Journal of the Korean Statistical Society (2024) 53:1091–1109
https://doi.org/10.1007/s42952-024-00279-y

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Kernel machine in semiparametric regression with nonignorable missing responses

Zhenzhen Fu[1] · Ke Yang[1] · Yaohua Rong[1] · Yu Shu[1]

Received: 5 September 2023 / Accepted: 4 July 2024 / Published online: 26 July 2024

Abstract

Missing data is prevalent in many fields. Among all missing mechanisms, nonignorable missing data is more challenging for model identification. In this paper, we propose a semiparametric regression model estimation method with nonignorable missing responses. To be specific, we first construct a parametric model for the propensity score and apply the generalized method of moments to obtain the estimated propensity score. For nonignorable missing responses, based on the inverse probability weighting approach, we propose the penalized garrotized kernel machine method to flexibly depict the complex nonlinear relationships between the response and the predictors, allow for interactions between the predictors, and eliminate the redundant variables automatically. The cyclical coordinate descent algorithm is provided to solve the corresponding optimization problems. Numerical results and real data analysis indicate that our proposed method achieves better prediction performance compared with the competing ones.

Keywords Semiparametric model · Missing not at random · Reproducing Kernel Hilbert Space · Regularized estimation · Inverse probability weighting

Mathematics Subject Classification 62D10 · 62G05 · 62G08 · 62P10

1 Introduction

Missing data occurs frequently in various fields, including clinical trials, surveys, and social sciences. For instance, in clinical trials, participants experiencing deterioration are more likely to drop out compared to those improving. Similarly, high-income individuals may be hesitant to disclose earnings in questionnaires. Handling missing data often requires some assumptions about the missing

✉ Yaohua Rong
rongyaohua@bjut.edu.cn

[1] School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing 100124, China

mechanism. The missing mechanism, as outlined by Little and Rubin (1976), plays a crucial role in statistical modeling. If the response probability of the study variable does not directly depend on the study variable, the missing mechanism is called missing at random (MAR). In contrast, the response probability of the study variable depends directly on the study variable, the missing mechanism is called nonignorable or missing not at random (MNAR) (Little and Rubin 2019). Compared to MAR case, the nonignorable missingness is associated with unobserved values, which makes subsequent statistical inference more complicated. The missing mechanism should be taken into account in statistical inference to avoid estimation bias and prevent erroneous conclusions.

In recent years, various efficient methodologies have emerged to deal with missing data. For instance, imputation methods (Rubin and Schenker 1986; Chen and Van Keilegom 2013), likelihood methods (Lv and Li 2013), and robust estimation techniques (Bianco et al. 2011; Liu and Goldberg 2020) have primarily been employed to handle ignorable missing data. However, addressing nonignorable missing data presents a more intricate challenge compared to ignorable cases. Recent advancements have introduced approaches tailored for MNAR data. For example, Morikawa et al. (2017) proposed a semiparametric maximum likelihood method for estimating parameters within the propensity score model. Shao and Wang (2022) devised bias-corrected generalized estimating equations, leveraging inverse propensity weighting to address nonignorable dropout. As highlighted by Wang et al. (2014), Zhao and Shao (2015), and Shao and Wang (2016), a critical aspect in handling nonignorable missing response data involves addressing identifiability issues. To tackle this challenge, the introduction of nonresponse instrument variables has been proposed (Wang et al. 2014). Despite the difficulties, there is some literature considering statistical models with nonignorable missing responses, including but not limited to the following papers. Bahari et al. (2021) studied the general linear model based on the empirical likelihood ratio function with missing covariates or responses. Tang and Tang (2018) developed a penalized likelihood approach for the generalized partially nonlinear model with nonignorable missing responses. Zhang and Wang (2022) proposed a penalized empirical likelihood method for the partially linear quantile regression model with nonignorable missing responses. However, the methods used in these models rarely take into account the complex relationships between the covariates, especially for multidimensional data.

In view of the flexibility of nonparametric models and the easy interpretation of parametric models, semiparametric models have been widely used. Semiparametric models with missing responses at random are widely studied. For example, see Wang et al. (2004); Wang and Sun (2007); Bianco et al. (2011); Chen and Van Keilegom (2013), among others. However, to capture the complex relationships between the covariates and the response, we would like the model to be flexible enough to account for the nonlinearity. Under such circumstances, kernel machine methods have become a popular method in recent years. For example, for the complete data, see Liu et al. (2007); Chen et al. (2018); Rong et al. (2018); Zheng et al. (2021), among others. For missing data, Liu and Liu and Goldberg (2020) developed two new kernel machines to handle missing responses at random.

Recently, many works have extended the regularized estimation methods to nonparametric or semiparametric models. Actually, it is reasonable to assume that only a minority of covariates contribute to the response. Under the assumption of sparsity, many regularized methods built on the penalized least square regression or the likelihood function have been proposed to simultaneously estimate parameters and select important predictors. For example, see LASSO (Tibshirani 1996), SCAD (Fan and Li 2001), adaptive LASSO (Zou 2006), MCP (Zhang 2010) and others. As a nonparametric model estimating method, kernel machine method does not require predetermined function form. Compared to the general kernel machine, Rong et al. (2018) proposed a garrotized kernel machine method that allows for interactions between covariates in nonlinearity, which is an efficient method for complete data. However, this is not applicable when the data contains missing values. To our knowledge, kernel machine methods in existing studies have rarely been applied to the semiparametric models with nonignorable missing responses. Therefore, it is desirable to design a novel method that can concurrently perform parameter estimation and select important predictors in nonlinearity with nonignorable missing responses.

The main contributions of this paper are as follows. Compare with the work of Rong et al. (2018), we address the problem of estimating the semiparametric model in scenarios where the response is missing, yet the covariates remain observable, and the mechanism for missing data is nonignorable. Of particular significance is our proposal for a two-step estimation approach aimed at modeling the semiparametric model with nonignorable missing responses. In the first step, we impose a parametric model and utilize the generalized method of moments (GMM) (Hansen 1982) to estimate the propensity score model. We employ nonresponse instrument variables to address the identifiability problem, extending the approach proposed by Wang et al. (2014) to other models. In the second step, recognizing the complex relationships among variables and the presence of redundant variables in the model, we develop a novel penalized garrotized kernel machine method capable of handling nonignorable missing responses within the semiparametric model framework. We construct a penalized objective function that integrates missing responses using the inverse probability weighting approach, thereby ensuring improved prediction accuracy by eliminating redundant variables. Additionally, we design an effective cyclical coordinate descent algorithm to solve the corresponding optimization problem.

The rest of the paper is organized as follows. In Sect. 2, we propose the penalized garrotized kernel machine method with nonignorable missing responses (NMGKM). To be specific, we describe the model formulation and discuss the identifiability of the propensity score model and introduce the nonresponse instrument approach to estimate the propensity score model of missing data under MNAR. Then, by inverse probability weighting, we construct an penalty objective function with missing responses. In Sect. 3, we propose an efficient algorithm for the solution of the proposed method. Numerical studies are conducted to compare the performance of the proposed method with the existing ones in Sect. 4. In Sect. 5, we apply the proposed NMGKM method to analyse a real data example. Finally, we conclude the paper with a brief discussion in Sect. 6.

2 Methods

2.1 Models with nonignorable missing responses

For subjects $i = 1, \dots, n$, let $(Y_i, R_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ be independent random samples from $(y, r, \boldsymbol{x}, \boldsymbol{z})$, where y is the response variable subject to missingness, r is the response indicator of y ($r = 1$ if y is observed and $r = 0$ otherwise), $\boldsymbol{x} = (X_1, \dots, X_P)^{\mathrm{T}}$ and $\boldsymbol{z} = (Z_1, \dots, Z_Q)^{\mathrm{T}}$ are P- and Q-dimensional vectors of covariates that are fully observed. We consider the semiparametric relationship between the covariates $\boldsymbol{x}, \boldsymbol{z}$ and the response y as

$$y = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta} + f(\boldsymbol{z}) + \epsilon, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)^{\mathrm{T}}$ is a P-dimensional unknown parameter vector, $f(\cdot)$ is an unknown smooth function and ϵ is the random error distributed with $N(0, \sigma^2)$.

For the nonparametric component $f(\cdot)$, we assume that it lies in the Reproducing Kernel Hilbert Space $\mathcal{H}_K$ generated by a positive definite kernel function $K(\cdot, \cdot)$. Generally, the positive definite kernel function is called the inner product kernel or Mercer kernel. According to Mercer's theorem (Cristianini and Shawe-Taylor 2000), under some regularity conditions, the kernel function $K(\cdot, \cdot)$ implicitly determines a unique function space. For example, a popular Gaussian kernel function is $K(\boldsymbol{z}, \boldsymbol{z}') = \exp\left[-\sum_{q=1}^{Q} \left(Z_q - Z'_q\right)^2 / \rho\right]$, where $\rho > 0$ is known as the bandwidth.

In this paper, we adopt the garrotized kernel $K^{(g)}(\boldsymbol{z}, \boldsymbol{z}'; \boldsymbol{\delta})$ proposed by Rong et al. (2018). Specifically, the garrotized kernel is defined as

$$K^{(g)}\left(\boldsymbol{z}, \boldsymbol{z}'; \boldsymbol{\delta}\right) = K^{(g)}\left(\boldsymbol{\delta}^{\frac{1}{2}} \circ \boldsymbol{z}, \boldsymbol{\delta}^{\frac{1}{2}} \circ \boldsymbol{z}'\right),$$

where $\boldsymbol{\delta} = (\delta_1, \dots, \delta_Q)^{\mathrm{T}}$ is an unknown vector with $\delta_q \geq 0$ for $q = 1, \dots, Q$, and $\circ$ is called Hadamard product. Specifically, let A and B be $m \times n$ matrices, the Hadamard product of A and B is defined by $[A \circ B]_{ij} = [A]_{ij}[B]_{ij}$ for all $1 \leq i \leq m$, $1 \leq j \leq n$ (Styan 1973). Actually, the common Mercer kernel is a special case of the garrotized kernel with $\delta_q = 1$ for $q = 1, \dots, Q$. On the other hand, we can also obtain the garrotized version of the Gaussian kernel for the general case as $K^{(g)}(\boldsymbol{z}, \boldsymbol{z}'; \boldsymbol{\delta}) = \exp\left[-\sum_{q=1}^{Q} \delta_q (Z_q - Z'_q)^2\right]$, which is applied in this paper.

Compared with the common Mercer kernel, the garrotized kernel can better depict the complex relationship between the covariates and the response. In fact, δ_q reflects the marginal effect of covariate Z_q on the response for $q = 1, \dots, Q$. For example, $\delta_q = 0$ implies that Z_q has no prediction performance on the response, which indicates that the garrotized kernel machine method can be used to provide a variable selection scheme.

In this paper, we consider the situation where the missing mechanism $\pi(\boldsymbol{x}, \boldsymbol{z}, y) = \Pr(r = 1 | y, \boldsymbol{x}, \boldsymbol{z})$ is MNAR. The conditional probability $\Pr(r = 1 | y, \boldsymbol{x}, \boldsymbol{z})$ is called the propensity score of missingness (Rosenbaum and Rubin 1983). By the identification condition proposed by (Wang et al. 2014), we assume that the covariates $\boldsymbol{x}$ and z can be partitioned into two components $\boldsymbol{u}$ and $\boldsymbol{v}$, such that

$$\Pr(r = 1|y, \boldsymbol{x}, z) = \Pr(r = 1|y, \boldsymbol{u}, \boldsymbol{v}) = \Pr(r = 1|y, \boldsymbol{u}). \tag{2}$$

Equation (2) indicates that the covariate $\boldsymbol{v}$, which is called the nonresponse instrument variable, is independent of the propensity score of missingness given y and $\boldsymbol{u}$. Then, we assume that the parametric propensity score for a generalized linear model is

$$\pi(y, \boldsymbol{u}; \boldsymbol{\gamma}) = P(r = 1|y, \boldsymbol{u}; \boldsymbol{\gamma}) = \Psi(\gamma_0 + \boldsymbol{\gamma_1}^{\mathrm{T}}\ \boldsymbol{u} + \gamma_2\, y),$$

where $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma_1}^{\mathrm{T}}, \gamma_2)^{\mathrm{T}}$ is an m-dimensional unknown parameter vector to be estimated, and $\Psi(\cdot)$ is a pre-specified function taking values at $[0, 1]$. Due to the existence of the instrument variable $\boldsymbol{v}$, the dimension $m \leq 1 + P + Q$. When $\gamma_2 = 0$, the missing mechanism becomes MAR. As for $\Psi(\cdot)$, popular models include the exponential tilting model with $\Psi(s) = \left[1 + \exp(s)\right]^{-1}$, the Logistic model with $\Psi(s) = \exp(s)/\left[1 + \exp(s)\right]$ and complementary Log-log (cLog-log) model with $\Psi(s) = 1 - \exp\left[-\exp(s)\right]$, the probit model with $\Psi(\cdot)$ being the standard normal distribution function.

Following the nonresponse instrument approach proposed by Wang et al. (2014), we apply the generalized method of moments to obtain an estimator of $\boldsymbol{\gamma}$. To be specific, construct following estimation functions

$$g(\boldsymbol{x}, z, y, r, \boldsymbol{\gamma}) = s(\boldsymbol{x}, z)\left[\frac{r}{\pi(y, \boldsymbol{u}; \boldsymbol{\gamma})} - 1\right],$$

where $s(\boldsymbol{x}, z)$ is a known vector-valued function, and in this paper we take $s(\boldsymbol{x}, z) = (1, \boldsymbol{x}, z)$. Then, we can construct estimation equations based on the fact that $E\left[g(\boldsymbol{x}, z, y, r, \boldsymbol{\gamma})\right] = \mathbf{0}$. Let

$$G(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i, R_i, \boldsymbol{\gamma}),$$

where $G(\boldsymbol{\gamma})$ is an L-dimensional vector with m-dimensional unknown parameters.

When $L = m$ (indicating that the dimension of the instrument variable $\boldsymbol{v}$ is only one), we can directly estimate $\boldsymbol{\gamma}$ by solving the equations

$$G(\gamma) = \mathbf{0}. \tag{3}$$

When $L > m$ (indicating that the dimension of the instrument variable $\boldsymbol{v}$ is larger than one), the estimation Eq. (3) are over identified. Thus, we employ the two-step GMM to estimate $\boldsymbol{\gamma}$. Specifically, the first-step generalized moment estimator of $\boldsymbol{\gamma}$ is

$$\hat{\boldsymbol{\gamma}}^{(1)} = \arg\min_{\gamma \in \Gamma} G(\boldsymbol{\gamma})^{\mathrm{T}}\, G(\boldsymbol{\gamma}),$$

where Γ is the parameter space of $\boldsymbol{\gamma}$. Let $\hat{W}_\gamma$ be the $(1+P+Q)\times(1+P+Q)$ matrix with (l,l')th element being $1/n\sum_{i=1}^{n} g_l(\boldsymbol{X_i},\boldsymbol{Z_i},Y_i,R_i,\hat{\boldsymbol{\gamma}}^{(1)})g_{l'}(\boldsymbol{X_i},\boldsymbol{Z_i},Y_i,R_i,\hat{\boldsymbol{\gamma}}^{(1)})$, where $g_l(\cdot)$ is the lth element of the vector $g(\cdot)$. Then, the second-step generalized moment estimator of $\boldsymbol{\gamma}$ is

$$\hat{\boldsymbol{\gamma}} = \arg\min_{\gamma\in\Gamma} G(\boldsymbol{\gamma})^{\mathrm{T}}\,\hat{W}_\gamma^{-1} G(\boldsymbol{\gamma}).$$

2.2 Loss function for NMGKM

To estimate the unknown parameters as well as the nonparametric component for model (1), we construct the loss function. In contrast to the work of Rong et al. (2018), the objective function needs to be constructed based on inverse probability weighting since the response is missing. In this paper, the following penalized objective function is constructed as follows

$$h(f,\boldsymbol{\beta},\boldsymbol{\delta}) = \frac{1}{2n}\sum_{i=1}^{n}\frac{R_i}{\pi_i}[Y_i-\boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{\beta}-f(\boldsymbol{Z_i})]^2+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\|\boldsymbol{\delta}\|_1+\frac{1}{2}\lambda_3\|f\|^2_{\mathcal{H}_{K^{(g)}}}, \tag{4}$$

where $\pi_i=\pi(Y_i,\boldsymbol{U}_i;\boldsymbol{\gamma})$ is the propensity score of missingness for the ith subject, $\|\cdot\|_1$ denotes the L$_1$-norm and $\|\cdot\|_{\mathcal{H}_{K^{(g)}}}$ denotes the functional norm in the space $\mathcal{H}_{K^{(g)}}$ generated by the garrotized kernel. The first term in (4) is the quadratic loss weighted by the inverse probability of missingness. The inverse probability weighting method is used to avoid estimation bias due to missing data. The non-negative tuning parameters λ_1,λ_2 and λ_3 strike a balance between the complexity and the goodness of fit of the model. The L$_1$-norm penalty on $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ can obtain sparse estimation, which implements variable selection on $\boldsymbol{x}$ and $\boldsymbol{z}$. The functional norm penalty on $f(\cdot)$ controls the complexity and smoothness of the nonparametric component.

By the representer theorem (Kimeldorf and Wahba 1971), the nonparametric component $f(\cdot)$ that minimizes the loss function (4) can be written as

$$f(z)=\sum_{i=1}^{n}\alpha_i K^{(g)}\left(z,\boldsymbol{Z}_i;\boldsymbol{\delta}\right),$$

where $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_n)^{\mathrm{T}}$. Let $\boldsymbol{Y}=(Y_1,\dots,Y_n)^{\mathrm{T}},\boldsymbol{X}=(\boldsymbol{X_1},\dots,\boldsymbol{X_n})^{\mathrm{T}}$, $\boldsymbol{W}=\mathrm{diag}(R_i/\pi_i)$, $\boldsymbol{A}=\boldsymbol{W}^{1/2}$ and $\boldsymbol{K}(\boldsymbol{\delta})$ be an $n\times n$ Gram matrix with (i,j)th element being $K^{(g)}(\boldsymbol{Z_i},\boldsymbol{Z_j};\boldsymbol{\delta})$. Then minimizing the loss function (4) is equivalent to minimizing

$$h(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta})=\frac{1}{2n}\|\boldsymbol{A}[\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{K}(\boldsymbol{\delta})\boldsymbol{\alpha}]\|^2+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\|\boldsymbol{\delta}\|_1+\frac{1}{2}\lambda_3\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{\delta})\boldsymbol{\alpha}. \tag{5}$$

Together with the non-negative constraints on δ_q, the estimation for model (1) can be obtained by solving the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta}} \quad & h(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta}), \\ \text{s.t.} \quad & \delta_q \geq 0, \ \ q = 1,\ldots,Q. \end{aligned} \tag{6}$$

The solution proposed above concerning the objective function (6) is referred to as the NMGKM method. For the nonignorable missing response, an instrument variable is used to identify the parameters in the propensity score model. The proposed estimation method not only portrays the potential nonlinear relationship between the predictors and the missing response and achieves the interaction between nonparametric predictors, but also automatically removes the redundant variables in the parametric and nonparametric components, thus improving the prediction performance of the model.

3 Algorithm

The proposed NMGKM method is estimated by solving the optimization problem (6). In practice, the propensity score of missingness π_i is usually unknown and needs to be estimated. Therefore, this paper proposes a two-step estimation scheme. For the unknown propensity score of missingness π_i in the objective function $h(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta})$, we plug-in their estimates $\hat{\pi}_i = \pi(Y_i, \boldsymbol{U}_i;\hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ is obtained by GMM, and denote the corresponding matrix $\boldsymbol{W}, \boldsymbol{A}$ by $\hat{\boldsymbol{W}}, \hat{\boldsymbol{A}}$. Once the missing probability π_i is obtained, then the cyclic coordinate descent algorithm (Friedman et al. 2010) is proposed to solve the unknown parameters $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta})$ in optimization problem (6) with fixed tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$. Furthermore, the validation procedure is introduced for the selection of the tuning parameters.

3.1 Estimation for NMGKM

By plugging $\hat{\pi}_i$ in the objective function (5) and fixed tuning parameters, we obtain the unknown parameters $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta})$ by the following cyclical coordinate descent algorithm. Specifically, we first give initial values $(\boldsymbol{\alpha}^{(0)},\boldsymbol{\beta}^{(0)},\boldsymbol{\delta}^{(0)})$. Then successively update the parameters one at a time by the cyclically coordinate descent algorithm.

- Given $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(t-1)}$ and $\boldsymbol{\delta} = \boldsymbol{\delta}^{(t-1)}$, update $\boldsymbol{\beta}$ by solving

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(t)} = \arg\min_{\boldsymbol{\beta}} & \frac{1}{2n} \|\hat{\boldsymbol{A}}[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}]\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \\ & + \lambda_2 \|\boldsymbol{\delta}^{(t-1)}\|_1 + \frac{\lambda_3}{2} (\boldsymbol{\alpha}^{(t-1)})^{\mathrm{T}} \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}, \end{aligned}$$

which is equivalent to solving

$$\hat{\boldsymbol{\beta}}^{(t)} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1, \tag{7}$$

where $\boldsymbol{Y}^* = \hat{\boldsymbol{A}}\boldsymbol{Y} - \hat{\boldsymbol{A}}\boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}^{(t-1)}$ and $\boldsymbol{X}^* = \hat{\boldsymbol{A}}\boldsymbol{X}$. The optimization (7) can be solved by LASSO regression and we denote the updated value for $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(t)}$.

- Given $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\delta} = \boldsymbol{\delta}^{(t-1)}$, update $\boldsymbol{\alpha}$ by solving

$$\arg\min_{\boldsymbol{\alpha}} \frac{1}{2n} \|\hat{\boldsymbol{A}}[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)} - \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}]\|^2 + \lambda_1 \|\boldsymbol{\beta}^{(t)}\|_1 + \lambda_2 \|\boldsymbol{\delta}^{(t-1)}\|_1 + \frac{\lambda_3}{2} \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha},$$

which is equivalent to solving

$$\hat{\boldsymbol{\alpha}}^{(t)} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2n} \|\hat{\boldsymbol{A}}[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)} - \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}]\|^2 + \frac{\lambda_3}{2} \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\boldsymbol{\alpha}. \tag{8}$$

Noting that the objective function in (8) is a quadratic function of $\boldsymbol{\alpha}$, take the first derivative of (8) with respect to $\boldsymbol{\alpha}$ and obtain the updated $\boldsymbol{\alpha}^{(t)}$ by solving the linear equations

$$[\boldsymbol{K}^{\mathrm{T}}(\boldsymbol{\delta}^{(t-1)})\hat{\boldsymbol{W}}\boldsymbol{K}(\boldsymbol{\delta}^{(t-1)}) + n\lambda_3 \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})]\boldsymbol{\alpha} = \boldsymbol{K}(\boldsymbol{\delta}^{(t-1)})\hat{\boldsymbol{W}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}). \tag{9}$$

When the coefficient matrix of $\boldsymbol{\alpha}$ on the left-hand side of (9) is singular, a diagonal matrix with the element being 10^{-5} is added to make it invertible.

- Given $\boldsymbol{\beta} = \boldsymbol{\beta}^{(\boldsymbol{t})}$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(\boldsymbol{t})}$, update $\boldsymbol{\delta}$ by minimizing

$$\arg\min_{\boldsymbol{\delta}} \frac{1}{2n} \|\hat{\boldsymbol{A}}[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)} - \boldsymbol{K}(\boldsymbol{\delta})\boldsymbol{\alpha}^{(t)}]\|^2 + \lambda_1 \|\boldsymbol{\beta}^{(t)}\|_1 + \lambda_2 \|\boldsymbol{\delta}\|_1 + \frac{\lambda_3}{2} (\boldsymbol{\alpha}^{(t)})^{\mathrm{T}} \boldsymbol{K}(\boldsymbol{\delta})\boldsymbol{\alpha}^{(t)}.$$

To update $\boldsymbol{\delta}$, we take the cyclic coordinate descent method one at a time. Specifically, to estimate δ_s, we fix $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}, \boldsymbol{\alpha} = \boldsymbol{\alpha}^{(t)}$ and $\delta_q = \delta_q^{(t-1)}, q = 1, \ldots, Q, q \neq s$. Then minimize

$$\begin{aligned} &\frac{1}{2n} \|\hat{\boldsymbol{A}}[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)} - \boldsymbol{K}(\delta_s;\delta_q^{(t-1)})\boldsymbol{\alpha}^{(t)}]\|^2 + \lambda_2 \sum_{q=1, q\neq s}^{Q} \delta_q^{(t-1)} \\ &\quad + \lambda_2 \delta_s + \frac{\lambda_3}{2} (\boldsymbol{\alpha}^{(t)})^{\mathrm{T}} \boldsymbol{K}(\delta_s;\delta_q^{(t-1)})\boldsymbol{\alpha}^{(t)}, \end{aligned} \tag{10}$$

which is a nonlinear optimization problem of δ_s. The univariate nonlinear constrained optimization in R is applied to obtain the updated $\delta_s^{(t)}$.

The above algorithms can be briefly described in Algorithm 1.

Algorithm 1 Solve optimization problem (6)

```
1: Give the instrument variable $\boldsymbol{v}$, $\hat{\gamma}$ is obtained by GMM.
2: Compute $\hat{\pi}_i = \pi(Y_i, \boldsymbol{U}_i; \hat{\gamma})$.
3: Initialize $\boldsymbol{\theta}^0 = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\delta}^{(0)})$ with $\delta_q^{(0)} \geq 0$ for $q = 1, \ldots, Q$.
4: $t \Leftarrow t + 1$.
5: Compute $\boldsymbol{\beta}^{(t)}$ by minimizing (7).
6: Compute $\boldsymbol{\alpha}^{(t)}$ from the Eq. (9).
7: $q \Leftarrow q + 1$.
8: Compute $\delta_q^{(t)}$ by minimizing (10).
9: Repeat 4-7 until convergence.
```

3.2 Selection of the tuning parameters and evaluation of the model performance

The validation method is applied to select the tuning parameters and evaluate the model performance. Cross-validation is often used for tuning parameter selection but can be computationally inconvenient. Instead, to select the optimal tuning parameters, the model performance is calculated on the validation dataset. To be specific, given a fixed set of the tuning parameters λ, we estimate the proposed NMGKM method based on the training set. With the missing response, we examine the performance of the competing estimation methods on the validation set by the mean square prediction error (MSPE) as follows

$$\mathrm{MSPE} = \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} R_i \left[Y_i - \boldsymbol{X_i}^{\mathrm{T}} \hat{\boldsymbol{\beta}} - \hat{f}(\boldsymbol{Z}_i) \right]^2.$$

For different values of the tuning parameters, a smaller MSPE on the validation set indicates the better choice of λ. In practice, we set a large range of λ for a rough search, constantly narrow the range based on the MSPE of the validation set, and eventually obtain a set of relatively reasonable values. With the chosen tuning parameters and the estimated NMGKM method, we can evaluate the prediction performance of the model by MSPE on the test set.

4 Simulation studies

In this section, we evaluate the finite sample performance of the proposed method under different scenarios. We compare the proposed method with the weighted-complete-case kernel machines (WCC) method proposed by Liu and Liu and Goldberg (2020) and the linear model with nonignorable missing responses (LMN) proposed by Shao and Wang (2016). The WCC method mainly considered the estimation problem of approximating nonparametric functions with Gaussian kernel in MAR.

In all the following simulation settings, ϵ_i is independently distributed with the standard normal distribution $N(0,\ 1)$, and $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are generated independently from uniform distribution $U(-1, 1)$ and $U(0,\ 1)$, respectively. The indicator variable r is generated from Bernoulli distribution with probability function being specified as a exponential tilting (ExpT) form

$$P(r = 1|y, \boldsymbol{u}; \boldsymbol{\gamma}) = 1/[1 + \exp(\gamma_0 + \boldsymbol{\gamma_1}^{\mathrm{T}}\ \boldsymbol{u} + \gamma_2\ y)],$$

or a Logistic form

$$P(r = 1|y, \boldsymbol{u}; \boldsymbol{\gamma}) = \exp(\gamma_0 + \boldsymbol{\gamma_1}^{\mathrm{T}}\ \boldsymbol{u} + \gamma_2\ y)/[1 + \exp(\gamma_0 + \boldsymbol{\gamma_1}^{\mathrm{T}}\ \boldsymbol{u} + \gamma_2\ y)],$$

or a cLog-log form

$$P(r=1|y,\boldsymbol{u};\boldsymbol{\gamma}) = 1-\exp[-\exp(\gamma_0+\boldsymbol{\gamma_1}^{\mathrm{T}}\ \boldsymbol{u}+\gamma_2\, y)],$$

when $\gamma_2 = 0$, the missing mechanism is MAR.

The prediction performances of the fitted models are measured using the MSPE and the average absolute Bias based on the test sets. The average absolute Bias with missing responses is defined as

$$\text{Bias} = \frac{1}{\sum_{i=1}^{n} R_i}\sum_{i=1}^{n} R_i|Y_i - \boldsymbol{X_i}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - \hat{f}(\mathbf{Z}_i)|.$$

4.1 Comparison with WCC

The WCC method focused on the estimation of the nonparametric model in MAR. To better highlight the advantages of the kernel function in the NMGKM method, we first consider the data from the nonparametric model as follows

$$y = f(z) + \epsilon. \tag{11}$$

We conduct the four settings (**S1**–**S4**) to generate data with missing responses for model (11), where QT denotes the true number of relevant covariates in z.

S1: $Q=5,\ QT=5,\ \boldsymbol{u}=(Z_4,Z_5)^{\mathrm{T}},\ f(z)=\cos(Z_1)-1.5Z_2^2+\exp(-Z_3)Z_4-0.8\sin(Z_5)\cos(Z_3)+2Z_1Z_5.$

S2: $Q=10,\ QT=5,\ \boldsymbol{u}=(Z_6,\dots,Z_{10})^{\mathrm{T}}$, and $f(z)$ is the same as **S1**. Thus, it implies that the fitted model contains 5 additional irrelevant variables in z.

S3: $Q=10,\ QT=10,\ \boldsymbol{u}=(Z_6,\dots,Z_{10})^{\mathrm{T}},\ f(z)=\cos(Z_1)-1.5Z_2^2+\exp(-Z_3)Z_4-0.8\sin(Z_5)\cos(Z_3)$ $+2Z_1Z_5+0.9Z_6\sin(Z_7)-0.8\cos(Z_6)Z_7+2Z_8\sin(Z_9)\sin(Z_{10})-1.5Z_8^3-Z_8Z_9-0.1\exp(Z_{10})\cos(Z_{10}).$

S4: $Q=20,\ QT=10,\ \boldsymbol{u}=(Z_{11},\dots,Z_{20})^{\mathrm{T}}$, and $f(z)$ is the same as **S3**. Thus, it implies that the fitted model contains 10 additional irrelevant variables in z.

The settings **S1**–**S4** consider different complex forms with the interaction between the covariates in the nonparametric component. The settings **S2** and **S4** take into account the addition of redundant variables in the nonparametric component.

We consider the following missing mechanism in three categories

ExpT1: $\gamma_0=-4(\log 3)/3,\ \boldsymbol{\gamma_1}=(2(\log 3)/15,\dots,2(\log 3)/15)^{\mathrm{T}},\ \gamma_2=0.$

Logistic1: $\gamma_0=0.85,\ \boldsymbol{\gamma_1}=(-(\log 5)/5,\dots,-(\log 5)/5)^{\mathrm{T}},\ \gamma_2=0.$

cLog-log1: $\gamma_0=0.82,\ \boldsymbol{\gamma_1}=(-0.08,\dots,-0.08)^{\mathrm{T}},\ \gamma_2=0.$

For **S1**–**S4**, the observed percentages under each category are approximately 70% to 80%, 50% to 60%, 80% to 90%, respectively. For the NMGKM method and WCC method, we need validation sets to select tuning parameters and test sets to evaluate the prediction performance of the model. All results are based on the sample sizes $n=300$, 600 and 900, and the samples are divided into a training set, a validation set and a test set with equal sample size. With 100 simulation replications, we report the MSPE, Bias and the corresponding standard deviation(SD) for the NMGKM method and WCC method in Table 1.

Table 1 The average MSPE(SD) and Bias(SD) of 100 Monte Carlo simulations by the NMGKM method and the WCC method in nonparametric model

n	Set	(Q, QT)	NMGKM		WCC	
			MSPE(SD)	Bias (SD)	MSPE (SD)	Bias (SD)
ExpT1						
300	**S1**	(5, 5)	**16.0 (7.56)**	**31.1 (7.88)**	18.8 (6.07)	34.1 (5.89)
	S2	(10, 5)	**21.7 (10.3)**	**36.3 (9.25)**	25.6 (6.01)	40.8 (5.23)
	S3	(10, 10)	**21.0 (7.37)**	**36.2 (6.44)**	23.5 (6.61)	38.7 (5.94)
	S4	(20, 10)	**28.8 (10.2)**	**42.4 (8.03)**	31.9 (7.13)	45.6 (5.79)
600	**S1**	(5, 5)	**7.57 (3.46)**	**21.3 (4.86)**	9.98 (3.79)	24.7 (4.45)
	S2	(10, 5)	**10.2 (3.85)**	**24.7 (4.88)**	13.6 (3.41)	29.0 (3.80)
	S3	(10, 10)	**15.0 (5.99)**	**30.2 (6.09)**	15.9 (3.57)	31.4 (3.80)
	S4	(20, 10)	**23.3 (7.43)**	**38.1 (5.99)**	25.9 (4.43)	40.8 (3.67)
900	**S1**	(5, 5)	**6.10 (2.09)**	**19.2 (3.41)**	7.92 (1.73)	21.7 (2.62)
	S2	(10, 5)	**8.77 (2.69)**	**23.0 (3.86)**	12.2 (2.71)	27.6 (3.00)
	S3	(10, 10)	**10.7 (2.83)**	**25.5 (3.58)**	13.0 (2.81)	28.2 (3.10)
	S4	(20, 10)	**12.6 (3.45)**	**27.8 (3.85)**	17.8 (3.31)	33.5 (3.25)
Logistic1						
300	**S1**	(5, 5)	**19.7 (9.66)**	**34.5 (8.78)**	26.7 (9.29)	41.1 (8.05)
	S2	(10, 5)	**25.0 (10.5)**	**39.5 (9.22)**	28.5 (12.9)	42.2 (10.2)
	S3	(10, 10)	**25.2 (10.1)**	**39.6 (8.35)**	25.8 (9.48)	40.2 (7.62)
	S4	(20, 10)	**34.5 (11.6)**	**46.9 (8.35)**	37.7 (11.3)	49.4 (7.97)
600	**S1**	(5, 5)	**8.79 (4.78)**	**22.7 (5.91)**	9.86 (3.93)	24.5 (4.89)
	S2	(10, 5)	**19.0 (7.71)**	**34.2 (7.36)**	25.2 (4.84)	40.7 (4.02)
	S3	(10, 10)	**15.6 (4.70)**	**31.1 (4.77)**	17.6 (4.56)	33.0 (4.40)
	S4	(20, 10)	**29.0 (10.3)**	**42.8 (8.12)**	32.7 (6.61)	46.3 (5.33)
900	**S1**	(5, 5)	**7.71 (4.78)**	**21.3 (6.00)**	6.85 (1.96)	20.4 (3.19)
	S2	(10, 5)	**11.0 (3.75)**	**25.9 (4.77)**	13.7 (3.68)	29.2 (3.98)
	S3	(10, 10)	**13.1 (3.70)**	**28.2 (4.05)**	15.4 (3.83)	30.9 (3.92)
	S4	(20, 10)	**21.5 (6.13)**	**36.8 (5.61)**	24.0 (5.28)	39.1 (4.76)
cLog-log1						
300	**S1**	(5, 5)	**15.2 (7.32)**	**30.3 (7.52)**	13.0 (6.02)	28.1 (6.70)
	S2	(10, 5)	**20.6 (9.64)**	**35.5 (9.00)**	21.0 (6.18)	36.4 (5.65)
	S3	(10, 10)	**20.3 (7.33)**	**35.5 (6.40)**	26.1 (6.90)	40.7 (5.82)
	S4	(20, 10)	**28.2 (9.28)**	**42.1 (7.58)**	25.7 (5.82)	40.5 (5.11)
600	**S1**	(5, 5)	**7.98 (5.42)**	**21.3 (6.41)**	7.47 (2.94)	21.1 (3.92)
	S2	(10, 5)	**10.8 (4.27)**	**25.5 (5.22)**	12.7 (2.95)	28.0 (3.44)
	S3	(10, 10)	**12.0 (4.43)**	**27.1 (4.80)**	13.6 (3.11)	29.0 (3.57)
	S4	(20, 10)	**15.6 (4.45)**	**31.2 (4.59)**	20.1 (3.44)	35.8 (3.21)
900	**S1**	(5, 5)	**6.11 (3.72)**	**19.0 (5.20)**	5.53 (1.62)	18.3 (2.79)
	S2	(10, 5)	**8.63 (2.68)**	**22.8 (3.90)**	10.7 (2.44)	25.8 (2.93)
	S3	(10, 10)	**10.2 (2.81)**	**24.9 (3.56)**	11.8 (2.56)	26.9 (2.91)
	S4	(20, 10)	**13.2 (3.64)**	**28.5 (4.09)**	17.0 (2.75)	32.7 (2.76)

All entries in MSPE(SD) and Bias(SD) are the result of multiplying by 100

The bolded results in the table represent the MSPE and Bias of the proposed method respectively, which means better prediction performance in different settings

From Table 1, it can be seen that the NMGKM method has smaller average MSPE and Bias than that of the WCC method in the four settings. This may be because the garrotized kernel in the NMGKM method is more flexible than the Gaussian kernel, which allows all δ_q to be unequal. When irrelevant variables are added, as in **S2** and **S4**, it can be seen that the average MSPE and Bias of the NMGKM method is significantly smaller than the WCC method. The reason is that the NMGKM method can eliminate irrelevant variables to implement variable selection. And increasing the sample size n improves the accuracy of prediction as expected. In addition, the NMGKM method performs better if we know the relevant variables in the true set.

Since the proposed NMGKM method is applicable to the semiparametric model, we further compare the prediction performance of the two methods for the following semiparametric model

$$y = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta} + f(z) + \epsilon.$$

The missing categories are the same as ExpT1, Logistic1 and cLog-log1, and we consider the following simulation settings (**S5-S8**) with PT and QT being the true number of relevant variables in $\boldsymbol{x}$ and z, respectively.

S5: $P = 1,\ Q = 5,\ PT = 1,\ QT = 5,\ \boldsymbol{\beta} = 1,\ \boldsymbol{u} = (Z_3, Z_4, Z_5)^{\mathrm{T}}$, and $f(z)$ is the same as **S1**.

S6: $P = 5,\ Q = 10,\ PT = 1,\ QT = 5,\ \boldsymbol{\beta} = (1, 0, 0, 0, 0)^{\mathrm{T}},\ \boldsymbol{u} = (X_2, \dots, X_5, Z_6, \dots, Z_{10})^{\mathrm{T}}$, and $f(z)$ is the same as **S1**. Thus, it implies that the fitted model contains 4 additional irrelevant variables in $\boldsymbol{x}$ and 5 additional irrelevant variables in z.

S7: $P = 2,\ Q = 10,\ PT = 2,\ QT = 10,\ \boldsymbol{\beta} = (1, 1)^{\mathrm{T}},\ \boldsymbol{u} = (Z_5, \dots, Z_{10})^{\mathrm{T}}$, and $f(z)$ is the same as **S3**.

S8: $P = 5,\ Q = 20,\ PT = 2,\ QT = 10,\ \boldsymbol{\beta} = (1, 1, 0, 0, 0)^{\mathrm{T}},\ \boldsymbol{u} = (X_3, \dots, X_5, Z_{11}, \dots, Z_{20})^{\mathrm{T}}$, and $f(z)$ is the same as **S3**. Thus, it implies that the fitted model contains 3 additional irrelevant variables in $\boldsymbol{x}$ and 10 additional irrelevant variables in z.

For **S5**–**S8**, the observed percentages under each category are approximately 70% to 80%, 40% to 60%, 80% to 90%, respectively. Since the WCC method does not consider the linear part, we put all the covariates $\boldsymbol{x}$ and z into the nonparametric component. For 100 simulation repetitions, we report the MSPE(SD) and Bias(SD) for the NMGKM method and WCC method in Table 2.

From Table 2, we observe that the average MSPE and Bias of the NMGKM method is always smaller than the WCC method for the four simulation settings (**S5**–**S8**). In fact, the proposed NMGKM method can be applied to the semiparametric model, and the correct model hypothesis is conducive to improving the prediction accuracy of the model. Moreover, as the dimension of covariates and the number of irrelevant variables increase, the prediction performance of the NMGKM method is more excellent. When the missing proportion is high, it can be concluded that increasing the sample size can improve the prediction performance of the model.

4.2 Comparison with the LMN

Shao and Wang (2016) considered the estimation of the mean of response in the linear model with the nonignorable missing response based on inverse probability

Table 2 The average MSPE(SD) and Bias(SD) of 100 Monte Carlo simulations by the NMGKM method and the WCC method in semiparametric model

n	Set	(P, Q)	NMGKM		WCC	
			MSPE (SD)	Bias (SD)	MSPE (SD)	Bias (SD)
ExpT1						
300	**S5**	(1, 5)	**19.3 (9.90)**	**34.1 (8.93)**	25.4 (18.2)	38.5 (12.2)
	S6	(5, 10)	**33.7 (9.31)**	**46.5 (7.26)**	37.8 (11.0)	48.7 (7.21)
	S7	(2, 10)	**26.3 (10.7)**	**40.5 (8.33)**	33.8 (11.3)	45.4 (7.41)
	S8	(5, 20)	**42.7 (12.4)**	**52.2 (8.24)**	52.4 (15.9)	57.1 (9.07)
600	**S5**	(1, 5)	**8.05 (2.79)**	**22.1 (3.93)**	12.1 (3.18)	26.9 (3.66)
	S6	(5, 10)	**20.3 (9.72)**	**35.1 (8.99)**	26.6 (5.02)	40.6 (4.04)
	S7	(2, 10)	**15.2 (4.74)**	**30.8 (4.95)**	22.8 (6.37)	37.1 (5.20)
	S8	(5, 20)	**21.3 (6.64)**	**36.4 (5.72)**	37.6 (8.03)	48.3 (5.36)
900	**S5**	(1, 5)	**6.86 (3.01)**	**20.4 (3.92)**	10.0 (2.34)	24.3 (2.89)
	S6	(5, 10)	**10.2 (2.89)**	**25.0 (3.92)**	22.2 (4.36)	37.1 (3.76)
	S7	(2, 10)	**17.7 (9.38)**	**32.6 (8.74)**	20.2 (4.02)	35.1 (3.50)
	S8	(5, 20)	**15.3 (3.75)**	**30.7 (3.93)**	29.2 (5.46)	42.5 (4.06)
Logistic1						
300	**S5**	(1, 5)	**25.8 (12.0)**	**39.9 (10.0)**	26.2 (13.1)	39.7 (10.3)
	S6	(5, 10)	**37.6 (14.6)**	**48.7 (9.96)**	45.0 (11.5)	53.4 (8.74)
	S7	(2, 10)	**37.5 (13.5)**	**49.0 (9.29)**	45.2 (21.0)	52.7 (12.5)
	S8	(5, 20)	**64.2 (28.3)**	**63.8 (13.8)**	73.4 (25.4)	67.7 (12.0)
600	**S5**	(1, 5)	**11.4 (6.51)**	**26.0 (7.21)**	14.6 (4.14)	29.6 (4.26)
	S6	(5, 10)	**17.4 (6.33)**	**32.8 (6.13)**	31.2 (6.66)	44.1 (4.87)
	S7	(2, 10)	**24.2 (11.6)**	**38.5 (9.37)**	20.0 (7.61)	42.0 (5.56)
	S8	(5, 20)	**49.2 (14.7)**	**56.2 (8.70)**	54.4 (18.7)	58.4 (9.50)
900	**S5**	(1, 5)	**7.54 (2.35)**	**21.4 (3.36)**	12.1 (2.78)	26.9 (3.25)
	S6	(5, 10)	**32.3 (7.55)**	**45.9 (6.17)**	35.6 (6.27)	47.6 (4.37)
	S7	(2, 10)	**16.1 (4.76)**	**16.1 (4.76)**	26.0 (6.61)	39.7 (4.92)
	S8	(5, 20)	**40.4 (7.49)**	**51.1 (4.73)**	42.6 (11.7)	51.5 (6.62)
cLog-log1						
300	**S5**	(1, 5)	**16.6 (7.82)**	**31.7 (7.82)**	19.8 (10.2)	34.5 (10.2)
	S6	(5, 10)	**32.9 (8.96)**	**45.9 (7.24)**	36.2 (10.8)	47.5 (7.02)
	S7	(2, 10)	**23.5 (9.42)**	**38.2 (7.91)**	33.1 (13.6)	44.9 (8.43)
	S8	(5, 20)	**37.9 (12.4)**	**49.0 (9.00)**	51.3 (18.6)	56.4 (10.2)
600	**S5**	(1, 5)	**7.98 (3.40)**	**21.8 (4.69)**	11.9 (3.41)	26.8 (3.77)
	S6	(5, 10)	**11.4 (3.22)**	**26.5 (4.12)**	24.3 (4.36)	38.8 (3.68)
	S7	(2, 10)	**14.5 (5.54)**	**29.9 (5.60)**	22.3 (5.23)	36.7 (4.42)
	S8	(5, 20)	**22.4 (7.44)**	**37.3 (6.26)**	34.0 (6.40)	45.9 (4.42)
900	**S5**	(1, 5)	**6.22 (3.03)**	**19.2 (4.50)**	9.37 (2.17)	23.6 (2.80)
	S6	(5, 10)	**11.4 (4.76)**	**26.3 (5.03)**	21.0 (3.83)	36.1 (3.42)
	S7	(2, 10)	**11.3 (2.78)**	**26.4 (3.38)**	18.7 (4.08)	33.6 (3.57)
	S8	(5, 20)	**20.4 (7.34)**	**35.4 (6.52)**	27.6 (5.04)	41.3 (3.73)

All entries in MSPE (SD) and Bias (SD) are the result of multiplying by 100

The bolded results in the table represent the MSPE and Bias of the proposed method respectively, which means better prediction performance in different settings

weighting. In this section, we compare the proposed NMGKM method with the LMN method.

We consider the following three missing cases

ExpT2: $\gamma_0 = -0.8,\ \boldsymbol{\gamma_1} = (-0.1, 0.1, \dots, -0.1, 0.1, \dots)^{\mathrm{T}},\ \gamma_2 = -0.4.$

Logistic2: $\gamma_0 = 0.4,\ \boldsymbol{\gamma_1} = (-0.1, \dots, -0.1)^{\mathrm{T}},\ \gamma_2 = -0.2.$

cLog-log2: $\gamma_0 = 1.1,\ \boldsymbol{\gamma_1} = (-0.5, \dots, -0.5)^{\mathrm{T}},\ \gamma_2 = 0.8.$

We compare the prediction performance of the two methods in the following semiparametric model with different missing categories

$$y = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta} + f(z) + \epsilon.$$

We consider the same simulation settings as for **S5**–**S8**. For the above three missing categories, the propensity score depends on Y_i, which represents three nonignorable missing data cases. For **S5**–**S8**, the observed percentages under each category are approximately 70% to 75%, 45% to 55%, 70% to 85%, respectively.

Since the LMN method does not consider the nonparametric component, we assume the linear relationship for all the covariates $\boldsymbol{x}$ and $\boldsymbol{z}$. For each simulation, we consider $n = 300$, 600 and 900. For the NMGKM method, the sample is divided into a training set, a validation set and a test set with equal sample size. For the LMN method, since there is no tuning parameter to be selected, two-thirds of the sample is divided into the training set for model estimation and the rest into the test set for prediction performance evaluation. With 100 simulation repetitions, we report the MSPE(SD) and Bias(SD) for the NMGKM method and LMN method in Table 3.

Table 3 shows the average MSPE (SD) and Bias (SD) of the proposed method and the LMN method for the four settings in three missing categories. The simulation results clearly indicate that the proposed method almost always outperforms the LMN method in terms of MSPE and Bias. This is possible because the proposed method correctly specifies the model, while the LMN method misspecifies the model when the data is generated from the semiparametric model. That is to say if there are complex nonlinear structures, the NMGKM method captures these effects better than the LMN method and hence gains higher prediction accuracy. Moreover, when the redundant variables are included in the model, we can significantly see that the average MSPE and Bias of the NMGKM method is smaller than the LMN method. Interestingly, the proposed method has better prediction performance even when the missing probability is large.

5 Analysis of the real data sets

In this section, we compare the performance of our proposed NMGKM method, the WCC method and the LMN method on the AIDS clinical trial data from the AIDS Clinical Trials Group Study 175 (ACTG 175) (Hammer et al. 1996), which is available in the R package speff2trial. ACTG 175 is a randomized clinical trial to compare monotherapy with zidovudine or didanosine with combination therapy with zidovudine and didanosine or zidovudine and zalcitabine in adults infected with the human

Table 3 The average MSPE(SD) and Bias(SD) of 100 Monte Carlo simulations by the NMGKM method and the LMN method in semiparametric model

			NMGKM		LMN	
n	Set	(P, Q)	MSPE (SD)	Bias (SD)	MSPE (SD)	Bias (SD)
ExpT2						
300	**S5**	(1, 5)	**21.0 (9.67)**	**35.8 (8.74)**	28.1 (18.0)	42.9 (14.1)
	S6	(5, 10)	**28.7 (9.42)**	**42.4 (7.86)**	36.8 (17.3)	47.7 (10.5)
	S7	(2, 10)	**18.4 (7.24)**	**33.6 (6.51)**	27.2 (9.52)	41.3 (7.24)
	S8	(5, 20)	**42.1 (9.63)**	**52.0 (6.65)**	62.1 (20.6)	62.7 (10.6)
600	**S5**	(1, 5)	**19.4 (10.6)**	**19.1 (7.43)**	23.7 (12.4)	34.2 (10.5)
	S6	(5, 10)	**13.9 (4.98)**	**29.2 (5.56)**	16.3 (4.52)	32.6 (5.16)
	S7	(2, 10)	**16.0 (5.45)**	**31.6 (5.53)**	19.8 (7.89)	35.4 (7.06)
	S8	(5, 20)	**51.7 (10.9)**	**55.2 (6.08)**	53.3 (12.6)	57.7 (7.11)
900	**S5**	(1, 5)	**8.17 (3.69)**	**22.1 (4.77)**	18.4 (6.38)	35.9 (7.13)
	S6	(5, 10)	**10.9 (4.18)**	**25.8 (4.89)**	15.1 (4.46)	31.4 (5.10)
	S7	(2, 10)	**13.8 (3.67)**	**29.2 (4.13)**	15.7 (4.76)	31.9 (5.21)
	S8	(5, 20)	**16.4 (4.77)**	**31.8 (4.46)**	22.2 (5.55)	37.8 (5.00)
Logistic2						
300	**S5**	(1, 5)	**22.9 (13.7)**	**22.9 (13.7)**	30.5 (14.5)	43.9 (11.3)
	S6	(5, 10)	**24.9 (11.0)**	**39.3 (8.80)**	54.2 (24.5)	57.9 (12.5)
	S7	(2, 10)	**27.1 (11.5)**	**41.5 (8.35)**	41.4 (17.9)	50.9 (11.3)
	S8	(5, 20)	**48.0 (12.6)**	**54.9 (7.38)**	93.5 (39.3)	76.9 (16.1)
600	**S5**	(1, 5)	**13.1 (7.07)**	**27.9 (6.85)**	14.4 (3.94)	30.5 (4.74)
	S6	(5, 10)	**17.6 (7.44)**	**33.0 (7.02)**	18.1 (4.68)	33.8 (4.42)
	S7	(2, 10)	**18.7 (6.64)**	**34.1 (6.17)**	23.3 (8.85)	38.0 (7.17)
	S8	(5, 20)	**23.8 (6.80)**	**38.6 (5.45)**	41.5 (14.0)	51.1 (8.61)
900	**S5**	(1, 5)	**9.86 (4.12)**	**24.5 (5.26)**	12.9 (2.91)	28.0 (3.58)
	S6	(5, 10)	**17.8 (9.66)**	**32.9 (8.83)**	20.3 (4.69)	36.0 (4.22)
	S7	(2, 10)	**15.8 (5.95)**	**31.2 (5.62)**	16.8 (5.40)	32.4 (5.31)
	S8	(5, 20)	**21.0 (7.11)**	**36.3 (5.83)**	24.6 (8.04)	39.2 (6.34)
cLog-log2						
300	**S5**	(1, 5)	**29.2 (8.06)**	**43.4 (7.17)**	30.0 (15.1)	46.6 (13.6)
	S6	(5, 10)	**33.0 (6.35)**	**45.9 (4.71)**	34.9 (13.3)	46.5 (8.81)
	S7	(2, 10)	**35.3 (8.82)**	**47.0 (6.35)**	34.8 (12.9)	46.7 (9.17)
	S8	(5, 20)	**48.7 (16.5)**	**55.6 (9.96)**	65.3 (28.0)	63.9 (12.0)
600	**S5**	(1, 5)	**21.0 (10.1)**	**35.6 (10.0)**	35.6 (13.7)	53.0 (12.0)
	S6	(5, 10)	**18.5 (6.23)**	**34.1 (5.72)**	25.5 (9.15)	40.2 (7.26)
	S7	(2, 10)	**20.6 (6.66)**	**35.9 (5.65)**	24.7 (8.79)	39.6 (7.32)
	S8	(5, 20)	**35.0 (10.7)**	**46.7 (7.29)**	37.8 (12.7)	48.6 (8.40)
900	**S5**	(1, 5)	**10.7 (6.42)**	**24.8 (7.30)**	38.0 (13.7)	55.4 (11.7)
	S6	(5, 10)	**15.3 (6.11)**	**30.5 (6.00)**	17.3 (5.77)	33.7 (6.18)
	S7	(2, 10)	**18.5 (6.36)**	**33.8 (5.73)**	21.9 (8.30)	37.6 (7.74)
	S8	(5, 20)	**33.9 (10.6)**	**46.1 (7.41)**	47.6 (18.0)	54.5 (10.5)

All entries in MSPE (SD) and Bias (SD) are the result of multiplying by 100

The bolded results in the table represent the MSPE and Bias of the proposed method respectively, which means better prediction performance in different settings

Table 4 Average MSPE(SD) and Bias(SD) of the three methods for 100 replications of the ACTG 175 data

Methods	Missing Mechanism	MSPE(SD)	Bias(SD)
NMGKM	**MNAR**	**64.8 (9.44)**	**63.6 (4.51)**
WCC	**MAR**	69.4 (8.88)	66.0 (4.40)
LMN	**MNAR**	66.5 (10.6)	64.1 (5.40)

All entries in MSPE(SD) and Bias(SD) are the result of multiplying by 100

The bolded results in the table represent the MSPE and Bias of the proposed method respectively, which means better prediction results in real data set

immunodeficiency virus whose CD4 cell counts is between 200 and 500 per cubic millimeter.

Among all the 2139 patients, We consider $n = 532$ patients with treatment zidovudine monotherapy. Let the CD4 counts at 96 ± 5 weeks be the response variable Y, and consider the following eight covariates: gender (X_1), CD4 counts at baseline (CD40, X_2), CD4 counts at 20 ± 5 weeks (CD420, X_3), age (Z_1), weight (Z_2), number of days of previously received antiretroviral therapy (preanti, Z_3), CD8 counts at baseline (CD80, Z_4), and CD8 counts at 20 ± 5 weeks (CD420, Z_5). Due to death and dropout, some observations on the response Y are subject to missingness while the covariates are completely observed. The missing proportion is about 40%. Previous experiences from doctors indicated that HIV-infected patients with low CD4 counts are more likely to drop out of the trial. Thus, the missing CD4 counts is nonignorable.

As an illustration, we consider the following semiparametric regression model $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + f(z)$, where the covariates $\boldsymbol{x} = (X_1, X_2, X_3)^{\mathrm{T}}$ and $z = (Z_1, \ldots, Z_5)^{\mathrm{T}}$. To unify the scales roughly, all of the continuous covariates are standardized. For the mechanism of missingness, we consider the model $\pi(y, \boldsymbol{u}; \boldsymbol{\gamma}) = 1/\left[1 + \exp(\gamma_0 + \boldsymbol{\gamma_1}^{\mathrm{T}}\ \boldsymbol{u} + \gamma_2\ y)\right]$, where $\boldsymbol{u} = (X_1, X_2, X_3, Z_3, Z_4, Z_5)^{\mathrm{T}}$. Also, the age and weight are always observed and thus can be used as the instrument variable $\boldsymbol{v}$. For the categorical variable gender, we introduce one dummy variable into the linear part of the model.

We randomly assign the 532 patients 100 times and follow the estimation and model evaluation procedure the same as those of the simulation section. The average MSPE and Bias for the 100 replications is given in Table 4.

From Table 4, we can clearly see that the NMGKM method has better prediction performance. Compared to the WCC method, our garrotized kernel allows for interactions between the covariates. The NMGKM method has better prediction performance compared to the linear model method, which suggests that the linear procedure is not flexible enough for capturing the complex relationship.

6 Conclusion

In this paper, we propose a penalized garrotized kernel machine method with nonignorable missing responses. Current researches mainly focus on ignorable missing responses. Dealing with nonignorable missing responses is a highly challenging problem, mainly because of the identifiability of the propensity score model. To deal with the issue, nonresponse instrument variables play a crucial role. In this paper, we construct a parametric propensity score model not involving instrument variables and apply the GMM approach to estimate the unknown parameters. Furthermore, based on the penalized garrotized kernel machine, our proposed NMGKM method can not only capture the complex relationships between the covariates and the response and allow for possible interactions among covariates, but also remove the irrelevant predictors automatically. The ability for model selection reduces the model complexity so that our proposed method has the potential to achieve better prediction accuracy compared to the competing methods.

In the future, several issues need to be further investigated to construct complex relationship between the response and the covariates for nonignorable missing data. For example, we will explore the extension of the NMGKM method to other models, such as the quantile regression model. Besides, the mixed-effects model with longitudinal data also deserves further study for its wide application.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No.11701021, No.72371241, No.11971001), National Social Science Fund of China (No.23 & ZD127), National Statistical Science Research Project (No. 2022LZ22) and Science and Technology Program of Beijing Education Commission (No. KM202110005013).

Funding National Natural Science Foundation of China (No. 11701021, No. 72371241, No.11971001), National Statistical Science Research Project (No. 2022LZ22), Science and Technology Program of Beijing Education Commission (No. KM202110005013), National Social Science Fund of China (No.23&ZD127).

Declarations

Conflict of interest No Conflict of interest nor Conflict of interest.

References

Bahari, F., Parsi, S., & Ganjali, M. (2021). Empirical likelihood inference in general linear model with missing values in response and covariates by MNAR mechanism. *Statistical Papers, 62*(2), 591–622.

Bianco, A., Boente, G., González-Manteiga, W., & Pérez-González, A. (2011). Asymptotic behavior of robust estimators in partially linear models with missing responses: the effect of estimating the missing probability on the simplified marginal estimators. *Test, 20*(3), 524–548.

Chen, J., Zhang, C., Kosorok, M. R., & Liu, Y. (2018). Double sparsity kernel learning with automatic variable selection and data extraction. *Statistics and Its Interface, 11*(3), 401.

Chen, S. X., & Van Keilegom, I. (2013). Estimation in semiparametric models with missing data. *Annals of the Institute of Statistical Mathematics, 65*(4), 785–805.

Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., & Niu, M. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine, 335*(15), 1081–1090.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society, 50*(4), 1029–1054.

Kimeldorf, G., & Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications, 33*(1), 82–95.

Little, R. J., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. Hoboken: Wiley.

Liu, D., Lin, X., & Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics, 63*(4), 1079–1088.

Liu, T., & Goldberg, Y. (2020). Kernel machines with missing responses. *Electronic Journal of Statistics, 14*(2), 3766–3820.

Lv, X., & Li, R. (2013). Smoothed empirical likelihood analysis of partially linear quantile regression models with missing response variables. *AStA Advances in Statistical Analysis, 97*, 317–347.

Morikawa, K., Kim, J. K., & Kano, Y. (2017). Semiparametric maximum likelihood estimation with data missing not at random. *Canadian Journal of Statistics, 45*(4), 393–409.

Rong, Y., Zhao, S. D., Zhu, J., Yuan, W., Cheng, W., & Li, Y. (2018). More accurate semiparametric regression in pharmacogenomics. *Statistics and Its Interface, 11*(4), 573.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American statistical Association, 81*(394), 366–374.

Shao, J., & Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika, 103*(1), 175–187.

Shao, Y., & Wang, L. (2022). Generalized partial linear models with nonignorable dropouts. *Metrika, 85*(2), 223–252.

Styan, G. P. (1973). Hadamard products and multivariate statistical analysis. *Linear Algebra and Its Applications, 6*, 217–240.

Tang, N., & Tang, L. (2018). Estimation and variable selection in generalized partially nonlinear models with nonignorable missing responses. *Statistics and Its Interface, 11*(1), 1–18.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Wang, Q., Linton, O., & Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association, 99*(466), 334–345.

Wang, Q., & Sun, Z. (2007). Estimation in partially linear models with missing responses at random. *Journal of Multivariate Analysis, 98*(7), 1470–1493.

Wang, S., Shao, J., & Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica, 24*, 1097–1116.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*(2), 894–942.

Zhang, T., & Wang, L. (2022). Smoothed partially linear quantile regression with nonignorable missing response. *Journal of the Korean Statistical Society, 51*(2), 441–479.

Zhao, J., & Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association, 110*(512), 1577–1590.

Zheng, X., Rong, Y., Liu, L., & Cheng, W. (2021). A more accurate estimation of semiparametric logistic regression. *Mathematics, 9*(19), 2376.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.