/2}\max_j\{n^{-1}|\boldsymbol{x}_j^\top(\hat{\boldsymbol{Y}} - \mathbf{X}\boldsymbol{\beta}_0)|\} \to_p 0$. Similar to the proof of Theorem 5 in [22], by Conditions 1–3, we can show that $\max_j(n^{-1}|\boldsymbol{x}_j^\top \hat{\boldsymbol{Y}}^{(M+1)}|) \to_p 0$. This together with Lemma A.1 in [36]

proves that $\boldsymbol{x}_j^\top \hat{\boldsymbol{Y}}^{(M+1)}/\sqrt{n} = O_p(1)$ for each $j = 1, \ldots, p_n$. Then we have $|\mathcal{A}_n|^{1/2} \max_j(n^{-1}|\boldsymbol{x}_j^\top \hat{\boldsymbol{Y}}^{(M+1)}|) \to_p 0$. Accordingly,

$$|\mathcal{A}_n|^{1/2} \max_j(n^{-1}|\boldsymbol{x}_j^\top \mathbf{X}\boldsymbol{\Lambda}|) \to_p 0. \tag{A.6}$$

On the other hand, by the Cauchy-Schwarz inequality and $\|\boldsymbol{\Lambda}_{\mathcal{A}_n^c}\| \to_p 0$, we obtain

$$|\mathcal{A}_n|\{n^{-1} \max_j |\boldsymbol{X}_j^\top \mathbf{X}_{\mathcal{A}_n^c}\boldsymbol{\Lambda}_{\mathcal{A}_n^c}|\}^2 \le 2|\mathcal{A}_n|\lambda_{\max}\{n^{-1}\mathbf{X}_{\mathcal{A}_n^c}^\top \mathbf{X}_{\mathcal{A}_n^c}\}\|\boldsymbol{\Lambda}_{\mathcal{A}_n^c}\|^2 \to_p 0.$$

This in conjunction with Eq (A.6) implies that $|\mathcal{A}_n|^{1/2} \max_j(n^{-1}|\boldsymbol{x}_j^\top \mathbf{X}_{\mathcal{A}_n}\boldsymbol{\Lambda}_{\mathcal{A}_n}|) \to_p 0$. Then we obtain that Eq (A.5) $\to_p 0$. Therefore, we have that $\|\boldsymbol{\Lambda}\| = \|\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^M\| \to_p 0$, which completes the proof. □

AIMS Press