Journal of Advanced Statistics, Vol. 2, No. 1, March 2017
https://dx.doi.org/10.22606/jas.2017.21001

Variable Selection for Additive Models with Missing Response at Random

Jian Wu[1], Junhua Zhang[2] and Gaorong Li[3*]

[1]College of Science, Northeastern University, Shenyang 110189, China
[2]College of Mechanical Engineering, Beijing Information Science and Technology University, Beijing 100192, China
[3]Beijing Institute for Scientific and Engineering Computing, Beijing University of Technology, Beijing 100124, China
Email: ligaorong@gmail.com

Abstract This paper studies the problems of variable selection and estimation in the additive models with missing response at random. Based on the centered spline basis function approximation, we propose two new imputed estimating equation methods to implement the variable selection for the additive models with missing response at random by using the smooth-threshold estimating equation. Two new imputed methods can select the significant variables and estimate the unknown functions simultaneously. The proposed methods not only avoid the problem of solving a convex optimization, but also reduce the burden of computation. With the proper choices of the regularization parameter, we show that the resulting estimators enjoy the oracle property. The data driven method is used to choose the tuning parameter. A numerical study is analyzed to confirm the performance of the proposed methods.

Keywords: Additive model, smooth-threshold estimating equations, variable selection, missing data, oracle property.

1 Introduction

In this paper, we consider the following additive model:

$$Y_i = \mu + \sum_{l=1}^{p} m_l(X_{il}) + \varepsilon_i, \qquad i = 1, \cdots, n, \tag{1}$$

where Y is the response variable, $\boldsymbol{X} = (X_1, \cdots, X_p)^\tau$ is a p dimensional vector of the predictor variables, μ is an unknown constant, and $m(\boldsymbol{X}) = \sum_{l=1}^{p} m_l(X_l)$ with the unknown nonparametric functions $m_l(X_l)$, where X_l is the lth component of $\boldsymbol{X}, l = 1, \cdots, p$. ε is model error with mean zero and variance σ^2. Without loss of generality, we assume that the estimation of $m(\boldsymbol{X}) = \sum_{l=1}^{p} m_l(X_l)$ is conducted on a compact support, and let the compact set be $\chi = [0, 1]^p$. We further also impose the condition $\mathbb{E}[m_l(X_l)] = 0$ which is required for identifiability of model (1). Let $r = 1$ if Y is observed and $r = 0$ if Y is missing. The missing mechanism associated with the missingness of Y is characterized by the conditional distribution of r given $\boldsymbol{X}$, which is assumed to satisfy

$$\mathbb{P}(r = 1|\boldsymbol{X}, Y) = \mathbb{P}(r = 1|\boldsymbol{X}) = \pi(\boldsymbol{X}, \boldsymbol{\eta}) = \frac{\exp\{(1, \boldsymbol{X}^\tau)\boldsymbol{\eta}\}}{1 + \exp\{(1, \boldsymbol{X}^\tau)\boldsymbol{\eta}\}}, \tag{2}$$

where π is a known function. In this paper, we assume that π is logistic function, $\boldsymbol{\eta}$ can be consistently estimated by the complete data. Recently, some researches had studied the problems on the statistical inference for model (1). For example, You, Chen and Zhou [1] considered the estimation and goodness-of-fit test of the additive model with measurement error in covariates. Wang and Yang [2] adopted the spline backfitting kernel method to estimate the unknown component functions, and established the asymptotic properties. Opersomer and Ruppert [3] proposed the local linear smoothing method to make inference for the additive model. Xue [4] proposed a regularized estimation procedure for variable selection that combines the basis function approximation with the SCAD penalty. Wu and Xue [5] studied the model detection for the additive models with longitudinal data. In addition, some variable selection methods

were proposed to study model (1) under the complete data, such as COSSO (Lin and Zhang [6]), SpAM (Ravikumar et al. [7]), gamsel (Chouldechova and Hastie [8]), and among others. It is very difficult for the variable selection with missing data using the penalty strategy.

In this paper, we study the problems of variable selection and estimation in the additive model (1) with missing response at random. Based on the centered spline basis function approximation, we propose two new imputed estimating equation methods to estimate the component functions using the smooth-threshold estimating equation proposed in Ueki [9]. The variable selection method based on the smooth-threshold estimating equation can avoid the convex optimization of the penalized variable selection procedure, and has been extended to some models, such as Li et al. [10], Zhao and Li [11], Lv et al. [12] and Geronimia and Saportab [13]. In the present paper, based on the marginal imputed estimating equation and the maximum correlation imputed estimating equation, we define two smooth-threshold estimating equations to study model (1) with missing response at random. Two methods can automatically select the important variables in the model, while simultaneously estimate the nonzero functions. The proposed methods can yield the sparse solutions, avoid the problems of solving convex optimization and reduce the burden of computation. Some asymptotic properties are established under some regularity conditions, and the simulation studies are carried out to illustrate the efficacy of the proposed methods.

The rest of the paper is organized as follows. In Section 2, we introduce our methods, study the asymptotic properties under some regularity conditions and propose the data driven method to select the tuning parameter. The simulation studies are carried out to illustrate the proposed methods in Section 3. The technical details are presented in the Appendix.

2 Methodology and Asymptotic Properties

2.1 Estimation

We approximate the nonparametric functions $m_l(\cdot)$'s by the centered B splines basis. Let $\mathbf{B}(\boldsymbol{x}) = \{B_{s,l}(x_l) : 1 \le l \le p, 1 \le s \le J\}$ is a basis system, where $\boldsymbol{x} = (x_l)_{l=1}^p$, $J = K + q + 1$, K is the number of interior knots and q is the order of B spline basis. The equally-spaced knots are used in this paper for the simplicity of proofs. Other regular knot sequences can also be used with similar asymptotic results. Suppose that $m_l(\cdot)$ can be approximated by the spline basis functions, that is

$$m_l(x) \approx m_l^{\text{sp}}(x) = \sum_{s=1}^{J} \beta_{sl} B_{s,l}(x), \qquad l = 1, \cdots, p. \tag{3}$$

Substituting (3) into model (1), we can get

$$Y_i \approx \mathbf{B}_i^\tau \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \cdots, n, \tag{4}$$

where $\boldsymbol{\beta} = (\mu, \beta_{11}, \cdots, \beta_{J1}, \cdots, \beta_{1p}, \cdots, \beta_{Jp})^\tau$ is the collection of the coefficients in (3), and $\mathbf{B}_i = (1, B_{1,1}(X_{i1}), \cdots, B_{1,J}(X_{i1}), \cdots, B_{p,1}(X_{ip}), \cdots, B_{p,J}(X_{ip}))^\tau$. Thus, model (4) is a standard linear regression model, and we note that each $m_l(x)$ is characterized by $(\beta_{1l} \cdots, \beta_{Jl})^\tau$. Therefore, we adopt the imputed estimating equation as follows

$$\sum_{i=1}^{n} \tilde{\phi}_i^{(I)}(\boldsymbol{\beta}, \boldsymbol{X}_i) = \sum_{i=1}^{n} \Big\{ r_i \phi_i(\boldsymbol{\beta}) + \{1 - r_i\} m_\phi(\boldsymbol{\beta}, \boldsymbol{X}_i) \Big\} = 0, \tag{5}$$

where $\phi_i(\boldsymbol{\beta}) = \mathbf{B}_i(Y_i - \mathbf{B}_i^\tau \boldsymbol{\beta})$, and $m_\phi(\boldsymbol{\beta}, \boldsymbol{x}) = \mathbb{E}\{\phi_i(\boldsymbol{\beta}) | \boldsymbol{X}_i = \boldsymbol{x}\}$ is a function vector of p-dimensional variable $\boldsymbol{x}$. Then we will suffer from the curse of dimensionality to estimate the function $m_\phi(\boldsymbol{\beta}, \boldsymbol{x})$. To overcome this problem, a new marginal imputation (MI) is used to solve the following estimating equation

$$\sum_{i=1}^{n} \phi_i^{(\text{MI})}(\boldsymbol{\beta}, \boldsymbol{X}_i) = \sum_{i=1}^{n} \Big\{ r_i \phi_i(\boldsymbol{\beta}) + \{1 - r_i\} \hat{m}_\phi^{(\text{MI})}(\boldsymbol{\beta}, \boldsymbol{X}_i) \Big\} = 0, \tag{6}$$

where $m_{\phi}^{(\mathrm{MI})}(\boldsymbol{\beta},\boldsymbol{x}) = \frac{1}{p}\sum_{l=1}^{p}\mathbb{E}\{\phi_i(\boldsymbol{\beta})|X_{il}=x_l\}$ can be estimated by

$$\hat{m}_{\phi}^{(\mathrm{MI})}(\boldsymbol{\beta},\boldsymbol{x}) = \frac{1}{p}\sum_{l=1}^{p}\sum_{i=1}^{n}W_{li}(x_l)\phi_i(\boldsymbol{\beta}),$$

where $W_{li}(x) = r_i K_{h_l}(X_{il}-x)/\sum_{j=1}^{n} r_j K_{h_l}(X_{jl}-x)$, $K_{h_l}(\cdot) = h_l^{-1}K(\cdot/h_l)$, $K(\cdot)$ is a kernel function, and h_l is a bandwidth, $l = 1, 2, \cdots, p$. By solving the marginal imputed estimating equation (6), we can obtain the estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{MI})} = \left(\hat{\mu}^{(\mathrm{MI})}, \hat{\beta}_{11}^{(\mathrm{MI})}, \cdots, \hat{\beta}_{J1}^{(\mathrm{MI})}, \cdots, \hat{\beta}_{1p}^{(\mathrm{MI})}, \cdots, \hat{\beta}_{Jp}^{(\mathrm{MI})}\right)^{\tau}$ of $\boldsymbol{\beta}$. Thus, we can obtain the estimator of function $m_l(\cdot)$ as follows

$$\hat{m}_l^{(\mathrm{MI})}(x) = \sum_{s=1}^{J}\hat{\beta}_{sl}^{(\mathrm{MI})}B_{s,l}(x), \quad l = 1, \cdots, p. \tag{7}$$

Since the above proposed procedure needs to select $h_1, \cdots, h_p$ with p bandwidths and estimate $m_1(x_1), \cdots, m_p(x_p)$ with p unknown functions, this will lead to the heavy computational burden to implement this procedure. So we propose a simple method called as the maximum correlation imputation method (MRI). We first calculate the correlation coefficients $\rho_l = \mathrm{cor}(Y, X_l)$, $l = 1, \cdots, p$, and choose u by maximizing the correlation coefficients as follows

$$u = \arg\max_{1\le l\le p}\rho_l.$$

Thus, we can construct the following maximum correlation imputed estimating equation:

$$\sum_{i=1}^{n}\phi_i^{(\mathrm{MRI})}(\boldsymbol{\beta}, X_{iu}) = \sum_{i=1}^{n}\left\{r_i\phi_i(\boldsymbol{\beta}) + \{1-r_i\}\hat{m}_{\phi}^{(\mathrm{MRI})}(\boldsymbol{\beta}, X_{iu})\right\} = 0, \tag{8}$$

where $\hat{m}_{\phi}^{(\mathrm{MRI})}(\boldsymbol{\beta}, X_{iu})$ is the consistent estimator of $m_{\phi}^{(\mathrm{MRI})}(\boldsymbol{\beta}, x) = \mathbb{E}\{\phi_i(\boldsymbol{\beta})|X_{iu}=x\}$, which can be estimated by the nonparametric smoothing methods. The estimator is defined as

$$\hat{m}_{\phi}^{(\mathrm{MRI})}(\boldsymbol{\beta},\boldsymbol{x}) = \sum_{i=1}^{n}W_{ui}(x)\phi_i(\boldsymbol{\beta}),$$

where $W_{ui}(x) = r_i K_{h_u}(X_{iu}-x)/\sum_{j=1}^{n} r_j K_{h_u}(X_{ju}-x)$, $K_{h_u}(\cdot) = h_u^{-1}K(\cdot/h_u)$, $K(\cdot)$ is a kernel function, and h_u is a bandwidth. By solving the estimating equation (8), we can obtain the estimator $\hat{\boldsymbol{\beta}}^{(\mathrm{MRI})} = \left(\hat{\mu}^{(\mathrm{MRI})}, \hat{\beta}_{11}^{(\mathrm{MRI})}, \cdots, \hat{\beta}_{J1}^{(\mathrm{MRI})}, \cdots, \hat{\beta}_{1p}^{(\mathrm{MRI})}, \cdots, \hat{\beta}_{Jp}^{(\mathrm{MRI})}\right)^{\tau}$ of $\boldsymbol{\beta}$. Similar to (7), the estimator of function $m_l(\cdot)$ is defined by

$$\hat{m}_l^{(\mathrm{MRI})}(x) = \sum_{s=1}^{J}\hat{\beta}_{sl}^{(\mathrm{MRI})}B_{s,l}(x), \quad l = 1, \cdots, p. \tag{9}$$

It is easy to see that this procedure will reduce the computing burden once we can obtain u by maximizing the correlation coefficients. Then we only need to choose a bandwidth h_u to estimate the nonparametric function $m_{\phi}^{(\mathrm{MRI})}(\boldsymbol{\beta}, x)$.

2.2 Variable Selection

In this subsection, we consider the problem of variable selection to select the important functional predictors for the additive model (1) with missing response at random. Motivated by the idea in Ueki [9] and Li et al. [10], we use the smooth-threshold estimating equations to perform the variable selection. Based on the marginal imputed estimating equation (6) and the maximum correlation imputed estimating equation (8), we define two smooth-threshold estimating equations as follows, respectively

$$(I_{pJ} - \triangle)\sum_{i=1}^{n}\phi_i^{(\mathrm{MI})}(\boldsymbol{\beta}, \boldsymbol{X}_i) + \triangle\boldsymbol{\beta} = 0 \tag{10}$$

and

$$(I_{pJ} - \triangle)\sum_{i=1}^{n}\phi_i^{(\mathrm{MRI})}(\boldsymbol{\beta}, \boldsymbol{X}_i) + \triangle\boldsymbol{\beta} = 0, \tag{11}$$

where I_{pJ} is the pJ-dimensional identity matrix and $\triangle = \mathrm{diag}\{\delta_1, \cdots, \delta_1, \cdots, \delta_p, \cdots, \delta_p\}$ is the $pJ \times pJ$ diagonal matrix. Note that $\|\beta_k\| = 0$ if $\delta_k = 1$, hence (10) and (11) can yield the sparse solution. In order to finish the variable selection, we first need to choose the threshold parameters $\delta_k, k = 1, \cdots, p$ using some data-driven criteria. We choose the threshold parameters as $\hat{\delta}_k = \min\{1, \lambda/\|\hat{\beta}_k^{(0)}\|\}$, where $\hat{\beta}_k^{(0)}$ is the initial estimator of β_k, and λ is the tuning parameter. $\hat{\beta}_k^{(0)}$ can be obtained by solving the marginal imputed estimating equation (6) and the maximum correlation imputed estimating equation (8), respectively. In Subsection 2.4, we will provide the data driven method to select the tuning parameter λ. Thus, we can obtain the following two estimating equations by replacing $\triangle$ by its estimator $\hat{\triangle} = \mathrm{diag}\{\hat{\delta}_1, \cdots, \hat{\delta}_1, \cdots, \hat{\delta}_p, \cdots, \hat{\delta}_p\}$ in (10) and (11), respectively

$$(I_{pJ} - \hat{\triangle})\sum_{i=1}^{n}\phi_i^{(\mathrm{MI})}(\boldsymbol{\beta}, \boldsymbol{X}_i) + \hat{\triangle}\boldsymbol{\beta} = 0 \tag{12}$$

and

$$(I_{pJ} - \hat{\triangle})\sum_{i=1}^{n}\phi_i^{(\mathrm{MRI})}(\boldsymbol{\beta}, \boldsymbol{X}_i) + \hat{\triangle}\boldsymbol{\beta} = 0. \tag{13}$$

By solving the estimating equations (12) and (13) respectively, we can obtain the estimator of $\boldsymbol{\beta}$, and we define the estimator of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}^{(\mathrm{MIVS})}$ and $\hat{\boldsymbol{\beta}}^{(\mathrm{MRIVS})}$, respectively. Thus, we can obtain the two estimators of function $m_l(\cdot)$ as follows, respectively

$$\hat{m}_l^{(\mathrm{MIVS})}(x) = \sum_{s=1}^{J}\hat{\beta}_{sl}^{(\mathrm{MIVS})}B_{s,l}(x), \quad l = 1, \cdots, p \tag{14}$$

and

$$\hat{m}_l^{(\mathrm{MRIVS})}(x) = \sum_{s=1}^{J}\hat{\beta}_{sl}^{(\mathrm{MRIVS})}B_{s,l}(x), \quad l = 1, \cdots, p. \tag{15}$$

2.3 Asymptotic Properties

In this Subsection, we will study the asymptotic properties of the estimators defined by (7), (9), (14) and (15). Let $m_l^0(\cdot)$ be the true function of $m_l(\cdot)$, and corresponding true value of $\boldsymbol{\beta}$ is denoted by $\boldsymbol{\beta}^0$. Without loss of generality, we assume that $m_l^0(\cdot) \equiv 0, l = d+1, \cdots, p$ and $m_l^0(\cdot), l = 1, \cdots, d$ are all nonzero components of $m^0(\cdot)$. We will provide the main theorems as follows.

Theorem 1. Suppose that the regularity conditions (A1)–(A5) given in the Appendix hold and the number of knots $K = O_p(n^{1/(2r+1)})$. As $n \to \infty$, then the estimators defined by (7) and (9) satisfy the following results

$$\|\hat{m}_l^{(\mathrm{MI})}(\cdot) - m_l^0(\cdot)\| = O_p(n^{-r/(2r+1)}), \quad l = 1, \cdots, p$$

and

$$\|\hat{m}_l^{(\mathrm{MRI})}(\cdot) - m_l^0(\cdot)\| = O_p(n^{-r/(2r+1)}), \quad l = 1, \cdots, p,$$

where r is defined in Condition (A4) in the Appendix.

Theorem 1 implies that the estimators defined by (7) and (9) of the unknown functions achieve the optimal convergence rate. The following theorem gives the convergence rate of the estimators defined by (14) and (15) based on two variable selection procedures.

Theorem 2. Suppose that the regularity conditions (A1)–(A5) given in the Appendix hold and the number of knots $K = O_p(n^{1/(2r+1)})$. As $n \to \infty$, then the estimators defined by (14) and (15) satisfy the following results

$$\|\hat{m}_l^{(\mathrm{MIVS})}(\cdot) - m_l^0(\cdot)\| = O_p(n^{-r/(2r+1)}), \quad l = 1, \cdots, d$$

and

$$\|\hat{m}_l^{(\mathrm{MRIVS})}(\cdot) - m_l^0(\cdot)\| = O_p(n^{-r/(2r+1)}), \quad l = 1, \cdots, d.$$

Theorem 2 implies that the estimators defined by (14) and (15) also achieve the optimal convergence rate if the subset of true zero coefficients are already known by choosing the proper tuning parameters. Under some regularity conditions, we show that such consistent estimators must possess the sparsity property, which is stated as follows.

Theorem 3. Suppose that the regularity conditions (A1)–(A5) given in the Appendix hold and the number of knots $K = O_p(n^{1/(2r+1)})$. If $\lambda \to 0$ and $n^{r/(2r+1)}\lambda \to \infty$ as $n \to \infty$, with probability tending to 1, then the estimators defined by (14) and (15) must satisfy $\hat{m}_l^{(\mathrm{MIVS})}(\cdot) \equiv 0$ and $\hat{m}_l^{(\mathrm{MRIVS})}(\cdot) \equiv 0$, $l = d+1, \cdots, p$.

Remark 1. Invoked by Theorem 2 and Theorem 3, it is clear that by choosing the proper tuning parameter, our variable selection procedures are consistent, and the estimators of coefficient functions achieve the optimal convergence rate if the subset of true zero functions is already known.

2.4 Tuning Parameters Selection

To implement the variable selection, we need to choose the number of interior knots K, the tuning parameter λ and bandwidth h. We can use the classical methods such as CV, GCV, BIC-type criterion to select these parameters. To simplify the computation, we choose the number of interior knots K using the procedure in Zhao and Xue [14]. We first can choose K by minimizing the following cross-validation score.

$$\mathrm{CV}(K) = \sum_{i=1}^{n} r_i \{Y_i - \mathbf{B}_i^\tau \hat{\boldsymbol{\beta}}_{[i]}\}^2, \tag{16}$$

where $\hat{\boldsymbol{\beta}}_{[i]}$ is the solution of $\sum_{i=1}^n r_i \phi_i(\boldsymbol{\beta}) = 0$ after deleting the ith subject. When K is determined, we can select the turning parameter λ by minimizing the following BIC criterion.

$$\mathrm{BIC}(\lambda) = \Big\| \sum_{i=1}^{n} r_i \mathbf{B}_i (Y_i - \mathbf{B}_i^\tau \hat{\boldsymbol{\beta}}^\lambda) \Big\|^2 + \mathrm{DF}(\lambda) \log n, \tag{17}$$

where $\hat{\boldsymbol{\beta}}^\lambda$ denotes the estimators defined by (14) and (15) given λ. As in Zhou, Alan and Wang [15], we take $h = Cn^{-1/3}$, and $h_l = \hat{\sigma}_{X_l} h$, where σ_{X_l} is the estimated standard deviation of X_l in the sample.

3 Numerical Results

In this section, we conduct some simulations to evaluate the finite sample performances of the proposed methods. The performances of estimators for $m(\cdot)$ will be assessed by using the square root of average square errors (RASE) defined by

$$\mathrm{RASE} = \left\{ \frac{1}{N} \sum_{s=1}^{N} \sum_{l=1}^{p} [\hat{m}_l(x_s) - m_l^0(x_s)]^2 \right\}^{1/2},$$

where x_s, $s = 1, \cdots, N$, are the grid points at which the function $\hat{m}(u)$ is evaluated. In our simulation, $N = 200$ is used. We simulate data from model (1), and we take $p = 10$, $m_1(x) = 4x - 1$, $m_2(x) = \cos(2\pi x)$ and $m_3(x) = \sin(2\pi x)$. The remaining functions, corresponding to the irrelevant variables, are given as zeros. We take $\boldsymbol{Z} = (Z_1, \cdots, Z_{10})^\tau \sim N(0, \Sigma_z)$, and the covariates $X_k = \Phi(Z_k)$, $k = 1, \cdots, 10$, where $\Phi(\cdot)$ denotes the distribution function of standard normal distribution. Y is generated according to model

(1) with $\varepsilon \sim N(0, 0.5)$. We also generated r_i, the response indicator variable, from Bernoulli distribution with probability $\pi(\cdot)$. We consider two response models for $\pi(\cdot)$:

Missing Mechenism 1:

$$\pi(\boldsymbol{X}, \boldsymbol{\eta}) = \mathbb{P}(r = 1|\boldsymbol{X}) = \frac{\exp\{(1, \boldsymbol{X}^\tau)\boldsymbol{\eta}\}}{1 + \exp\{(1, \boldsymbol{X}^\tau)\boldsymbol{\eta}\}},$$

where $\boldsymbol{\eta} = (1, 2, 3, 0, 0, 0, 0, 0, 0, 0, 0.5)^\tau$, and the response rate is about 60%.

Missing Mechenism 2:

$$\pi(\boldsymbol{X}, \boldsymbol{\eta}) = \mathbb{P}(r = 1|\boldsymbol{X}) = \frac{\exp\{(1, \boldsymbol{X}^\tau)\boldsymbol{\eta}\}}{1 + \exp\{(1, \boldsymbol{X}^\tau)\boldsymbol{\eta}\}},$$

where $\boldsymbol{\eta} = (0.3, 0.6, 1, 1, 0, 0, 0, 0, 0, 0, 0)^\tau$, and the response rate is about 80%.

In this simulation, we use the cubic B-splines and consider the sample size $n = 150, 300$ and 500, respectively. The simulation results are reported in Table 1. In Table 1, we first compare the performances of two estimators defined in (7) and (9) with the naive estimator, which is estimated by using the complete samples. We also compute the oracle estimator as a benchmark, and the oracle estimator is only available in simulation studies, where the true data information is known.

Table 1. The performances of MI and MRI, NAIVE (naive estimator) and ORACLE (oracle estimator)

		Modle 1					Modle 2				
n	Method	μ	$m_1(x)$	$m_2(x)$	$m_3(x)$	$m_4(x)$	μ	$m_1(x)$	$m_2(x)$	$m_3(x)$	$m_4(x)$
150	NAIVE	17.07	32.65	31.99	32.09	31.12	13.06	25.65	24.87	24.03	24.32
	MI	15.31	27.32	27.01	27.12	26.83	12.36	22.53	21.25	21.32	21.58
	MRI	15.16	26.23	25.28	25.39	24.63	13.16	22.26	21.36	21.25	21.63
	ORACLE	14.51	24.50	23.51	23.42	22.47	11.35	18.86	18.38	18.86	18.26
300	NAIVE	11.24	19.48	19.44	18.88	18.31	10.25	17.47	18.27	17.85	17.31
	MI	10.60	17.56	17.49	17.32	17.66	9.66	15.17	15.64	15.06	15.01
	MRI	10.55	17.00	16.98	16.58	15.86	9.54	15.17	15.54	15.03	14.94
	ORACLE	10.20	15.72	15.53	15.31	14.42	8.84	13.63	13.86	13.35	13.31
500	NAIVE	8.44	14.19	13.83	13.62	13.39	8.44	13.38	12.75	12.66	12.34
	MI	8.07	12.98	12.84	12.64	12.73	8.07	11.31	11.26	11.23	11.03
	MRI	8.06	12.83	12.41	12.17	11.96	6.54	11.28	11.21	11.20	11.01
	ORACLE	7.87	11.91	11.41	11.18	10.91	5.32	9.63	9.45	9.36	9.29

The MSE of parameter μ and RASE of component functions are reported in Table 1, the results demonstrates our proposed two methods have the same effectiveness as the empirical method.

The average number of the estimated zero component function for component functions, with 1000 simulation runs, is listed in Table 2, in which the column labeled “C" gives the average number of the true zero component function correctly set to zero, and the column labeled “I" gives the average number of the true nonzero function incorrectly set to zero. Furthermore, Table 2 also presents the median of RASE of semiparametric $\mu + \sum_{i=1}^{n} m_i(x_i)$ over the 1000 simulations. From Table 2, we can find the following results.

(i) For the given two missing mechanism models, the performances of the estimators defined by (14) and (15) become closer to the oracle estimator as the sample size n increases. This indicates that the proposed methods can eliminate the unimportant variables effectively and estimate the nonzero functions simultaneously.

(ii) For the given sample size n, the performances of the estimators defined by (14) and (15) are similar, but the latter is easy to implement.

Table 2. The performances of MIVS and MRIVS, NAIVE (naive estimator) and ORACLE (oracle estimator)

		NAIVE			MIVS			MRIVS			ORACLE		
	n	C	I	RASE	C	I	RASE	C	I	RASE	C	I	RASE
M1	150	5.463	0	0.251	6.609	0	0.179	6.479	0	0.186	6.785	0	0.172
	300	6.828	0	0.145	6.984	0	0.126	6.952	0	0.129	6.996	0	0.124
	500	7	0	0.087	7	0	0.085	7	0	0.085	7	0	0.084
M2	150	6.479	0	0.186	6.867	0	0.153	6.860	0	0.156	6.982	0	0.136
	300	6.974	0	0.126	6.997	0	0.103	6.992	0	0.110	6.998	0	0.096
	500	7	0	0.089	7	0	0.080	7	0	0.080	7	0	0.079

4 Concluding Remark

In summary, we extend the variable selection method called as the smooth-threshold estimating equation in Ueki [9] and Li et al. [10] to the additive models with missing response at random, and we propose two new variable selection procedures for model (1) with missing response at random. The proposed methods can select the important variables and estimate the unknown component functions simultaneously. The main advantage of the proposed methods is to avoid the problem of solving a convex optimization. The simulation studies show that the proposed variable selection procedures can effectively select the important variables in the model with missing data. In addition, there is a useful form of expansion of the additive model for additive partially linear model and the generalized additive model. Therefore, further study of the problem is how to use the smooth-threshold estimating equations method to study the variable selection in some semiparametric models with missing data.

Acknowledgments. Jian Wu's research was supported by the Doctoral Scientific Research Foundation of Liaoning Province (No. 201601009) and the Grant of Central Universities of China (No. N160504004). Junhua Zhang's research was supported by the National Natural Science Foundation of China (No. 11472057). Gaorong Li's research was supported by the National Natural Science Foundation of China (No. 11471029).

Appendix

For convenience and simplicity, let C denote a positive constant that may be different at each appearance throughout this paper. Before we prove our main theorems, we list some regularity conditions that are used in this paper.

(A1) The covariate $\boldsymbol{X} = (X_1, \cdots, X_p)^\tau$ has a compact support set, and let the compact set be $\chi = [0,1]^p$.

(A2) The density function of X_l, denoted by $f_l(x)$, is absolutely continuous. There exist constants C_1 and C_2 such that $0 < C_1 \leq \min_{x\in[0,1]} f_l(x) \leq \max_{x\in[0,1]} f_l(x) \leq C_2 < \infty$ for all $l = 1, \ldots, p$.

(A3) The bandwidth h satisfies $h \to 0, nh^2 \to \infty$, and $nh^4 \to 0$ as $n \to \infty$.

(A4) For each $l = 1, \ldots, p$, $m_l(\cdot)$ has r continuous derivatives for some $r \geq 2$.

(A5) The tuning parameter λ satisfies that $n^{r/(2r+1)}\lambda \to \infty$ and $n^{1/2}\lambda \to 0$ as $n \to \infty$.

These conditions are commonly used in the additive model and variable selection references, such as Conditions (A1)–(A4) are similar to those used in Xue [16], and Condition (A5) is a necessary condition to prove the consistency of variable selection.

Proof of Theorem 1. Since the proof of the estimator $\hat{m}_l^{(\mathrm{MRI})}(\cdot)$ defined in (9) is similar to the proof of the estimator $\hat{m}_l^{(\mathrm{MI})}(\cdot)$, we only prove the result of the estimator $\hat{m}_l^{(\mathrm{MI})}(\cdot)$ defined in (7). By (6),

we have

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{X}) &= \frac{1}{n}\sum_{i=1}^{n}\phi_i^{(\mathrm{MI})}(\boldsymbol{\beta}, \boldsymbol{X}_i) = \frac{1}{n}\sum_{i=1}^{n}\Big\{r_i\phi_i(\boldsymbol{\beta}) + \{1-r_i\}\hat{m}_\phi^{(\mathrm{MI})}(\boldsymbol{\beta}, \boldsymbol{X}_i)\Big\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big\{r_i\phi_i(\boldsymbol{\beta}) + \{1-r_i\}m_\phi(\boldsymbol{\beta}, \boldsymbol{X}_i) + \{1-r_i\}\{\hat{m}_\phi^{(\mathrm{MI})}(\boldsymbol{\beta}, \boldsymbol{X}_i) - m_\phi(\boldsymbol{\beta}, \boldsymbol{X}_i)\}\Big\} \\ &=: I_1 + I_2 + I_3. \end{aligned}$$

For I_1, we have

$$\begin{aligned} I_1 &= \frac{1}{n}\sum_{i=1}^{n}\{r_i\phi_i(\boldsymbol{\beta})\} = \frac{1}{n}\sum_{i=1}^{n}\{r_i\mathbf{B}_i(Y_i - \mathbf{B}_i^\tau\boldsymbol{\beta})\} \\ &= \frac{1}{n}\sum_{i=1}^{n} r_i\mathbf{B}_i\left\{\left[Y_i - \Big(\mu + \sum_{l=1}^{p} m_l(X_{il})\Big)\right] + \left(\mu + \sum_{l=1}^{p} m_l(X_{il}) - \mathbf{B}_i^\tau\boldsymbol{\beta}^0\right) + r_i\mathbf{B}_i\mathbf{B}_i^\tau(\boldsymbol{\beta}^0 - \boldsymbol{\beta})\right\} \\ &=: I_{11} + I_{12} + I_{13}. \end{aligned}$$

By some elementary calculations, we have

$$I_{11} = \frac{1}{n}\sum_{i=1}^{n} r_i\mathbf{B}_i\varepsilon_i \to O_p(n^{-1/2}).$$

Note that $\mathbf{B}_i^\tau\boldsymbol{\beta}^0$ represents that the function $\mu + \sum_{l=1}^{p} m_l(x_l)$ projects into the additive spline space. By the spline result in de Boor [17], we have

$$I_{12} = \max_{1\le i\le n}\Big\|\mu + \sum_{l=1}^{p} m_l(X_{il}) - \mathbf{B}_i^\tau\boldsymbol{\beta}^0\Big\| = O_p(K^{-r})$$

and

$$I_{13} = O_p(1)(\boldsymbol{\beta} - \boldsymbol{\beta}^0).$$

Then, we have $I_1 = O_p(K^{-r} + n^{-1/2}) + O_p(1)(\boldsymbol{\beta} - \boldsymbol{\beta}^0)$. Using the same argument of I_1, we can obtain that $I_2 = O_p(K^{-r} + n^{-1/2}) + O_p(1)(\boldsymbol{\beta} - \boldsymbol{\beta}^0)$. By the results of the kernel smoothing method, we have $I_3 = O_p((nh)^{-1/2})$. Note that $\hat{\boldsymbol{\beta}}^{(\mathrm{MI})}$ is the solution of the equation $L(\boldsymbol{\beta}, \boldsymbol{X}) = 0$, and $K = O_p(n^{1/(2r+1)})$ and $nh^4 \to 0$. Thus it is easy to see that

$$\|\hat{m}_l^{(\mathrm{MI})}(\cdot) - m_l^0(\cdot)\| = O_p(n^{-r/(2r+1)}), \qquad l = 1, \cdots, p.$$

By these results, we finish the proof of Theorem 1.

Proof of Theorem 2. By Theorem 1, we can get that $\tilde{\beta}_j^0 = O_p(n^{-r/(2r+1)})$, $j = 1, \cdots, p$. For any $\epsilon > 0$, we have

$$\mathbb{P}(\hat{\delta}_j > n^{-1/2}\epsilon) = \mathbb{P}\left(\frac{\lambda}{\|\tilde{\beta}_k^0\|} > n^{-1/2}\epsilon\right) = \mathbb{P}\Big(\|\tilde{\beta}_k^0\| < n^{1/2}\epsilon^{-1}\lambda\Big) \to 0, \quad j = 1, \cdots, p.$$

By the condition $n^{1/2}\lambda \to 0$, and $\|\tilde{\beta}_k^0\|$ is bound away from zero. Thus, we have $\hat{\delta}_j = o_p(n^{-1/2})$, for $j = 1, \cdots, p$. A slight symbol abuse, we have $\hat{\triangle} = o_p(n^{-1/2})$ and

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (I_{pJ} - \hat{\triangle})\sum_{i=1}^{n}\Big\{r_i\mathbf{B}_i(Y_i - \mathbf{B}_i^\tau\boldsymbol{\beta}) + (1-r_i)\hat{m}_\phi^{(\mathrm{MI})}(\boldsymbol{X}_i)\Big\} + \hat{\triangle}\boldsymbol{\beta} \\ &= (I_{pJ} - \hat{\triangle})\sum_{i=1}^{n} r_i\mathbf{B}_i(Y_i - \mathbf{B}_i^\tau\boldsymbol{\beta}) + (I_{pJ} - \hat{\triangle})\sum_{i=1}^{n}(1-r_i)m_\phi(\boldsymbol{X}_i) \\ &\quad + (I_{pJ} - \hat{\triangle})\sum_{i=1}^{n}(1-r_i)(\hat{m}_\phi^{(\mathrm{MI})}(\boldsymbol{X}_i) - m_\phi(\boldsymbol{X}_i)) + \hat{\triangle}\boldsymbol{\beta}. \end{aligned}$$

Similar to the proof of Theorem 1, and note that $\hat{\delta}_j = o_p(n^{-1/2})$, the proof of the convergence rate of $\hat{m}_l^{(\mathrm{MIVS})}(\cdot)$ is similar to that of the convergence rate of $\hat{m}_l^{(\mathrm{MI})}(\cdot)$, $l = 1, \cdots, p$. Similarly, we also can prove the convergence rate of $\hat{m}_l^{(\mathrm{MRIVS})}(\cdot)$. Therefore, we finish the proof of Theorem 2.

Proof of Theorem 3. By condition (A5) $n^{r/(2r+1)}\lambda \to \infty$ as $n \to \infty$, we have

$$\sum_{j=d+1}^{p} \mathbb{P}(\lambda/\|\hat{\beta}_j^0\| < 1) \leq \lambda^{-1} O(n^{-r/(2r+1)}) \to 0.$$

By the above result, it is easy to see that

$$\mathbb{P}(\hat{\delta}_j = 1, \ for\ all\ j = d+1, \cdots, p) \to 1,$$

which implies that $\hat{m}_j^{(\mathrm{MIVS})}(\cdot) \equiv 0$ and $\hat{m}_j^{(\mathrm{MRIVS})}(\cdot) \equiv 0, j = d+1, \cdots, p$. Hence we complete the proof of Theorem 3.

References

1. J. You, G. Chen, and Y. Zhou, "Block empirical likelihood for longitudinal partially linear regression models," *Canadian Journal of Statistics*, vol. 34, pp. 79–96, 2006.
2. L. Wang and L. Yang, "Spline backfitted kernel smoothing of nonlinear additive autoregression model," *The Annals of Statistics*, vol. 35, pp. 2474–2503, 2007.
3. J. D. Opsomer and D. Ruppert, "Fitting a bivariate additive model by local polynomial regression," *The Annals of Statistics*, vol. 25, no. 1, pp. 186–211, 1997.
4. L. Xue, "Consistent variable selection in additive models," *Statistica Sinica*, vol. 19, pp. 1281–1296, 2009.
5. J. Wu and L. Xue, "Model detection for additive models with longitudinal data," *Open Journal of Statistics*, vol. 4, pp. 868–878, 2014.
6. Y. Lin and H. Zhang, "Component selection and smoothing in multivariate nonparametric regression," *The Annals of Statistics*, vol. 34, pp. 2272–2297, 2006.
7. P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman, "Spam: sparse additive models," *Journal of the Royal Statistical Society, Series B*, vol. 71, pp. 1009–1030, 2009.
8. A. Chouldechova and T. Hastie, "Generalized additive model selection," *arXiv preprint arXiv: 1506.03850*, 2015.
9. M. Ueki, "A note on automatic variable selection suing smooth-threshold estimating equations," *Biometrika*, vol. 96, pp. 1005–1011, 2009.
10. G. Li, H. Lian, S. Feng, and L. Zhu, "Automatic variable selection for longitudinal generalized linear models," *Computational Statistics & Data Analysis*, vol. 61, pp. 174–186, 2013.
11. P. Zhao and G. Li, "Modified see variable selection for varying coefficient instrumental variable models," *Statistical Methodology*, vol. 12, pp. 60–70, 2013.
12. J. Lv, H. Yang, and C. Guo, "Smoothing combined generalized estimating equations in quantile partially linear additive models with longitudinal data," *Computational Statistics*, vol. 31, pp. 1203–1234, 2016.
13. J. Geronimia and G. Saportab, "Variable selection for multiply-imputed data with penalized generalized estimating equations," *Computational Statistics & Data Analysis*, vol. 110, pp. 103–114, 2017.
14. P. Zhao and L. Xue, "Variable selection for semiparametric varying coefficient partially linear models," *Statistics and Probability Letters*, vol. 79, pp. 2148–2157, 2009.
15. Y. Zhou, A. T. K. Wan, and X. Wang, "Estimating equations inference with missing data," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1187–1199, 2009.
16. L. Xue, "Consistent model selection for marginal generalized additive model for correlated data," *Journal of the American Statistical Association*, vol. 105, pp. 1518–1530, 2010.
17. C. de Boor, *A Practical Guide to Splines*. Springer, 2001.

