Statistica Sinica **31** (2021), 1005-1026
doi:https://doi.org/10.5705/ss.202018.0297

A MODEL-AVERAGING METHOD FOR HIGH-DIMENSIONAL REGRESSION WITH MISSING RESPONSES AT RANDOM

Jinhan Xie[1], Xiaodong Yan[2] and Niansheng Tang[1]

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

Abstract: This study considers the ultrahigh-dimensional prediction problem in the presence of responses missing at random. A two-step model-averaging procedure is proposed to improve the prediction accuracy of the conditional mean of the response variable. The first step specifies several candidate models, each with low-dimensional predictors. To implement this step, a new feature-screening method is developed to distinguish between the active and inactive predictors. The method uses the multiple-imputation sure independence screening (MI-SIS) procedure, and candidate models are formed by grouping covariates with similar size MI-SIS values. The second step develops a new criterion to find the optimal weights for averaging a set of candidate models using weighted delete-one cross-validation (WDCV). Under some regularity conditions, we show that the proposed screening statistic enjoys the ranking consistency property, and that the WDCV criterion asymptotically achieves the lowest possible prediction loss. Simulation studies and an example demonstrate the proposed methodology.

Key words and phrases: High-dimensional data, missing at random, model averaging, multiple imputation, screening, weighted delete-one cross-validation.

1. Introduction

Model selection and model averaging are two popular approaches to improving the prediction accuracy in a regression analysis. Model selection is often implemented by using some proper criterion, such as the Akaike information criterion (AIC) (Akaike (1973)) or Bayesian information criterion (BIC) (Schwarz (1978)), to select the best model from among a set of candidate models. Because these model-selection methods ignore the contributions of other candidate models, they may suffer from the model selection uncertainty and bias problem when a single model is not overwhelmingly supported by the data (Hjort and Claeskens (2003)). More importantly, different model-selection methods or

Corresponding author: Niansheng Tang, Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, Yunnan University, Kunming, 650500, P. R. of China. E-mail: nstang@ynu.edu.cn.

criteria may lead to different best models; thus statistical inferences based on the final model would vary between data sets. To address this issue, a model-averaging approach has been proposed to improve the prediction accuracy that pools predictions by giving higher weights to better models. As such, it often reduces the bias in the regression prediction, by not depending on only one best model. Furthermore it ensures that we do not ignore useful information from the relationship between the response and the covariates (Zhang (2013)). Various model-averaging approaches have been proposed, including AIC model averaging (Akaike (1979)), BIC model averaging (Hoeting et al. (1999)), Mallows model averaging (Hansen (2007); Wan, Zhang and Zou (2010)), jackknife model averaging (Hansen and Racine (2012)), Kullback–Leibler (KL) loss model averaging (Zhang et al. (2016)), and generalized least squares model averaging with heteroskedastic errors (Liu, Okui and Yoshimura (2016)). However, the aforementioned methods are applicable only when the dimension of the predictors is less than the sample size, and thus cannot be applied directly to ultrahigh-dimensional data.

High-dimensional data in which the number of predictors is much larger than the sample size, are often encountered in fields such as biomedicine, social science, and economics. A statistical analysis of high-dimensional data is quite challenging. For conducting inferences on statistical models with high-dimensional data, many penalized methods have been developed that simultaneously select important predictors and estimate unknown parameters in the considered models. These methods include the Lasso (Tibshirani (1996)), smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)), and minimax concave penalty (MCP) (Zhang (2010)). When the dimensionality of the predictors grows exponentially fast with the sample size, we use feature-screening methods to reduce the dimensionality of the predictors to a moderate scale, allowing us to apply classical statistical inference methods to the reduced models. For example, see Fan and Lv (2008), Fan and Song (2010), and Chang, Tang and Wu (2013) for model-based feature-screening methods; and Zhu et al. (2011); Li, Zhong and Zhu (2012); He, Wang and Hong (2013); Chang, Tang and Wu (2016); Xie et al. (2020) for model-free feature-screening approaches. A few works have investigated model averaging in ultrahigh-dimensional data. For instance, Ando and Li (2014) proposed a two-step model-averaging procedure for ultrahigh-dimensional regression models using a delete-one cross-validation procedure to estimate the model weights; Lan et al. (2018) proposed a sequential model-averaging approach to making stable predictions for high-dimensional linear regression models. However, existing model-averaging methods for high-dimensional regression models

focus mainly on fully observed data.

Missing data are relatively common in surveys, clinical trials, and longitudinal studies. For example, some individuals may be unwilling to answer sensitive questions, information may be lost as a result of uncontrollable factors, and individuals may be surveyed intermittently or drop out of the study (Little and Rubin (2019)). Ignoring missing data may lead to prediction bias. To address this issue, model-selection and model-averaging methods have been developed to improve the prediction accuracy in the presence of missing data. For example, Ibrahim, Zhu and Tang (2008) developed a novel model-selection criterion for the missing data problem based on the EM algorithm; Schomaker, Wan and Heumann (2010) presented two approaches to handle missing data for the model-averaging problem; Dardanoni, Modica and Peracchi (2011) adopted a model-averaging approach to tackle the bias–precision trade-off in the presence of missing covariate values in linear regression models; Zhang (2013) proposed using the Mallows model-averaging approach to handle covariates missing completely at random; and Fang et al. (2017) presented a model-averaging approach in the context of fragmentary data. However, the aforementioned works all apply to the classical setting in which the number of predictors is fixed and less than the sample size. To the best of our knowledge, few works have examined model-averaging for ultrahigh-dimensional regression models with responses missing at random (MAR).

This study proposes a two-step model-averaging approach for ultrahigh-dimensional regression models in the presence of responses MAR. The first step constructs a set of candidate models, each with low-dimensional predictors. To implement this step, we develop a new feature-screening index, called the multiple-imputation sure independence screening (MI-SIS) index, to identify the active and inactive predictors. Thus, candidate models are formed by grouping predictors with similar size MI-SIS values. Under some mild regularity assumptions, we show its sure screening and ranking consistency properties. The proposed feature-screening procedure is robust to a misspecification of the propensity score function. The second step uses the weighted delete-one cross-validation (WDCV) criterion to identify the optimal weights for averaging a set of candidate models. Under some regularity assumptions, we prove that the derived weights are asymptotically optimal, in the sense that the corresponding weighted squared error is asymptotically identical to that of the infeasible best positive model averaging estimator, where the standard constraint that the sum of the weights is equal to one is removed.

For simplicity, we assume a parametric propensity score function with high-dimensional covariates. A penalized likelihood method with some proper penalty function is employed to simultaneously estimate the regression coefficients and select the significant covariates in the assumed parametric propensity score function. In addition, we present a data-driven approach (e.g., the BIC) in numerical studies to select the tuning parameter in the penalized likelihood function. Under some regularity assumptions, we prove the oracle properties of the proposed penalized likelihood estimators of the parameters, including the sparsity and asymptotic normality.

The rest of this paper is organized as follows. In Section 2, we describe the model setting and present our two-step model-averaging procedure in the presence of responses MAR. In Section 3, we systematically investigate the asymptotic properties of the proposed shrinkage estimators, establish the sure screening and rank consistency properties of the proposed screening procedure, and demonstrate the optimality of the weighted model-averaging estimator. In Section 4, we evaluate the proposed methods using simulation studies and a real-data example. Section 5 concludes the paper. All technical details are given in the online Supplementary Material.

2. Method

Consider a data set $\{(Y_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$ with n individuals, where Y_i is the response variable and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ is a $p \times 1$ vector of predictors. It is assumed that $\boldsymbol{X}_i$ are fully observed, whilst Y_i are subject to missingness. We define $\delta_i = 1$ if Y_i is observed, and $\delta_i = 0$ otherwise. Thus, the complete data set consists of observations $\{(\boldsymbol{X}_i, Y_i, \delta_i), i = 1, \ldots, n\}$. To quantify the relationship between the response variable and the predictors, we consider the following linear regression model:

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a $p \times 1$ vector of unknown regression coefficients, and ε_i is an independent random error with mean zero and finite variance σ_ε^2. Without loss of generality, we omit the intercept term. Throughout this paper, we assume that the number of predictors is allowed to grow with the sample size; that is, $\log(p) = o(n^v)$, for some constant $v \in (0, 1)$. In this case, it is recognized that only a few predictors may indeed contribute to Y_i; that is, model (2.1) has a sparse structure. Thus, a feature-screening method should be employed to select the important predictors.

For missing data Y_i, we assume that δ_i is independent of δ_j, for any $i \neq j$, and that δ_i depends only on some components of $\boldsymbol{X}_i$; that is, the missingness data mechanism is MAR. However, in practice, it is rather difficult to determine which components of $\boldsymbol{X}_i$ contribute to the missingness of Y_i. More importantly, it is recognized that only a few covariates may indeed contribute to the missingness of Y_i (Lee and Tang (2006)). In general, we initially incorporate many covariates to specify the missingness data mechanism, and then use a penalized method to identify those that contribute to the missingness of Y_i. For example, following much of the literature on missing data, we consider the following parametric model for δ_i:

$$\Pr(\delta_i = 1|\boldsymbol{U}_i, \boldsymbol{\gamma}) = \pi(\boldsymbol{U}_i; \boldsymbol{\gamma}) := \pi_i(\boldsymbol{\gamma}), \tag{2.2}$$

which defines a MAR mechanism, where $\boldsymbol{\gamma}$ is a $q \times 1$ vector of unknown parameters, and $\pi(\cdot)$ is the selection probability function. Furthermore $\boldsymbol{U}_i$ is a subvector of $\boldsymbol{X}_i$ (i.e., $\boldsymbol{U}_i$ is composed of some components of $\boldsymbol{X}_i$), but the true components of $\boldsymbol{U}_i$ (i.e., the covariates indeed contribute to the missingness of Y_i) may be different from those of $\boldsymbol{X}_i$ (i.e., the predictors indeed contribute to Y_i). As an illustration, consider $\text{logit}\{\pi_i(\boldsymbol{\gamma})\} = \boldsymbol{U}_i^\top \boldsymbol{\gamma}$, where $\text{logit}(\pi_i) = \log\{\pi_i/(1-\pi_i)\}$. For identification, we assume that q may be less than p, and $\log(q) = O(n^\alpha)$ for $\alpha \in (0, 1/2)$. We again assume that the aforementioned missingness data mechanism model has a sparse structure.

Under the MAR assumption defined above, penalized methods such as the Lasso, Adaptive Lasso, and SCAD methods can be employed to evaluate the maximum likelihood estimation (denoted as $\hat{\boldsymbol{\gamma}}$) of $\boldsymbol{\gamma}$. That is, $\hat{\boldsymbol{\gamma}}$ can be obtained by maximizing the following penalized log-likelihood function with respect to γ:

$$Q_n(\boldsymbol{\gamma}) = \frac{1}{n} l_n(\boldsymbol{\gamma}) - \sum_{j=1}^{q} f_{\lambda_n}(|\gamma_j|), \tag{2.3}$$

where $l_n(\boldsymbol{\gamma}) = \sum_{i=1}^{n} [\delta_i \log \pi(\boldsymbol{U}_i, \boldsymbol{\gamma}) + (1-\delta_i) \log\{1 - \pi(\boldsymbol{U}_i, \boldsymbol{\gamma})\}]$, $f_{\lambda_n}(t)$ is some proper penalty function, γ_j is the jth component of $\boldsymbol{\gamma}$, and $\lambda_n \geq 0$ is a regularization parameter controlling the trade-off between the bias and the model complexity. For example, one can take $f_{\lambda_n}(t)$ as the SCAD regularization (Fan and Li (2001)), which is defined in terms of its first derivative and is symmetric around the origin. For $\gamma > 0$, the first derivative of the SCAD regularization has the form

$$f'_{\lambda_n}(\gamma) = \lambda_n \left\{ I(\gamma \leq \lambda_n) + \frac{(a\lambda_n - \gamma)_+}{(a-1)\lambda_n} I(\gamma > \lambda_n) \right\},$$

where $a > 2$ and $\lambda_n > 0$ are the tuning parameters, $b_+ = bI(b \geq 0)$, and $I(\gamma \leq \lambda_n)$ is an indicator function of the event $\{\gamma \leq \lambda_n\}$, which takes one if $\gamma \leq \lambda_n$, and zero otherwise. Fan and Li (2001) proposed using $a = 3.7$, from a Bayesian point of view. The parameter λ_n can be determined using a data-driven method such as cross-validation (CV) or generalized cross-validation (GCV).

For the linear regression model defined in (2.1), we denote the number of true predictors (i.e., those with nonzero regression coefficients β_j) as d. In practice, both d and the set of true predictors $\mathscr{A}_\beta = \{j : |\beta_j| > 0\}$ are unknown. Model averaging is widely used to improve the prediction accuracy for the considered model (2.1). Prior studies on model averaging have mainly focused on settings with no missing data or a low-dimensional linear regression. In what follows, we extend the model-averaging approach to a setting that simultaneously includes responses MAR and a high-dimensional linear regression. Thus, to improve the accuracy of predicting the mean of Y in a high-dimensional linear regression in the presence of responses MAR, we propose the following two-step model-averaging procedure.

Step 1: Construct candidate models

In this step, we construct the candidate models. Denote a set of S candidate models $M_1, \ldots, M_S$ as

$$M_s : Y_i = \sum_{j \in A_s} X_{ij}\beta_j + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where A_s is the index set of predictors in the sth candidate model M_s, for $s = 1, \ldots, S$. Here, we assume that $\mathscr{A}_\beta \subset \{A_1 \cup A_2 \cup \cdots \cup A_S\} \subset A$ and $A_k \cap A_j = \phi$, for any $k \neq j$, where $A = \{X_1, \ldots, X_p\}$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{\beta}_s = \{\beta_j : j \in A_s\}$ be a $p_s \times 1$ vector of unknown regression coefficients, $\boldsymbol{X}_s = \{X_{ij} : i = 1, \ldots, n, j \in A_s\}$ be an $n \times p_s$ design matrix, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. Thus, the sth candidate model M_s can be written as $\boldsymbol{Y} = \boldsymbol{X}_s\boldsymbol{\beta}_s + \boldsymbol{\varepsilon}$.

For the sth candidate model M_s, we adopt the propensity score adjusted least squares (PS-LS) method to estimate $\boldsymbol{\beta}_s$. That is, under the aforementioned assumption, the PS-LS estimator $\tilde{\boldsymbol{\beta}}_s$ of $\boldsymbol{\beta}_s$ can be obtained using

$$\tilde{\boldsymbol{\beta}}_s = \underset{\boldsymbol{\beta}_s}{\operatorname{argmin}} (\boldsymbol{Y} - \boldsymbol{X}_s\boldsymbol{\beta}_s)^\top \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{X}_s\boldsymbol{\beta}_s),$$

where $\boldsymbol{W} = \operatorname{diag}(\delta_1/\pi_1, \ldots, \delta_n/\pi_n)$, in which $\pi_i = \pi(\boldsymbol{U}_i; \boldsymbol{\gamma})$, for $i = 1, \ldots, n$. It is easily shown that $\tilde{\boldsymbol{\beta}}_s = (\boldsymbol{X}_s^\top \boldsymbol{W} \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s^\top \boldsymbol{W} \boldsymbol{Y}$. Thus, based on the sth

candidate model M_s, the PS-LS prediction of the mean of the response variables $\boldsymbol{Y}$ is given by $\tilde{\boldsymbol{\mu}}_s = \boldsymbol{X}_s\tilde{\boldsymbol{\beta}}_s = \boldsymbol{X}_s(\boldsymbol{X}_s^\top \boldsymbol{W}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^\top \boldsymbol{W}\boldsymbol{Y}$. When $\boldsymbol{\gamma}$ is unknown, we use $\hat{\boldsymbol{\gamma}}$ in place of $\boldsymbol{\gamma}$. Thus, the corresponding estimator of $\boldsymbol{\mu}_s$ has the form $\hat{\boldsymbol{\mu}}_s = \boldsymbol{X}_s\hat{\boldsymbol{\beta}}_s = \boldsymbol{X}_s(\boldsymbol{X}_s^\top \widehat{\boldsymbol{W}}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^\top \widehat{\boldsymbol{W}}\boldsymbol{Y}$, where $\hat{\boldsymbol{\beta}}_s = (\boldsymbol{X}_s^\top \widehat{\boldsymbol{W}}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^\top \widehat{\boldsymbol{W}}\boldsymbol{Y}$, and $\widehat{\boldsymbol{W}} = \mathrm{diag}(\delta_1/\hat{\pi}_1, \ldots, \delta_n/\hat{\pi}_n)$, in which $\hat{\pi}_i = \pi(\boldsymbol{U}_i; \hat{\boldsymbol{\gamma}})$, for $i = 1, \ldots, n$.

After applying the PS-LS estimation procedure to the S candidate models introduced above, we obtain S PS-LS predictions of the mean of the response variable, that is, $\{\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_S\}$. Given a weight vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_S)^\top \in \mathcal{W} = \{\boldsymbol{\omega} \in [0,1]^S : 0 \le \omega_s \le 1\}$, the model-averaging predictor of the mean of the response variables is defined as

$$\hat{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \sum_{s=1}^{S} \omega_s \hat{\boldsymbol{\mu}}_s = \sum_{s=1}^{S} \omega_s \boldsymbol{X}_s(\boldsymbol{X}_s^\top \widehat{\boldsymbol{W}}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^\top \widehat{\boldsymbol{W}}\boldsymbol{Y} = \sum_{s=1}^{S} \omega_s \widehat{\boldsymbol{P}}_s\boldsymbol{Y} = \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y},$$

where $\widehat{\boldsymbol{P}}_s = \boldsymbol{X}_s(\boldsymbol{X}_s^\top \widehat{\boldsymbol{W}}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^\top \widehat{\boldsymbol{W}}$, for $s = 1, \ldots, S$, and $\widehat{\boldsymbol{P}}(\boldsymbol{\omega}) = \sum_{s=1}^{S} \omega_s \widehat{\boldsymbol{P}}_s$ is the corresponding hat matrix. In the literature on model averaging, one usually assumes that $\sum_{s=1}^{S} \omega_s = 1$. Here, we omit this assumption, following Ando and Li (2014).

When there are many candidate models, it is computationally intensive to evaluate the model-averaging estimator $\hat{\boldsymbol{\mu}}(\boldsymbol{\omega})$ in high-dimensional regression models. Thus, it is desirable to adopt a feature-screening approach to screen important predictors prior to the model averaging in the presence of responses MAR. To this end, a novel feature-screening procedure is developed, which we describe below.

Without loss of generality, we assume that the covariates have been standardized, and $Y \perp \delta | X_k$ (e.g., see He, Wang and Hong (2013)), for $k = 1, \ldots, p$, where $\perp$ represents statistical independence. Under the above assumption, we can use the information of X_k rather than $\boldsymbol{X}$ to impute the missing data in the marginal utility. Thus, for $k = 1, \ldots, p$, we define the estimated marginal MI-SIS index between Y and X_k as

$$\hat{r}_k = \frac{1}{n}\sum_{i=1}^{n} \left\{ \delta_i X_{ik} Y_i + (1-\delta_i)\frac{1}{m}\sum_{v=1}^{m} X_{ik}\tilde{Y}_{iv}^k \right\}, \tag{2.4}$$

where m is the number of multiple imputations, $\{\tilde{Y}_{iv}^k\}_{v=1}^m$ are m independent imputations for missing Y_i from $\hat{F}(y|X_{ik})$, $\hat{F}(y|X_{ik}) = \sum_{j=1}^{n} \vartheta_{ik}^j I(Y_i \le y)$ is a kernel estimator of $F(y|X_{ik})$, $\vartheta_{ik}^j = \delta_j K_h(X_{jk} - X_{ik}) / \sum_{\ell=1}^{n} \delta_\ell K_h(X_{\ell k} - X_{ik})$,

$F(y|X_{ik})$ is the conditional distribution of Y given $X_k = X_{ik}$, $K_h(u) = K(u/h)$, $K(\cdot)$ is a kernel function on the real line, $h = h_n$ is a positive smoothing bandwidth sequence, such as $h_n \to 0$, and $I(\cdot)$ is the indicator function. Following Wang and Chen (2009), $\tilde{Y}_{iv}^k$ effectively has a discrete distribution, where the probability of selecting Y_{jk} with $\delta_j = 1$ is ϑ_{ik}^j. Thus, for a complete data set $\{(\boldsymbol{X}_i, Y_i, \delta_i) : i = 1, \ldots, n\}$, it is easy to calculate $\hat{r}_k$ using (2.4), for $k = 1, \ldots, p$. Then, we can sort the magnitudes of $\hat{r}_k$ in decreasing order, and select the important predictors using the criterion $\widehat{\mathcal{M}}_{\varrho_n} = \{1 \le k \le p : |\hat{r}_k| > \varrho_n\}$, which is usually called the estimated active predictor subset, where ϱ_n is the prespecified threshold value. Based on the above feature-screening criterion, the full model with p predictors may shrink to a reduced model with fewer than n predictors.

Based on the calculated MI-SIS statistics between the response variable and each of the p predictors in the presence of missing responses, we partition the p predictors into $S + 1$ groups, where the first group has the highest MI-SIS value, and the $(S+1)$th group has the MI-SIS value closest to zero. Let the sth candidate model consist of those predictors with MI-SIS values in the sth group. We drop the $(S + 1)$th group, and use only the first S groups to conduct the model averaging. That is, the number of candidate models is S.

Step 2: Determine the optimal weights

The key task when evaluating $\hat{\boldsymbol{\mu}}(\boldsymbol{\omega})$ is to find the optimal weights ω_s. Many methods, such as the CV and GCV methods, can be used to implement this task. Here, the delete-one CV approach is adopted to evaluate the optimal weights, owing to its asymptotic optimality theory for heteroskedastic errors. Let $\tilde{\mu}_s^{(-i)}$ be the predicted value of the mean of the response variables computed after deleting the ith observation $(\boldsymbol{X}_i, Y_i, \delta_i)$ from the sample in the sth candidate model M_s. Denote $\tilde{\boldsymbol{\mu}}_s^d = (\tilde{\mu}_s^{(-1)}, \ldots, \tilde{\mu}_s^{(-n)})^\top$, and $\overline{\boldsymbol{P}}_s = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}_s(\boldsymbol{X}_s^\top\widehat{\boldsymbol{W}}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^\top\widehat{\boldsymbol{W}}^{1/2}$. It is easily shown that $\tilde{\boldsymbol{\mu}}_s^d$ can be written as $\tilde{\boldsymbol{\mu}}_s^d = \widetilde{\boldsymbol{P}}_s\boldsymbol{Y}$, where $\widetilde{\boldsymbol{P}}_s = \widehat{\boldsymbol{D}}_s(\widehat{\boldsymbol{P}}_s - \boldsymbol{I}) + \boldsymbol{I}$. Here, $\widehat{\boldsymbol{D}}_s = \mathrm{diag}(\hat{d}_1^s, \ldots, \hat{d}_n^s)$, where $\hat{d}_i^s = 1/(1 - \hat{h}_{ii}^s)$ and $\hat{h}_{ii}^s$ is the ith diagonal element of $\overline{\boldsymbol{P}}_s$, for $i = 1, \ldots, n$. Then, the delete-one predictor of the mean of the response variables is defined as

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \sum_{s=1}^{S} \omega_s \tilde{\boldsymbol{\mu}}_s^d = \sum_{s=1}^{S} \omega_s \widetilde{\boldsymbol{P}}_s\boldsymbol{Y} = \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y},$$

where $\widetilde{\boldsymbol{P}}(\boldsymbol{\omega}) = \sum_{s=1}^{S} \omega_s \widetilde{\boldsymbol{P}}_s$. Similarly to Hansen and Racine (2012), to incorporate the information associated with the missing data, we use the following

weighted squared error loss function to select the optimal weight vector $\boldsymbol{\omega}$:

$$\mathrm{WCV}(\boldsymbol{\omega}) = \{\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\omega})\}^\top \widehat{\boldsymbol{W}} \{\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\omega})\} = \{\boldsymbol{Y} - \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\}^\top \widehat{\boldsymbol{W}} \{\boldsymbol{Y} - \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\},$$

which is also referred to as the weighted delete-one CV criterion. According to the above definition, we can rewrite $\mathrm{WCV}(\boldsymbol{\omega})$ as

$$\begin{aligned}\mathrm{WCV}(\boldsymbol{\omega}) &= \boldsymbol{Y}^\top \widehat{\boldsymbol{W}} \boldsymbol{Y} - 2\sum_{s=1}^{S} \omega_s \boldsymbol{Y}^\top \widetilde{\boldsymbol{P}}_s \widehat{\boldsymbol{W}} \boldsymbol{Y} + \sum_{s=1}^{S}\sum_{k=1}^{S} \omega_s \omega_k \boldsymbol{Y}^\top \widetilde{\boldsymbol{P}}_s^\top \widehat{\boldsymbol{W}} \widetilde{\boldsymbol{P}}_k \boldsymbol{Y} \\ &= \boldsymbol{Y}^\top \widehat{\boldsymbol{W}} \boldsymbol{Y} - 2\boldsymbol{\omega}^\top \mathscr{A} + \boldsymbol{\omega}^\top \mathscr{B} \boldsymbol{\omega},\end{aligned}$$

which indicates that $\mathrm{WCV}(\boldsymbol{\omega})$ is a quadratic function of $\boldsymbol{\omega}$, where $\mathscr{A}$ is an $S \times 1$ vector with the sth component $\mathscr{A}_s = \boldsymbol{Y}^\top \widetilde{\boldsymbol{P}}_s \widehat{\boldsymbol{W}} \boldsymbol{Y}$, and $\mathscr{B}$ is an $S \times S$ matrix with the (s,k)th component $\mathscr{B}_{s,k} = \boldsymbol{Y}^\top \widetilde{\boldsymbol{P}}_s^\top \widehat{\boldsymbol{W}} \widetilde{\boldsymbol{P}}_k \boldsymbol{Y}$. Thus, the weight vector $\boldsymbol{\omega}$ is selected by minimizing $\mathrm{WCV}(\boldsymbol{\omega})$ over the set $\mathcal{W}$; that is,

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega} \in \mathcal{W}}{\operatorname{argmin}}\, \mathrm{WCV}(\boldsymbol{\omega}) = \underset{\boldsymbol{\omega} \in \mathcal{W}}{\operatorname{argmin}} \{-2\boldsymbol{\omega}^\top \mathscr{A} + \boldsymbol{\omega}^\top \mathscr{B} \boldsymbol{\omega}\}. \tag{2.5}$$

Unlike other cross-validation problems, which are often time-consuming, numerous software packages (e.g., the quadprog package in R and Matlab) are available to evaluate the above quadratic optimization problem in a short time, even if S is quite large. That is, the proposed optimization problem is computationally feasible. Based on the optimal weights evaluated above, the model-averaging predictor of the mean of the response variable can be expressed as $\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\omega}}) = \sum_{s=1}^{S} \hat{\omega}_s \hat{\boldsymbol{\mu}}_s$.

3. Asymptotic Properties

The theoretical properties of the penalized likelihood estimator $\hat{\boldsymbol{\gamma}}$ and the proposed feature-screening procedure can be found in the Supplementary Material. In what follows, we investigate the asymptotic properties of the proposed model-averaging procedure.

Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ and $\boldsymbol{\mu} = E(\boldsymbol{Y}|\boldsymbol{X})$. Consider the loss function $L(\boldsymbol{\omega}) = \{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}^\top \widehat{\boldsymbol{W}} \{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}$, with a risk function $R(\boldsymbol{\omega}) = E[\{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}^\top \widehat{\boldsymbol{W}} \{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}|\boldsymbol{X}]$. Let $\xi_n = \inf_{\boldsymbol{\omega} \in \mathcal{W}} R(\boldsymbol{\omega})$, which indicates that ξ_n is the lowest risk among all the considered weights. Here, C_0, $C_1, \ldots, C_5$ denote some appropriate constants, $\phi(\cdot)$ represents the maximal diagonal element of a matrix, and p_s denotes the number of columns of matrix $\boldsymbol{X}_s$. To obtain the asymptotic properties of the proposed model-averaging procedure using the WDCV approach, we need the following regularity conditions.

Assumption 1. *The propensity score function* $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma}) > C_0 > 0$, *for* $i = 1,\ldots,n$. *Its first three order derivations with respect to* $\boldsymbol{\gamma}$ *are continuous and bounded.*

Assumption 2. $E(\boldsymbol{X}_s^\top \boldsymbol{W}\boldsymbol{X}_s)$ *is nonsingular, and there exists a constant* $C_1 > 0$, *such that* $\mathbb{E}_{\min}(\sum_{i=1}^n X_{is}X_{is}^\top/n) \geq C_1$, *for any* s *and* n, *and* $\mathbb{E}_{\max}(\sum_{i=1}^n X_{is}X_{is}^\top/n)$ *is uniformly bounded with respect to* s *and* n, *where* $\mathbb{E}_{\min}(\boldsymbol{A})$ *and* $\mathbb{E}_{\max}(\boldsymbol{A})$ *represent the smallest and largest eigenvalues, respectively, of matrix* $\boldsymbol{A}$.

Assumption 3. *For some fixed integer* $1 \leq G < \infty$, *(i)* $E(\varepsilon_i^{4G}) \leq C_2 < \infty$, *for* $i = 1,\ldots,n$; *(ii)* $\sup_{1\leq s\leq S} p_s^2 d_m/n = o(1)$; *(iii)* $\sup_{1\leq s\leq S} p_s^{8/3} d_m/n \leq C_3 < \infty$; *(iv)* $\|\boldsymbol{\mu}\|^2/n \leq C_4 < \infty$; *(v)* $\sup_{1\leq s\leq S}\{\phi(\boldsymbol{P}_s)/p_s\} \leq C_5/n$; *and (vi)* $S^{4G+2}\|\boldsymbol{\mu}\|^{2G}/\xi_n^{2G} = o_p(1)$, *where* $\boldsymbol{P}_s = \boldsymbol{X}_s(\boldsymbol{X}_s^\top\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^\top$, *for* $s = 1,\ldots,S$.

Assumption 1 is necessary for missing data. The lower bound guarantees that the weights do not go to infinity as the sample size increases, and that the proposed parametric weights are asymptotically consistent. Assumption 2 states that the design matrix is uniformly bounded; the nonsingular assumption is necessary to ensure the existence of the hat matrix. Assumption 3(i) is a moment condition on the random error, and can be satisfied for Gaussian noise. Assumption 3(ii) limits the increasing rate of p_s as $n \to \infty$, and implies that the quantity $p_s^2 d_m$ increases at a slower rate than n for $s = 1,\ldots,S$. Thus, this assumption is stronger than assumption (6) of Ando and Li (2014). The cost of imposing this restriction is using the estimated propensity score function in the PS-LS estimation. Assumption 3(iii) shows that $p_s^{8/3} d_m$ has the same increasing rate as n. Assumption 3(iv) is a commonly used condition in linear regression models; for example, see Wan, Zhang and Zou (2010) and Ando and Li (2014). Assumption 3(v) excludes extremely unbalanced designs for each of the candidate models, and is the same as Condition (5.2) of Li (1987). Assumption 3(vi) indicates that $\xi_n \to \infty$, that is, there is no finite approximating model for which the bias is zero. If the number of candidate models S increases to infinity as the sample size increases, ξ_n should grow at a rate no slower than $\sqrt{n}$, under Assumption 3(iv). Suppose that the order of ξ_n is $n^{1-\phi}$, with $\phi \geq 0$. Then, Assumption 3(vi) reduces to $S^{(2+1/G)} = o_p(n^{(1-2\phi)/2})$. In particular, when G is fixed and $\phi < 1/2$, S is allowed to grow to infinity.

Theorem 1. *Suppose that Assumptions* 1–3 *hold. Then, as* $n \to \infty$, *we have*

$$\frac{L(\hat{\boldsymbol{\omega}})}{\inf_{\boldsymbol{\omega}\in\mathcal{W}} L(\boldsymbol{\omega})} \to 1, \tag{3.1}$$

where the convergence is in probability.

Theorem 1 shows that the proposed WDCV criterion for selecting the optimal weights is asymptotically equivalent to the weighted squared error. Thus, the proposed model-averaging estimator of $\boldsymbol{\mu}$ is asymptotically optimal in the class of model-averaging estimators, where the weight vector is restricted to the set $\mathcal{W}$.

4. Numerical Studies

In this section, we first conduct simulation studies to investigate the finite-sample performance of the proposed two-step model-averaging procedure and MI-SIS procedure for identifying the active and inactive predictors. We then use an example to illustrate the proposed methodologies.

4.1. Simulation studies

In this subsection, we use the weighted mean square error (WMSE) for 100 replications to measure the effectiveness of the proposed model-averaging approach. Here, the WMSE for 100 replications is defined as

$$\text{WMSE} = \frac{1}{100}\sum_{k=1}^{100}\sum_{i=1}^{n}\frac{\delta_i}{\pi(\boldsymbol{U}_i;\hat{\boldsymbol{\gamma}})}\left(\boldsymbol{\mu}_{i0} - \hat{\boldsymbol{\mu}}_i^{(k)}(\hat{\boldsymbol{\omega}})\right)^2,$$

where $\boldsymbol{\mu}_{i0}$ is the true value of the mean of the response variable $\boldsymbol{Y}$ given $\boldsymbol{X}_i$, and $\hat{\boldsymbol{\mu}}_i^{(k)}$ is the estimated mean of the response variable $\boldsymbol{Y}$ in the kth replication.

First, to investigate the sensitivity of the proposed model-averaging approach to the feature-screening methods used in step 1, we calculate the WMSEs for the MI-SIS method and for existing feature-screening methods, such as the inverse probability weighted sure independence screening method (IPW-SIS; Lai et al. (2017)), borrowing missingness information (BMI) containing missing indicator surrogate feature screening method (MI-S), and missing indicator imputation screening method (MI-I; Wang and Li (2018)). Second, for the predictors selected using the proposed feature-screening procedure, we compare the performance of the proposed model-averaging approach with that of the following methods: (A) model averaging with the AIC under the restriction $\sum_{s=1}^{S}\omega_s = 1$ (MAIC); (B) model averaging with the BIC under the restriction $\sum_{s=1}^{S}\omega_s = 1$ (MBIC); (C) weighted model-averaging method of Ando and Li (2014), without adjusting the missing data (MCV); (D) weighted model averaging with the CV method, without the restriction $\sum_{s=1}^{S}\omega_s = 1$ (WMCV1); (E) model averaging with the

CV method for the CC data, without the restriction $\sum_{s=1}^{S}\omega_s = 1$ (CC1); (F) weighted model averaging with the CV method under the restriction $\sum_{s=1}^{S}\omega_s = 1$ (WMCV2); (G) model averaging with the CV method for the CC data under the restriction $\sum_{s=1}^{S}\omega_s = 1$ (CC2); (H) the penalized likelihood method with the SCAD (SCAD); (I) the penalized likelihood method with the MCP (MCP); (J) the penalized likelihood method with the Lasso (LASSO); (K) the penalized likelihood method with the group Lasso (G-LASSO). The latter is implemented by partitioning p predictors into $S+1$ groups, and the first S groups are the same as those obtained in the model-averaging procedure. To implement the proposed feature-screening procedure, we employ the Gaussian kernel function $K(u) = \exp(-u^2/2)/(2\pi)^{1/2}$, and select the bandwidth using the CV method.

Experiment 1. Consider the following linear model:

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$ is a $p \times 1$ vector of predictors, and the noise ε_i is independent of the predictors. Here, $\boldsymbol{X}_i$ is generated from a multivariate normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, with components of $\boldsymbol{\Sigma} = (\sigma_{jk})_{p\times p}$ being $\sigma_{jk} = \rho^{|j-k|}$, for $1 \le j, k \le p$. The true values of nonzero β_j are independently sampled from the normal distribution $\mathcal{N}(0, 0.5^2)$. Thus, the mean of the response variables is $\boldsymbol{\mu} = (\boldsymbol{X}_1^\top\boldsymbol{\beta}, \ldots, \boldsymbol{X}_n^\top\boldsymbol{\beta})^\top$. We assume that $\boldsymbol{X}_i$ are completely observed, but that Y_i are subject to missingness. The missing indicators δ_i of Y_i are generated independently from the Bernoulli distribution with probability $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma}) = \Pr(\delta_i = 1|\boldsymbol{U}_i)$, where $\boldsymbol{U}_i = (X_{i1}, X_{i2})^\top$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^\top$. In this experiment, we take $n = 60$, $p = 1{,}000$, and $d = 50$, and assume that the true index set of nonzero β_j is $\mathcal{A}_\beta = \{j : j = 20(k-1)+1,\ k = 1, \ldots, 50\}$. Here, we consider the following four settings for ρ, $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})$ and the distribution of ε_i:

(a) $\rho = 0.7$, $\varepsilon_i \sim \mathcal{N}(0, 0.5)$, $\text{logit}\{\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})\} = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2}$, where the true value of $\boldsymbol{\gamma}$ is taken as $\boldsymbol{\gamma} = (2.2, 2.5, -1.9)^\top$, giving the average proportion of missing data of about 19.35%;

(b) $\rho = 0.5$, $\varepsilon_i \sim 0.7\mathcal{N}(0,1) + 0.3t(5)$, and the propensity score function $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})$ is taken as that given in setting (a), giving the average proportion of missing data of about 22.33%, where $t(5)$ denotes the Student's t distribution with five degrees of freedom;

(c) $\rho = 0.7$, $\varepsilon_i \sim \mathcal{N}(0, 0.5)$, and the propensity score function $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})$ is taken as $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma}) = \Phi(\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2})$, where $\Phi(\cdot)$ is the cumulative distribution

function of the standard normal distribution, and the true value of $\boldsymbol{\gamma}$ is set as $\boldsymbol{\gamma} = (1.3, 2.9, -1.9)^\top$, giving the average proportion of missing data of about 28.43%;

(d) $\rho = 0.7$, $\varepsilon_i \sim 0.7\mathcal{N}(0, 0.5) + 0.3t(5)$, and the propensity score function $\pi(\boldsymbol{U}_i; \boldsymbol{\gamma})$ is taken as that given in setting (c), giving the average proportion of missing data of about 28.25%.

For each of the 100 replicated data sets generated from each of the four settings, we use a penalized likelihood method and an appropriate data-driven approach to select the penalty parameter λ_n. This enables us to evaluate the estimate of $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_q)^\top$ for $q = p$, and we use the proposed model-averaging approach to compute $\hat{\boldsymbol{\mu}}$. To select the penalty parameter λ_n, we consider the following high-dimensional BIC-type criterion: $\mathrm{BIC}(\lambda_n) = -2l_n(\hat{\boldsymbol{\gamma}}_{\lambda_n}) + |\mathscr{A}_{\lambda_n}|\{\log(n) + 2\log(q)\}$, where $\hat{\boldsymbol{\gamma}}_{\lambda_n}$ is the PLE of $\boldsymbol{\gamma}$, given the penalty parameter λ_n, $\mathscr{A}_{\lambda_n}$ is the index set of nonzero components of $\hat{\boldsymbol{\gamma}}_{\lambda_n}$, and $|\mathscr{A}_{\lambda_n}|$ is the cardinality of the set $\mathscr{A}_{\lambda_n}$. Thus, we select the tuning parameter λ_n by minimizing $\mathrm{BIC}(\lambda_n)$. Prior to the model averaging, we sort the predictors using the proposed MI-SIS method, leading to $\widehat{\mathcal{M}}_{\varsigma_n}$ for $\varsigma_n = 100$. Then, we take $S = 10$, yielding a class of 10 candidate models, each with 10 predictors.

The results for the WMSE values under the four cases are given in Figures 1 and 2. First, the figures show that the proposed screening method behaves better than the IPW-SIS, MI-I, and MI-S methods, in the sense that it has the smallest WMSE median for the considered cases. This implies that the selection of the feature-screening methods in the initial step has a certain effect on the final model-averaging result (e.g., WMSE value). Second, the weighted model averaging with CV method behaves better than the model averaging with CV method for the CC data regardless of the restriction $\sum_{s=1}^{S} \omega_s = 1$. Third, the weighted model averaging with CV method without the restriction $\sum_{s=1}^{S} \omega_s = 1$ performs better than that with the restriction. Fourth, the weighted model averaging with CV method without the restriction $\sum_{s=1}^{S} \omega_s = 1$ has almost the same performance as the weighted model-averaging method of Ando and Li (2014), without adjusting the missing data. Fifth, the model averaging with AIC method behaves better than the model averaging with BIC method. Sixth, the group Lasso method performs best of the penalized likelihood methods, followed by the SCAD, MCP, and Lasso methods, in that order. Seventh, the group Lasso method outperforms the WMCV2 and CC2 methods. Eighth, our proposed weighted model-averaging method performs better than the model av-

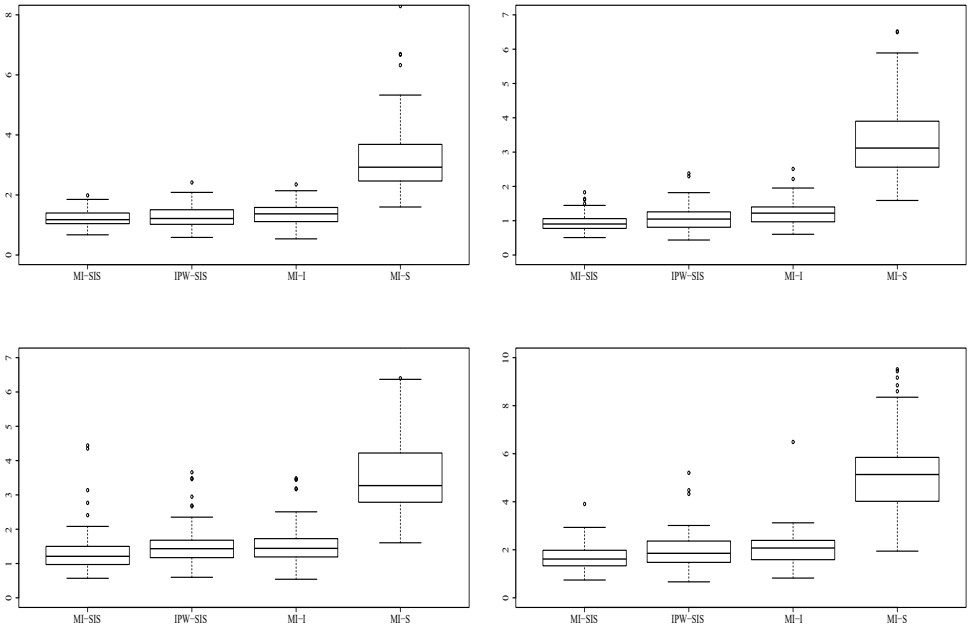

Figure 1. WMSE values of the proposed model-averaging method for four screening methods: case (a) (left upper panel), case (b) (right upper panel), case (c) (left lower panel), and case (d) (right lower panel) in Experiment 1.

eraging with the delete-one CV method for the CC data. That is, our proposed model-averaging procedure yields the best performance of the compared methods because it has the smallest median WMSE value.

Experiment 2. The main purpose of this experiment is to investigate the robustness of our proposed model-averaging method to a misspecified propensity score function. To this end, we consider the same linear regression as that given in Experiment 1, but we use different propensity score functions to create the missing data:

(e) $\text{logit}\{\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})\} = \gamma_0 + \sin(\gamma_1 X_{i1} + \gamma_2 X_{i2})$, with the true value of $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^\top$ taken as $\boldsymbol{\gamma} = (1.0, 1.8, -1.8)^\top$, giving the average proportion of missing data of about 27.38%;

(f) $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma}) = \Phi(\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2})$, with the true value of $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^\top$ taken as $\boldsymbol{\gamma} = (2.0, 2.2, -1.5)^\top$, giving the average missing proportion of about 13.60%.

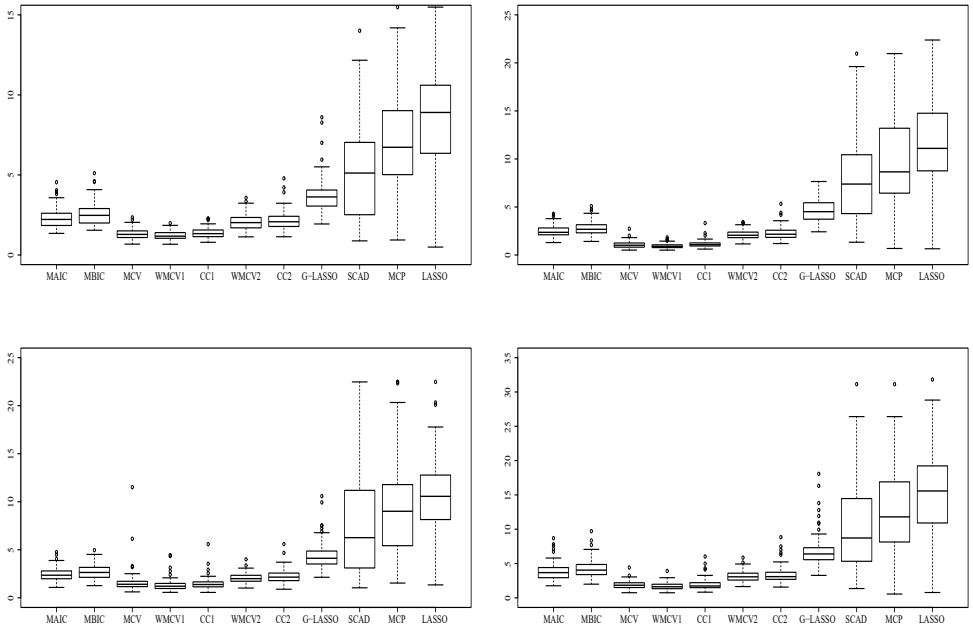

Figure 2. WMSE values of 11 model-averaging methods for four settings of ρ, the distribution of ε_i, and the propensity score function $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})$: case (a) (left upper panel), case (b) (right upper panel), case (c) (left lower panel), and case (d) (right lower panel) in Experiment 1.

For each of the 100 replicated data sets generated from each of the two settings, we calculate the corresponding results based on the propensity score function: $\text{logit}\{\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})\} = \gamma_1 X_{i1} + \gamma_2 X_{i2} + \cdots + \gamma_q X_{iq}$, with $q = p$, using the proposed model-averaging method. The results are presented in Figures 3 and 4. The figures show similar patterns to those of Figures 1 and 2. This implies that the proposed feature-screening method and model-averaging method are robust to a misspecification of the propensity score function.

4.2. Real-data example

In this subsection, we use the rate eye microarray expression data set (Scheetz et al. (2006)), available from `http://www.ncbi.nlm.nih.gov/geo`, to illustrate the proposed model-averaging method. For this data set, 120 12-week-old male rats were selected for tissue harvesting from the eyes and for microarray analysis. The microarrays used to analyze the RNA from the eyes of these rats contain over 31,042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Ar-

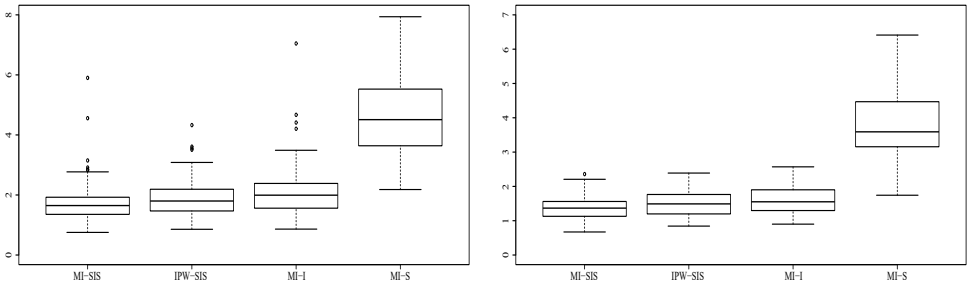

Figure 3. WMSE values of the proposed model-averaging method for four screening methods: case (e) (left panel) and case (f) (right panel) in Experiment 2.

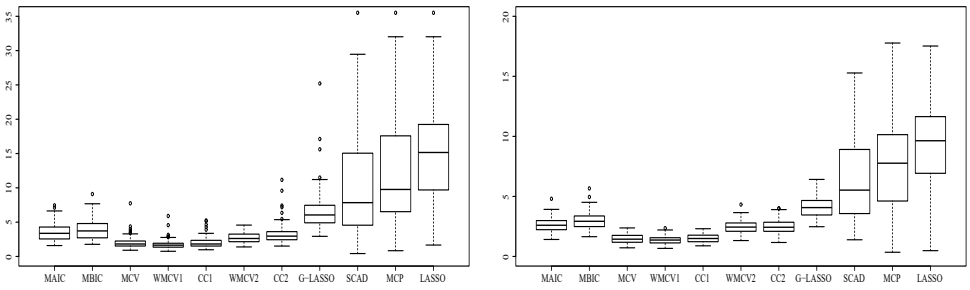

Figure 4. WMSE values of 11 model-averaging methods for two settings of ρ, the distribution of ε_i, and the propensity score function $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})$: case (e) (left panel) and case (f) (right panel) in Experiment 2.

ray). The intensity values were normalized using the robust multi-chip averaging method to obtain summary expression values for each probe set. Gene expression levels were analyzed on a logarithmic scale. To investigate genetic variation in human eye disease, Scheetz et al. (2006) applied the expression quantitative trait locus mapping method to 18,976 probes that are considered "sufficiently variable" and that exhibit at least a two-fold variation in expression level among the 120 male rats. The main interest of this study is to find the genes that are correlated with the gene TRIM32, which was recently found to cause Bardet–Biedl syndrome (Chiang et al. (2006); Huang, Ma and Zhang (2008)). Chiang et al. (2006) found that the gene TRIM32 at probe 1389163_at, which is regarded as the response variable ($\boldsymbol{Y}$), is critical to Bardet–Biedl syndrome, a genetic human disease concerning the retina. Our purpose is to find which probes among

the remaining 18,975 probes are most associated with TRIM32. In this case, the sample size is $n = 120$ and the number of probes is $p = 18{,}975$; thus, $p \gg n$, and this is a sparse, high-dimensional regression problem. Hence, a screening procedure is required to screen out most of relevant genes before an elaborative second-stage analysis. To roughly unify the scales, the selected gene expressions are standardized.

For this data set, we consider the linear regression model: $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i$, for $i = 1, \ldots, n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$. Because there is no missing data in the original data set, to illustrate the proposed model-averaging method in the presence of responses MAR, we artificially create missing responses using the following missingness data-mechanism model: $\text{logit}\{\pi(\boldsymbol{U}_i; \boldsymbol{\gamma})\} = \gamma_0 + \boldsymbol{\gamma}_1^\top \boldsymbol{U}_i$, where $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_1^\top)^\top$, with γ_0 an interception term and $\boldsymbol{\gamma}_1 = (\gamma_{11}, \ldots, \gamma_{1q})^\top$, and $\boldsymbol{U}_i = (X_{i1}, \ldots, X_{iq})^\top$ is a subvector of $\boldsymbol{X}_i$, with $q = 1{,}000$. The true value of $\boldsymbol{\gamma}$ is taken as $\boldsymbol{\gamma} = (1.5, 2.2, -1.9, 2.8, -1.8, 2.5, \mathbf{0}_{q-5}^\top)^\top$. Thus, the missing proportion is about 33%.

Our main interest is to investigate the prediction performance of the proposed model-averaging method. Therefore, we randomly divide the data into a training set with $n_1 = 80$ for model fitting and a testing set with $n_2 = 40$. To simultaneously estimate $\boldsymbol{\gamma}$ and identify the nonzero components in $\boldsymbol{\gamma}_1$ for the training set using the penalized likelihood method, we select the penalty parameter λ_n by minimizing the following BIC criterion: $\text{BIC}(\lambda_{n_1}) = -2l_{n_1}(\hat{\boldsymbol{\gamma}}_{\lambda_{n_1}}) + |\mathscr{A}_{\lambda_{n_1}}|\{\log(n_1) + 2\log(q)\}$, where $\hat{\boldsymbol{\gamma}}_{\lambda_{n_1}}$ is the penalized likelihood estimation of $\boldsymbol{\gamma}$, given the penalty parameter λ_{n_1}, $\mathscr{A}_{\lambda_{n_1}}$ is the index set of nonzero components of $\hat{\boldsymbol{\gamma}}_{\lambda_{n_1}}$, and $|\mathscr{A}_{\lambda_{n_1}}|$ is the cardinality of the set $\mathscr{A}_{\lambda_{n_1}}$. For comparison, we consider the 10 methods (MAIC, MBIC, WMCV1, CC1, WMCV2, CC2, G-LASSO, SCAD, MCP, LASSO) presented in the simulation studies for the training data set. For the MAIC, MBIC, WMCV1, CC1, WMCV2, and CC2, we first sort the genes using the MI-SIS procedure, yielding $\widehat{\mathcal{M}}_{\varsigma_n}$ for $\varsigma_n = 200$. Then we set $S = 20$, leading to a class of 20 candidate models, each with 10 genes.

We assess the prediction performance of the considered 10 methods using the following weighted mean squared prediction error (WMSPE):

$$\text{WMSPE} = \frac{1}{N_T} \sum_{1 \le i \le n, i \in \mathcal{T}} \frac{\delta_i}{\pi(\boldsymbol{U}_i; \hat{\boldsymbol{\gamma}}_{\lambda_{n_1}})} \{Y_i - \hat{\mu}_i(\hat{\boldsymbol{\omega}})\}^2,$$

where $N_T = \sum_{1 \le i \le n, i \in \mathcal{T}} \delta_i$, $\hat{\boldsymbol{\omega}}$ denotes the optimal weights evaluated by the CV method based on the training data set, and $\mathcal{T} = \{i\text{: the } i\text{th sample belongs to the}$

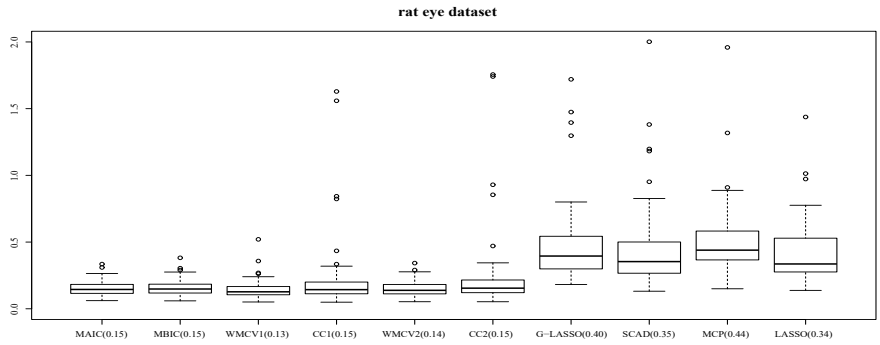

Figure 5. WMSPE values of 10 model-averaging methods in the rat eye data set. The number in brackets is the median of the distribution of the WMSPE.

testing set}. We repeat the entire procedure 100 times, and obtain 100 WMSPE values for each of 10 methods. The results are presented in Figure 5, and show that our proposed model-averaging method has best predictive efficiency of the methods considered, including those with the delete-one CV method based on the CC data, the classical model-averaging methods, and the penalized likelihood methods.

5. Conclusion

This study investigates the prediction accuracy problem for ultrahigh-dimensional linear regression models in the presence of responses MAR, and proposes a two-step model-averaging procedure to improve the prediction accuracy. The first step constructs the candidate models for averaging. To implement this step, we developed a novel feature-screening procedure in the presence of responses MAR to separate the active and inactive predictors based on the multiple-imputation sure independence index. Under some regularity assumptions, we showed its sure screening property and ranking consistency property. The proposed screening procedure is robust to a misspecification of the propensity score function. The second step identifies the optimal weights for averaging. To implement the second step, we first adopted the PS-LS method to estimate the regression parameters for each candidate model. Then we proposed a WDCV criterion without the restriction $\sum_{s=1}^{S} \omega_s = 1$ to select the optimal weights. Under some regularity assumptions, we proved that the proposed WDCV criterion is asymptotically equivalent to the weighted squared error, which is our theoretical basis for using

the model-averaging method.

In addition, to simultaneously estimate the regression coefficients in $\boldsymbol{\gamma}$ and select the important covariates in a parametric propensity score function in a high-dimensional setting, we have proposed a penalized likelihood method based on some proper penalty function. To select the tuning parameter λ_n in the penalized likelihood function, we use a data-driven approach, such as the BIC, in numerical studies. Under some regularity conditions, we proved the oracle properties, including the sparsity and asymptotic normality, of the proposed penalized likelihood estimator of $\boldsymbol{\gamma}$.

Simulation studies and an example are used to illustrate the proposed model-averaging method based on criteria such as the WMSE and the WMSPE. The results show that the proposed method outperforms 10 other approaches, including existing model-averaging methods.

The proposed MI-SIS approach used to screen the important predictors in an ultrahigh-dimensional linear regression model in the presence of responses MAR is a nonparametric screening method. However, it is unclear how to extend the proposed screening procedure to a non-ignorable missing data case, which is widely encountered in practice. In addition, their theoretical properties remain unknown in the presence of non-ignorable missing data.

Supplementary Material

The online Supplementary Material includes the properties of the penalized likelihood estimator, the proposed screening procedure, and all technical proofs.

Acknowledgments

The authors are grateful to the Editor, the Associate Editor, and two referees for their valuable suggestions. This work was supported by grants from the National Natural Science Foundation of China (Grant No.: 11671349), and the National Social Science Foundation of China (Grant No.: 17BTJ038) and the Key Projects of the National Natural Science Foundation of China (Grant No.: 11731011).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceeding of the Second International Symposium on Information Theory*, 267–281. Akademiai Kiado, Budapest.

Akaike, H. (1979). A Bayesian extension of minimum AIC procedure of autoregressive model fitting. *Biometrika* **66**, 237–242.

Ando, T. and Li, K. C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association* **109**, 254–265.

Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics* **41**, 2123–2148.

Chang, J., Tang, C. Y. and Wu, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *The Annals of Statistics* **44**, 515–539.

Chiang, A. P., Beck, J. S., Yen, H. J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, Kim, K.-Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D. C., Casavant, T. L., Stone, E. M. and Sheffield, V. C. (2006). Homozygosity mapping with SNP arrays identifies a novel gene for bardet-biedl syndrome gene (bbs11). *Proceedings of the National Academy of Sciences of the United States of America* **103**, 6287–6292.

Dardanoni, V., Modica, S. and Peracchi, F. (2011). Regression with imputed covariates: a generalized missing indicator approach. *Journal of Economics* **162**, 362–368.

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

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

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

Fang, F., Lan, W., Tong, J. and Shao, J. (2017). Model averaging for prediction with fragmentary data. *Journal of Business & Economic Statistics* **37**, 517–527.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica* **75**, 1175–1189.

Hansen, B. E. and Racine, J. (2012). Jackknife model averaging. *Journal of Economics* **167**, 38–46.

He, X., Wang, L. and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics* **41**, 342–369.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* **14**, 382–417.

Huang, J., Ma, S. and Zhang, C. H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603–1618.

Ibrahim, J. G., Zhu, H. and Tang, N. (2008). Model selection criteria for missing data problems using EM algorithm. *Journal of the American Statistical Association* **103**, 1648–1658.

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

Lan, W., Ma, Y., Zhao, J., Wang, H. and Tsai, C. L. (2018). Sequential model averaging for high dimensional linear regression models. *Statistica Sinica* **28**, 449–469.

Lee, S. Y. and Tang, N. S. (2006). Bayesian analysis of nonlinear structural equation models with nonignorable missing data. *Psychometrika* **71**, 541–564.

Li, K. C. (1987). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics* **14**, 1011–1112.

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

Little, R. J. A. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data.* 3rd Edition. John Wiley & Sons Inc., New York.

Liu, Q., Okui, R. and Yoshimura, A. (2016). Generalized least squares model averaging. *Econometric Reviews* **35**, 1692–1752.

Schomaker, M., Wan, A. T. K. and Heumann, C. (2010). Frequentist model averaging with missing observations. *Computational Statistics and Data Analysis* **54**, 3336–3347.

Scheetz, T. E., Kim, K.-Y., Swiderski, R. E., Philp1, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C. and Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to rye disease. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 14429–14434.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

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

Wan, A. T. K., Zhang, X. and Zou, G. (2010). Least squares model averaging by mallows criterion. *Journal of Economics* **156**, 277–283.

Wang, D. and Chen, S. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics* **37**, 490–517.

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

Xie, J., Lin, Y., Yan, X. and Tang, N. (2020). Category-adaptive variable screening for ultra-high dimensional heterogeneous categorical data. *Journal of the American Statistical Association* **115**, 747–760.

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

Zhang, X. (2013). Model averaging with covariates that are missing completely at random. *Economic Letters* **121**, 360–363.

Zhang, X., Yu, D., Zou, G. and Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association* **111**, 1775–1790.

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

Jinhan Xie

Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, Yunnan University, Kunming, 650500, P. R. of China.

E-mail: jinhanxie@163.com

Xiaodong Yan

School of Economics, Shandong University, Jinan, 250100, China.

E-mail: yanxiaodong@sdu.edu.cn

Niansheng Tang

Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, Yunnan University, Kunming, 650500, P. R. of China.

E-mail: nstang@ynu.edu.cn

(Received June 2018; accepted August 2019)