Statistica Sinica Preprint No: SS-2018-0297	
Title	A Model-averaging method for high-dimensional regression with missing responses at random
Manuscript ID	SS-2018-0297
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0297
Complete List of Authors	Jinhan Xie Xiaodong Yan and Niansheng Tang
Corresponding Author	Niansheng Tang
E-mail	nstang@ynu.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

A Model-averaging method for high-dimensional regression with missing responses at random

Jinhan Xie[a] Xiaodong Yan[b] and Niansheng Tang[a]

[a] *Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University*
[b] *School of Economics, Shandong University*

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

Key words and phrases: High-dimensional Data; Multiple Imputation; Missing at Random; Model Averaging; Screening; Weighted delete-one cross-validation.

1. Introduction

Model selection and model averaging are two popular approaches to improve prediction accuracy in regression analysis. Model selection is often implemented by using some proper criterion, such as the Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Schwarz, 1978), to select just the best model among a set of candidate models. Since these model selection methods ignore the contribution of other candidate models, it 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 criteria may lead to different best models, which indicates that statistical inference based on the final model would be variation from data set to data set. To address the aforementioned issue, a model averaging approach has been proposed to improve the prediction accuracy in that it provides the pooling of predictions by giving higher weights to the better models, by which, it often reduces the bias in regression prediction, instead of depending on only one best model, and it avoids ignoring useful information from the form of the relationship between response and covariates (Zhang, 2013). Various model averaging approaches have been developed over the past years. For example, see AIC model averaging (Akaike, 1979), BIC model averaging (Hoeting, 1999), Mallows model

averaging (Hansen, 2007; Wan et al., 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 et al., 2016). However, the aforementioned methods are developed for the case that the dimension of predictors is less than sample size, and can not be directly applied for high/ultrahigh dimensional data.

High dimensional data that the number of predictors is much larger than sample size are often encountered in various fields such as biomedicine, social science and economics. Statistical analysis of high-dimensional data is quite challenging. To make inference on statistical models with high-dimensional data, many penalized methods have been developed to simultaneously select important predictors and estimate unknown parameters in the considered models. For example, see Lasso (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), minimax concave penalty (MCP) (Zhang, 2010). For the case that the dimensionality of predictors grows exponentially fast with sample size, some feature screening methods have been developed to largely reduce the dimensionality of predictors to a moderate scale so that classical statistical inference methods can be applied to the reduced models. For example, see Fan and Lv (2008), Fan and Song (2010) and Chang et al. (2013) for model-based feature screening methods; Zhu et

al. (2011), Li et al. (2012), He et al. (2013), Chang et al. (2016), Yan et al. (2018) and Xie et al. (2019) for model-free feature screening approaches. A little work has been done on 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 make stable predictions for high dimensional linear regression models. However, existing model averaging methods for high-dimensional regression models mainly focus on the fully observed data.

Missing data are frequently encountered in various settings, such as surveys, clinical trials and longitudinal studies, due to various reasons such as unwillingness of some sampled individuals to answer sensitivity questions, loss of information caused by uncontrollable factors, some scheduled visits intermittently or drop out of the study (Little and Rubin, 2002). Ignoring missing data may lead to prediction bias. To address the issue, many model selection or model averaging methods have been developed to improve the prediction accuracy in the presence of missing data. For example, Ibrahim et al. (2008) developed a novel model selection criterion for missing data problem based on the EM algorithm; Schomaker et al. (2010) presented two approaches to

handle missing data for model averaging problem; Dardanoni et al. (2011) adopted model-averaging approach to tackle the bias-precision trade-off with in the presence of missing covariate values in linear regression models; Zhang (2013) proposed using Mallows model averaging approach to handle missing completely covariates at random; Fang et al. (2017) presented a model averaging approach in the context of fragmentary data. However, the aforementioned works have been developed for classical setting that the number of predictors is fixed and less than sample size. To our knowledge, there is little work on model-averaging for ultrahigh dimensional regression models with missing responses at random.

This paper proposes a two-step model averaging approach for ultrahigh dimensional regression models in the presence of missing responses at random. The first step is to construct 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 of 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 the misspecification of propensity score function. The second step

is to find the optimal weights for averaging a set of candidate models via the weighted delete-one cross-validation criterion (WDCV). 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 regression coefficients and select the significant covariates in the assumed parametric propensity score function. Also, a data-driven approach such as the BIC criterion in numerical studies is presented to select the tuning parameter in the penalized likelihood function. Under some regularity assumptions, we have proved oracle properties of the proposed penalized likelihood estimators of parameters including the sparsity and asymptotic normality.

The rest of this paper is organized as follows. In Section 2, we describe model setting and present a two-step model averaging procedure in the presence of missing responses at random. In Section 3, we systematically investigate 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 through some simulation studies and a real data example. Some concluding remarks are made in Section 5. Technical details are given in the Appendix.

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 $\boldsymbol{X}_i$'s are fully observed, whilst Y_i's 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 response variable and 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's are the independent random errors 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, i.e., $\log(p) = o(n^v)$ for some constant $v \in (0, 1)$. In this case,

it is recognized that only a few predictors may indeed contribute to Y_i, i.e., the model (2.1) has a sparse structure. Thus, some feature screening method should be employed to select the important predictors.

For missing data Y_i's, we assume that δ_i is independent of δ_j for any $i \neq j$, and δ_i only depends on some components of $\boldsymbol{X}_i$, i.e., missingness data mechanism is missing at random (MAR). But, in applications, it is rather difficult to determine which components of $\boldsymbol{X}_i$ contribute to missingness of Y_i. More importantly, it is recognized that only a few covariates may indeed contribute to missingness of Y_i (Lee and Tang, 2006). Generally, one may initially incorporate a lot of covariates to specify missingness data mechanism, and then adopt a penalized method to identify important covariates contributing to missingness of Y_i. For example, following a lot of missing data literatures, we consider a 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, $\pi(\cdot)$ is the selection probability function, and $\boldsymbol{U}_i$ 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., covariates indeed contribute to missingness of Y_i) may be different from those of $\boldsymbol{X}_i$ (i.e., predictors indeed contribute to Y_i). As an illustration, we may 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 $\alpha \in (0, 1/2)$. In this case, it is again assumed that the aforementioned missingness data mechanism model has a sparse structure.

Under MAR assumption defined above, some penalized methods such as Lasso, Adaptive Lasso and SCAD methods can be employed to evaluate maximum likelihood estimation (denoted as $\hat{\boldsymbol{\gamma}}$) of $\boldsymbol{\gamma}$. To wit, $\hat{\boldsymbol{\gamma}}$ 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 bias and 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 of

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

where $a > 2$ and $\lambda_n > 0$ are the tuning parameters, $b_+ = bI(b \geq 0)$, and $I(\gamma \leq \lambda_n)$ is an indictor function of the event $\{\gamma \leq \lambda_n\}$, which takes 1 if $\gamma \leq \lambda_n$ and 0 otherwise. Fan and Li (2001) proposed taking $a = 3.7$ from

a Bayesian point of view. Parameter λ_n can be taken by some 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 applications, both d and the set of true predictors $\mathscr{A}_\beta = \{j : |\beta_j| > 0\}$ are unknown. To improve the prediction accuracy for the considered model (**??**), a widely used method is the model averaging procedure. Existing literature on model averaging has mainly focused on the settings that there is no missing data or a low-dimensional linear regression. In what follows, we extend the model averaging approach to the setting simultaneously including missing responses at random and high-dimensional linear regression. Thus, to improve the accuracy in predicting mean of Y in a high-dimensional linear regression in the presence of missing responses at random, we propose the following two-step model averaging procedure.

(1) The first step: construct candidate models

The first-step model averaging procedure is to construct 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 a $n \times p_s$ 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 $\boldsymbol{\beta}_s$. To wit, under the aforementioned assumption, the PS-LS estimator $\tilde{\boldsymbol{\beta}}_s$ of $\boldsymbol{\beta}_s$ can be obtained by

$$\tilde{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta}_s} (\boldsymbol{Y} - \boldsymbol{X}_s\boldsymbol{\beta}_s)^\top \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{X}_s\boldsymbol{\beta}_s),$$

where $\boldsymbol{W} = \mathrm{diag}(\delta_1/\pi_1, \ldots, \delta_n/\pi_n)$ 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 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 $\hat{\boldsymbol{\gamma}}$ to replace $\boldsymbol{\gamma}$. Thus, the corresponding estimator of $\boldsymbol{\mu}_s$ has the form of $\hat{\boldsymbol{\mu}}_s = \boldsymbol{X}_s\hat{\boldsymbol{\beta}}_s = \boldsymbol{X}_s(\boldsymbol{X}_s^\top \widehat{\boldsymbol{W}} \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s^\top \widehat{\boldsymbol{W}} \boldsymbol{Y}$, 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 S candidate models introduced above, we obtain S PS-LS predictions of the mean of response variable, i.e., $\{\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_S\}$. Given a weight vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_S)^\top \in \mathcal{W} = \{\boldsymbol{\omega} \in [0,1]^S : 0 \leq \omega_s \leq 1\}$, the model averaging predictor of the mean of

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 of model averaging, one usually assumes that $\sum_{s=1}^{S} \omega_s = 1$. Here, we omit this assumption. The reasons for relaxing the restriction $\sum_{s=1}^{S} \omega_s = 1$ can see Ando and Li (2014).

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

Without loss of generality, it is assumed that covariates have been standardized, and $Y \perp \delta | X_k$ (e.g., see He et al., 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 missing data in the marginal utility. Thus, for $k = 1, \ldots, p$, we define the estimated marginal multiple-imputation sure independence screening (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 indictor function. Following the argument of Wang and Chen (2009), effectively $\tilde{Y}_{iv}^{k}$ 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 $\hat{r}_k$ via (2.4) for $k = 1, \ldots, p$. Then, we can sort the magnitudes of $\hat{r}_k$'s in a decreasing order, and select the important predictors via the following 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 are the pre-specified threshold value. Based on the above defined feature screening criterion, the full model with p predictors may shrink to a reduced model with the number of predictors less than n.

Based on the calculated MI-SIS statistics between response variable and each of 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 the predictors with the MI-SIS values falling into the sth group. We drop the $(S+1)$th group and only use the first S groups to conduct model averaging. To wit, the number of the candidate models is S.

(2) The second step: determine the optimal weights

The key task for evaluating $\hat{\boldsymbol{\mu}}(\boldsymbol{\omega})$ is to find the optimal weights ω_s's. Many methods such as the CV and GCV methods can be used to implement the task. Here the delete-one CV approach is adopted to evaluate the optimal weights due to its asymptotic optimality theory for heteroscedasticity error. Let $\tilde{\mu}_s^{(-i)}$ be the predicted value of the mean of response variables computed with the ith observation $(\boldsymbol{X}_i, Y_i, \delta_i)$ deleted 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 $\overline{\boldsymbol{P}}_s = \widehat{\boldsymbol{W}}^{\frac{1}{2}} \boldsymbol{X}_s (\boldsymbol{X}_s^\top \widehat{\boldsymbol{W}} \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s^\top \widehat{\boldsymbol{W}}^{\frac{1}{2}}$. 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 $\widetilde{\boldsymbol{P}}_s = \widehat{\boldsymbol{D}}_s (\widehat{\boldsymbol{P}}_s - \boldsymbol{I}) + \boldsymbol{I}$, and $\widehat{\boldsymbol{D}}_s = \mathrm{diag}(\hat{d}_1^s, \ldots, \hat{d}_n^s)$ in which $\hat{d}_i^s = 1/(1-\hat{h}_{ii}^s)$ 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 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 $\widetilde{\boldsymbol{P}}(\boldsymbol{\omega}) = \sum_{s=1}^{S} \omega_s \widetilde{\boldsymbol{P}}_s$. Similar to Hansen and Racine (2012), to incorporate the information associated with 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 a $S \times 1$ vector with the sth component $\mathscr{A}_s = \boldsymbol{Y}^\top \widetilde{\boldsymbol{P}}_s \widehat{\boldsymbol{W}} \boldsymbol{Y}$, and $\mathscr{B}$ is a $S \times S$ 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 $\mathcal{W}$, i.e.,

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

Unlike other cross validation problems that are often time-consuming, some numerous software packages, such as the quadprog package in R and Matlab, are available for evaluating a solution to the above quadratic optimization problem. Even if S is quite large, evaluating the solution to (2.5) is also fast using anyone of these packages. To wit, the above proposed optimization problem is computationally feasible. Based on the optimal weights evaluated above, the model averaging predictor of the mean of 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

Theoretical properties of the penalized likelihood estimator $\hat{\boldsymbol{\gamma}}$ and the preceding proposed feature screening procedure can be found in the supplemental materials. In what follows, we investigate asymptotic properties of the above 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 $L(\boldsymbol{\omega}) = \{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}^\top \widehat{\boldsymbol{W}} \{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}$, whose risk function is $R(\boldsymbol{\omega}) = E[\{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}^\top \widehat{\boldsymbol{W}} \{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}|\boldsymbol{X}]$. 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. Notations C_0, $C_1, \ldots, C_5$ are some appropriate constant, $\phi(\cdot)$ represents the maximal diagonal element of a matrix, and p_s denotes the number of columns of matrix $\boldsymbol{X}_s$. To obtain asymptotic properties of the above proposed model averaging procedure with the weighted delete-one CV 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 of matrix $\boldsymbol{A}$, respectively.

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) $\sup_{1 \leq s \leq S} \{\phi(\boldsymbol{P}_s)/p_s\} \leq C_5/n$; (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, and the lower bound guarantees that weights do not go to infinity as the sample size increases and the proposed parametric weights are asymptotically consistent. Assumption 2 states that design matrix is uniformly bounded, the nonsingular assumption is necessary for ensuring the existence of hat matrix. Assumption 3(i) is a moment condition on 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). Imposing on this restriction is the cost of 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 et al. (2010) and Ando and Li (2014). Assumption 3(v) excludes extremely unbalanced designs for each of candidate

models, which is the same as Condition (5.2) of Li (1987). Assumption 3(vi) indicates that $\xi_n \to \infty$, i.e., there is not a finite approximating model for which 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 $\phi \geq 0$, 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 growing 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 above proposed weighted delete-one CV criterion for selecting the optimal weights is asymptotically equivalent to the weighted squared error. Thus, the above 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 $\mathcal{W}$.

5. 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; and then an example is used to illustrate the proposed methodologies.

5.1 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

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

where $\boldsymbol{\mu}_{i0}$ is true value of the mean of response variable $\boldsymbol{Y}$ given $\boldsymbol{X}_i$, and $\hat{\boldsymbol{\mu}}_i^{(k)}$ is the estimated mean of 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 the initial step, we calculate the WMSEs for the proposed feature screening procedure (MI-SIS) and existing feature screening methods such as the inverse probability weighted sure independence screening method (denoted by 'IPW-SIS' method, Lai et al., 2017), the borrowing missingness information (BMI) containing missing indicator surrogate feature screening method (denoted by 'MI-S'), and the missing indicator imputation screening method (denoted by 'MI-I', Wang and Li, 2018). Second, for the selected predictors via the proposed feature screening procedure, we compare the performance of the proposed model-averaging approach together

with the following methods: (A) model-averaging with the Akaike information criterion (AIC) under the restriction $\sum_{s=1}^{S}\omega_s = 1$ (denoted as 'MAIC'); (B) model-averaging with the Bayesian information criterion (BIC) under the restriction $\sum_{s=1}^{S}\omega_s = 1$ (denoted as 'MBIC'); (C) weighted model averaging method of Ando and Li (2014) without adjusting the missing data (denoted as 'MCV'); (D) weighted model-averaging with the CV method without the restriction $\sum_{s=1}^{S}\omega_s = 1$ (denoted as 'WMCV1'); (E) model-averaging with the CV method for the CC data without the restriction $\sum_{s=1}^{S}\omega_s = 1$ (denoted as 'CC1'); (F) weighted model-averaging with the CV method under the restriction $\sum_{s=1}^{S}\omega_s = 1$ (denoted as 'WMCV2'); (G) model-averaging with the CV method for the CC data under the restriction $\sum_{s=1}^{S}\omega_s = 1$ (denoted as 'CC2'); (H) penalized likelihood method with the SCAD (denoted as 'SCAD'); (I) penalized likelihood method with the MCP (denoted as 'MCP'); (J) penalized likelihood method with the Lasso (denoted as 'Lasso'); (K) penalized likelihood method with the group Lasso (denoted as 'G-Lasso'), which is implemented by partitioning p predictors into $S+1$ groups, and the first S groups are the same as those obtained in model-averaging procedure. To implement the proposed feature screening procedure, we take the Gaussian kernel function $K(u) = \exp(-u^2/2)/(2\pi)^{1/2}$, and select the bandwidth via the cross-validation 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 predictors. Here, $\boldsymbol{X}_i$'s are generated from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with components of $\boldsymbol{\Sigma} = (\sigma_{jk})_{p\times p}$ being $\sigma_{jk} = \rho^{|j-k|}$ for $1 \le j, k \le p$. True values of nonzero β_j's are independently sampled from the normal distribution $\mathcal{N}(0, 0.5^2)$. Thus, the mean of response variables is $\boldsymbol{\mu} = (\boldsymbol{X}_1^\top \boldsymbol{\beta}, \ldots, \boldsymbol{X}_n^\top \boldsymbol{\beta})^\top$. We assume that $\boldsymbol{X}_i$'s are completely observed, but Y_i's are subject to missingness. The missing indicators δ_i of Y_i are independently generated 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 $n = 60$, $p = 1000$ and $d = 50$, and assume that the true index set of nonzero β_j's 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 true value of $\boldsymbol{\gamma}$ is taken as $\boldsymbol{\gamma} = (2.2, 2.5, -1.9)^\top$ giving the average proportion of missing data 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 the setting (a) giving the average proportion

of missing data about 22.33%, where $t(5)$ denotes the Student's t distribution with five degrees of freedom;

(c) $\rho = 0.7$, $\varepsilon_i \sim \mathcal{N}(0, 0.5)$, while 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, true value of $\boldsymbol{\gamma}$ is set as $\boldsymbol{\gamma} = (1.3, 2.9, -1.9)^\top$ giving the average proportion of missing data 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 the setting (c) giving the average proportion of missing data about 28.25%.

For each of 100 replicated data sets generated from each of the above presented four settings, the preceding developed penalized likelihood method together with some appropriate data-driven approach to select the penalty parameter λ_n is adopted to evaluate the estimate of $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_q)^\top$ with $q = p$, and the above developed model-averaging approach is employed 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, and $\mathscr{A}_{\lambda_n}$ is the index set of nonzero components of $\hat{\boldsymbol{\gamma}}_{\lambda_n}$, $|\mathscr{A}_{\lambda_n}|$ 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 model averaging, we first sort predictors utilizing the above developed MI-SIS method, which leads to $\widehat{\mathcal{M}}_{\varsigma_n}$ for $\varsigma_n = 100$, and then take $S = 10$ to yield a class of 10 candidate models, each with 10 predictors.

Results for WMSE values under four cases are given in Figures 1 and 2. Examination of Figures 1 and 2 shows that (i) the proposed screening method behaves better than the IPW-SIS, MI-I and MI-S methods in the sense that the former has the smaller WMSE median than the latter for the considered cases, which implies the selection of the feature screening methods in the initial step has a certain effect on the final result (e.g., WMSE value) for model-averaging; (ii) weighted model-averaging with CV method behaves better than model-averaging with CV method for the CC data regardless of with or without the restriction $\sum_{s=1}^{S} \omega_s = 1$; (iii) weighted model-averaging with CV method without the restriction $\sum_{s=1}^{S} \omega_s = 1$ performs better than that with the restriction; (iv) weighted model-averaging with CV method without the restriction $\sum_{s=1}^{S} \omega_s = 1$ has almost the same performance as weighted model averaging method of Ando and Li (2014) without adjusting the missing data; (v) model-averaging with AIC method behaves better than model-averaging with BIC method; (vi) group Lasso method outperforms other three penalized likelihood methods such as SCAD, MCP and Lasso, and SCAD method performs better than MCP and Lasso, while Lasso method behaves worst among

four penalized likelihood methods; (vii) group Lasso method has better performance than WMCV2 and CC2; (viii) our proposed weighted model-averaging method has better performance than model-averaging with the delete-one CV for the CC data. That is, our proposed model-averaging procedure yields best performance among the compared methods because our proposed model-averaging method has the smallest median of WMSE values.

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

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

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

For each of 100 replicated data sets generated from each of the above presented two settings, we calculate their corresponding results based on the propensity score function: $\text{logit}\{\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})\} = \gamma_1 X_{i1} + \gamma_2 X_{i2} + \ldots + \gamma_q X_{iq}$ with

$q = p$ using the above developed model-averaging method. Results are presented in Figures 3 and 4. Inspection of Figures 3 and 4 indicates that the same pattern is observed, which implies that the proposed feature screening method and model averaging method are robust to the misspecification of propensity score functions.

5.2 Real data example

In this subsection, the rate eye microarray expression dataset (Scheetz et al., 2006), which is available from http://www.ncbi.nlm.nih.gov/geo, is used to illustrate the above proposed model-averaging method. For this dataset, 120 twelve-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 rates contain over 31,042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array). 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 exhibit at least two-fold variation in expression level among the 120 male rates. The main

interest of this study is to find the genes that are correlated with the gene TRIM32, which was recently found to causes Bardet-Biedl syndrome (Chiang et al., 2006; Huang et al., 2008). Chiang et al. (2006) found that the gene TRIM32 at probe 1389163_at, which is regarded as response variable ($\boldsymbol{Y}$), is a critical gene to the 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 $p = 18,975$, which indicates that $p \gg n$. Thus, 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 dataset, we consider a linear regression model: $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \varepsilon_i$ for $i = 1, \ldots, n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$. Since there is not missing data in the original data set, to illustrate the above proposed model-averaging method in the presence of missing response at random, we artificially create missing response via the following missingness data mechanism model: $\text{logit}\{\pi(\boldsymbol{U}_i; \boldsymbol{\gamma})\} = \gamma_0 + \boldsymbol{\gamma}_1^\top \boldsymbol{U}_i$, where $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_1^\top)^\top$, γ_0 is 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 $q = 1000$. 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 above proposed model-averaging method, we randomly divide the data into a training set with $n_1 = 80$ for model fitting and a testing set with $n_2 = 40$. In simultaneously estimating $\boldsymbol{\gamma}$ and identifying 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: $\mathrm{BIC}(\lambda_{n_1}) = -2l_{n_1}(\hat{\boldsymbol{\gamma}}_{\lambda_{n_1}}) + |\mathscr{A}_{\lambda_{n_1}}|\{\log(n_1) + 2\log(q)\}$, where $\hat{\boldsymbol{\gamma}}_{\lambda_{n_1}}$ is the penalized likelihood estimation of $\boldsymbol{\gamma}$ given the penalty parameter λ_{n_1}, and $\mathscr{A}_{\lambda_{n_1}}$ is the index set of nonzero components of $\hat{\boldsymbol{\gamma}}_{\lambda_{n_1}}$, $|\mathscr{A}_{\lambda_{n_1}}|$ is the cardinality of the set $\mathscr{A}_{\lambda_{n_1}}$. For comparison, we consider ten methods (e.g., MAIC, MBIC, WMCV1, CC1, WMCV2, CC2, G-LASSO, SCAD, MCP, LASSO) presented in simulation studies for the training data set. For MAIC, MBIC, WMCV1, CC1, WMCV2 and CC2, we first sort genes utilizing the MI-SIS procedure yielding $\widehat{\mathcal{M}}_{\varsigma_n}$ for $\varsigma_n = 200$, and then set $S = 20$ to lead to a class of 20 candidate models, each with 10 genes.

We assess the prediction performance of the considered ten methods via the following weighted mean squared prediction errors (WMSPE):

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

where $N_T = \sum_{1 \le i \le n, i \in \mathcal{T}} \delta_i$, $\hat{\boldsymbol{\omega}}$ is the optimal weights evaluated by the CV

method based on the training data set, $\mathcal{T} = \{i$: the ith sample belongs to the testing set$\}$. We repeat the entire procedure 100 times, and obtain 100 WMSPE values for each of ten considered methods. Results are presented in Figure 5. Inspection of Figure 5 shows that our proposed model-averaging has best predictive efficiency among ten methods, and has better predictive efficiency than those with the delete-one CV method based on the CC data, the classical model-averaging methods and the penalized likelihood methods.

6. Discussion

This paper investigates the prediction accuracy problem for ultrahigh dimensional liner regression models in the presence of missing responses at random, and proposes a two-step model averaging procedure to improve the prediction accuracy. The first step is to construct the candidate models for averaging. To implement the first step, we have developed a novel feature screening procedure in the presence of missing responses at random to separate the active and inactive predictors based on the multiple-imputation sure independence index. Under some regularity assumptions, we have showed its sure screening property and ranking consistency property. The proposed novel screening procedure is robust to the misspecification of the propensity score function. The second step is to find the optimal weights for averaging. To implement the

second step, we have first adopted the PS-LS method to estimate regression parameters for each of candidate models, and have then proposed a weighted delete-one CV criterion without the restriction $\sum_{s=1}^{S} \omega_s = 1$ to select the optimal weights. Under some regularity assumptions, we have proved that the proposed weighted delete-one CV criterion for selecting the optimal weights is asymptotically equivalent to the weighted squared error, which is our theoretical basis for utilizing model-averaging method.

Also, to simultaneously estimate regression coefficients in $\boldsymbol{\gamma}$ and select 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 have given a data-driven approach such as the BIC in numerical studies. Under some regularity conditions, we have proved the oracle properties including the sparsity and asymptotic normality of the proposed penalized likelihood estimator of $\boldsymbol{\gamma}$.

Some simulation studies and an example are used to illustrate the proposed model-averaging method based on some criterions such the weighted mean squared errors and the weighted mean squared prediction errors. Results evidence that the proposed method behaves best among the considered ten approaches including the existing model-averaging methods.

The above proposed multiple-imputation sure independence screening approach to screen the important predictors in a ultrahigh-dimensional linear regression model in the presence of missing responses at random is a non-parametric screening method. But it is unclear how to extend the proposed screening procedure to a non-ignorable missing data case, which is widely encountered in various fields, and their theoretical properties still remain unknown in the presence of non-ignorable missing data.

Acknowledgements

The authors are grateful to the Editor, the Associate Editor and two referees for their valuable suggestions that greatly improved the manuscript. This work was supported by the grants from the National Natural Science Foundation of China (Grant No.: 11671349) and the Key Projects of the National Natural Science Foundation of China (Grant No.: 11731101).

Supplementary Materials

Supplementary materials include the properties of penalized likelihood estimator, the proposed screening procedure and all the technical proofs.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory* **pp**, 267-281.

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.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759-771.

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics* **64**, 1062-1069.

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* **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*, doi: 10.1080/07350015.2017.1383263.

Garcia, R. I., Ibrahim, J. G. and Zhu, H. (2010). Variable selection for regression models with missing data. *Statistica Sinica* **20**, 149-165.

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 Amer-*

ican 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. (1987). *Statistical analysis with missing data.* New York: John

Wiley, 2002.

Liu, Q., Okui, R. and Yoshimura, A. (2016). Generalized least squares model averaging. *Econometric Reviews* **0**, 1-61.

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

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* **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. (2019). Category-adaptive variable screening for ultra-high dimensional heterogeneous categorical data. *Journal of the American Statistical Association* (in press, doi.org/10.1080/01621459.2019.1573734).

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

Zhang, C. 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., Zou, G. and Liang, H. (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika* **101**, 205-218.

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, Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University,

Kunming, 650500, P. R. of China

E-mail: jinhanxie@163.com

School of Economics, Shandong University, Jinan, P. R. of China

E-mail: yanxiaodong@sdu.edu.cn

Niansheng Tang, Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming, 650500, P. R. of China

E-mail: nstang@ynu.edu.cn

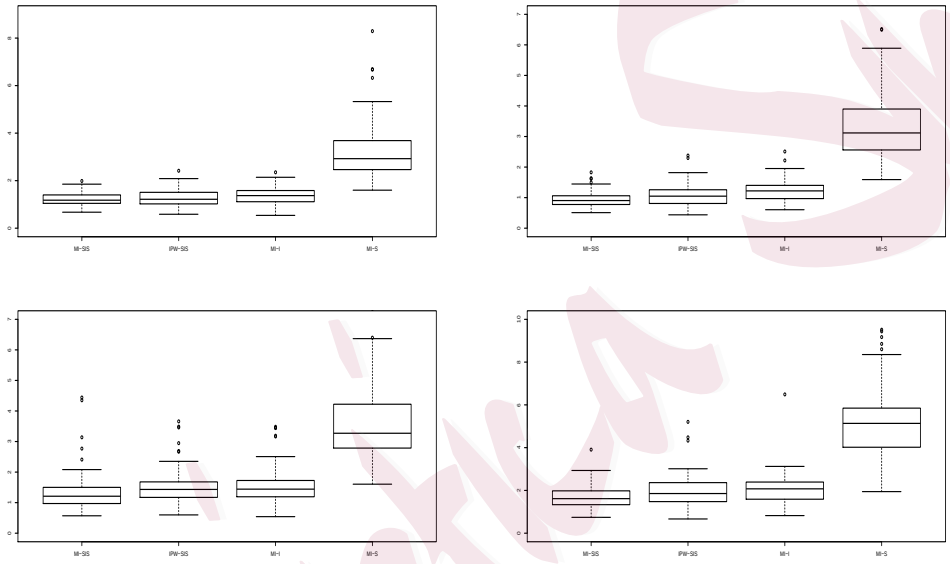

Figure 1: WMSE values of the proposed model averaging method for four different screening methods: case (a) (left upper panel), case (b) (right upper panel), case (c) (left lower panel) and case (d) (right lower panel) in Experiment 1

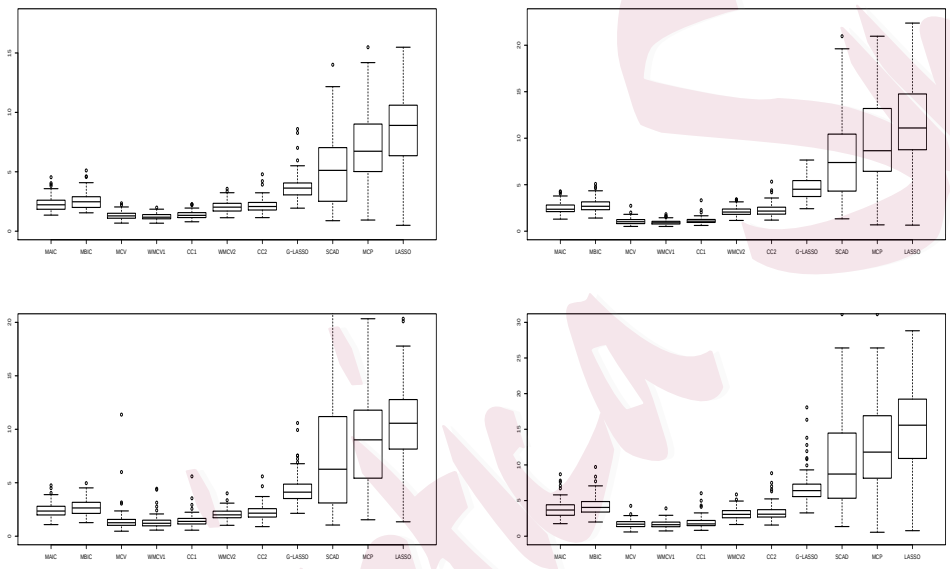

Figure 2: WMSE values of eleven model-averaging methods for four settings of ρ, the distribution of ε_i and propensity score function $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})$: case (a) (left upper panel), case (b) (right upper panel), case (c) (left lower panel) and case (d) (right lower panel) in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

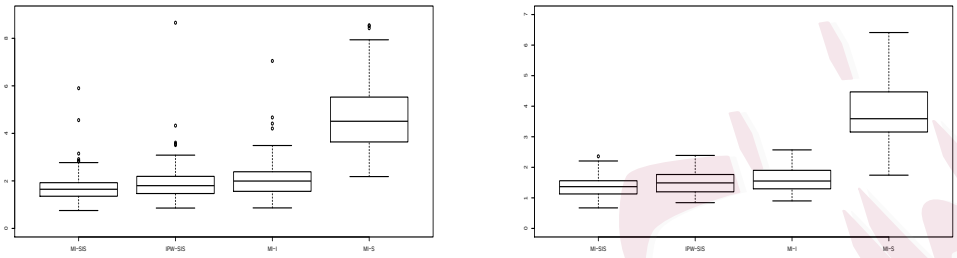

Figure 3: WMSE values of the proposed model averaging method for four different screening methods: case (e) (left panel) and case (f) (right panel) in Experiment 2

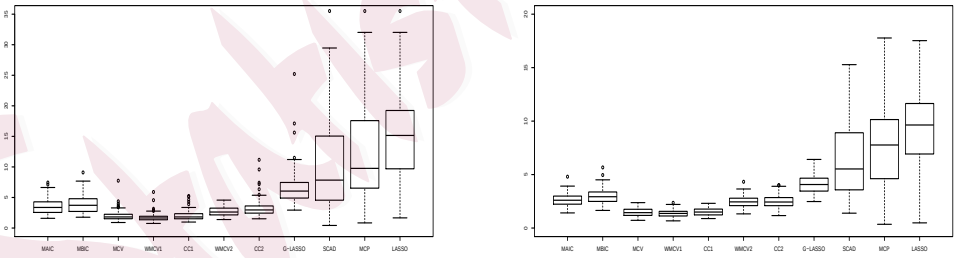

Figure 4: WMSE values of eleven model-averaging methods for two settings of ρ, the distribution of ε_i and propensity score function $\pi(\boldsymbol{U}_i;\boldsymbol{\gamma})$: case (e) (left panel) and case (f) (right panel) in Experiment 2

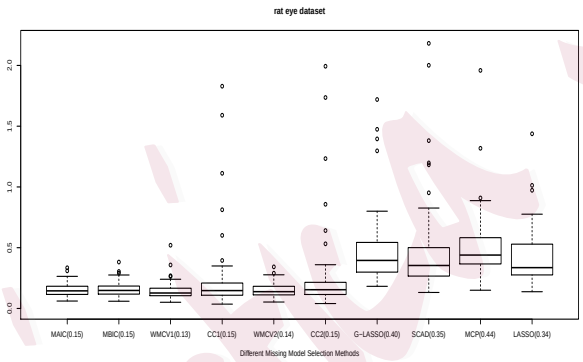

Figure 5: WMSPE values of ten model-averaging methods in the rat eye dataset. The number in brackets is the median of distribution for the WMSPE.