

# Statistical inference for missing data mechanisms

Yang Zhao 

Department of Mathematics and  
Statistics, University of Regina, Regina,  
Saskatchewan, Canada

## Correspondence

Yang Zhao, Department of Mathematics  
and Statistics, University of Regina,  
College West 307.14, Regina, SK S4S 0A2,  
Canada.

Email: zhaoyang@uregina.ca

## Funding information

Natural Sciences and Engineering  
Research Council of Canada,  
Grant/Award Number: YZ

In the literature of statistical analysis with missing data there is a significant gap in statistical inference for missing data mechanisms especially for nonmonotone missing data, which has essentially restricted the use of the estimation methods which require estimating the missing data mechanisms. For example, the inverse probability weighting methods (Horvitz & Thompson, 1952; Little & Rubin, 2002), including the popular augmented inverse probability weighting (Robins et al, 1994), depend on sufficient models for the missing data mechanisms to reduce estimation bias while improving estimation efficiency. This research proposes a semiparametric likelihood method for estimating missing data mechanisms where an EM algorithm with closed form expressions for both E-step and M-step is used in evaluating the estimate (Zhao et al, 2009; Zhao, 2020). The asymptotic variance of the proposed estimator is estimated from the profile score function. The methods are general and robust. Simulation studies in various missing data settings are performed to examine the finite sample performance of the proposed method. Finally, we analysis the missing data mechanism of Duke cardiac catheterization coronary artery disease diagnostic data to illustrate the method.

## KEYWORDS

EM algorithm, missing data mechanism, nonmonotone missing data pattern, pseudo-likelihood

## 1 | INTRODUCTION

Missing data are a common problem in statistical analysis. In practice data can be missing either by design or happenstance.<sup>1,2</sup> It's well known that missing data mechanism is a key point which needs to be considered in statistical analysis with missing data. The three classes of missing data mechanisms<sup>3</sup> are (i) missing complete at random where the missingness is independent of both observed data and unobserved data, (ii) missing at random (MAR) where the missingness is independent of missing data given observed data, and (iii) missing not at random where the missingness depends on unobserved data given observed data. One other feature about the missing data is missing data pattern which includes monotone and nonmonotone patterns. In general arbitrary nonmonotone missing data patterns increase the difficulties of statistical analysis compared with a simple monotone missing data pattern.

In statistical analysis, people often make assumptions about the missing data mechanism, for example, data are MAR, without proper testing of the assumption. An invalid assumption may produce a significant biased result. On the other hand, many methods for statistical analysis with missing data require estimating the missing data mechanisms, for example, the popular augmented inverse probability weighting estimating equation<sup>4,18</sup> depends on sufficient models for the missing data mechanisms to reduce estimation bias while improving

estimation efficiency. However, developing methods for modeling the missing data mechanisms for nonmonotone missing data patterns is challenging even for the commonly treated MAR data. There has been very limited research in this area.<sup>5,6</sup>

Estimation method<sup>7</sup> based on a class of models for the missing data mechanisms for the randomized monotone missingness processes is complex and computationally intensive, which is still not able to be implemented in any computing software. The multinomial model<sup>6</sup> for estimating the missing data probability for MAR data with nonmonotone missing data patterns assumes that for each missing data pattern the missingness depends on the fully observed variables with that missing data pattern. It models the probability for each missing data pattern separately using a logistic regression model based on the fully observed variables in that pattern and estimates the model using only the subset of data with those variables fully observed. Then the probability of being a complete observation is computed as a combined result from the set of logistic regression models. We note that (i) the method cannot estimate the missing data probabilities directly, (ii) estimating the missing mechanism for each pattern separately is not efficient as it only use a subset of the data especially when the sizes of some of the subsets are small, and (iii) the model has natural restrictions as discussed in the article, for example, in the simulation study it sets a limited range for the covariates as  $(X_1, X_2, X_3) \in [0, 2]^3$ , otherwise a constrained Bayesian estimation method is required for evaluating the estimates to overcome the natural restrictions of the model.

In general parametric regression models can only use fully observed variables to make inference for missing data mechanisms, which is often insufficient as the missingness may depend on the partially observed variables in a general MAR setting.<sup>8</sup> This research develops methods for statistical inference for missing data mechanisms through a parametric regression model based on both fully observed and partially observed variables. We propose estimating the regression model directly from a semiparametric likelihood model using an EM algorithm.<sup>9,10</sup> The rest of the article is organized as follows. Section 2 introduces notation and the regression model for the missing data mechanism. In Section 3, we start with a simple monotone missing data pattern to introduce a semiparametric likelihood method for estimating the missing data mechanism, then we extend the model to deal with arbitrary nonmonotone missing data patterns. Finally it introduces a method to estimate the asymptotic variance of the semiparametric maximum likelihood estimator through profile score functions. Section 4 examines the finite sample performance of the proposed method and the asymptotic variance estimator in simulation studies. In Section 5, we analysis the missing data mechanism of Duke cardiac catheterization coronary artery disease diagnostic data. Although in general the MAR assumption cannot be tested we will explain how to test it in the simple monotone missing data pattern using the real data example. Some comments and a brief discussion are given in Section 6.

## 2 | NOTATION

Let  $X$  be a vector of variables with data missing in nonmonotone missing data patterns,  $n$  be the sample size,  $i$  be the index for subject and  $i = 1, \dots, n$ ,  $V$  be an index set of complete observations and  $\bar{V}$  be the complement of  $V$ . That is, if  $i \in V$  we have a complete observation and we denote it as  $X_i$ , if  $i \in \bar{V}$  we have an incomplete observation. According to the observed variables for the incomplete observations in  $\bar{V}$  we further divide  $\bar{V}$  into  $K$  subsets,  $\bar{V}^k$ ,  $k = 1, \dots, K$  such that the observations in the same subset have the same variables being observed. Therefore,  $\bar{V}^k$  is an index set of incomplete observations with the same missing data pattern. Let  $m$  be the number of variables in  $X$ , then the number of missing data patterns  $K < 2^m$ . For convenience, we denote an incomplete observation as  $(U_i^k, Z_i^k)$ , for  $i \in \bar{V}^k$ , where  $Z_i^k$  represents the observed part of  $X_i$  and  $U_i^k$  is the unobserved part of  $X_i$ .

Let  $R$  be an indicator variable,  $R_i = 1$  if  $i \in V$ , and 0 otherwise. Without loss of generality, we assume that the probability of being a complete observation is bounded away from zero with probability 1, that is,

$$Pr(R_i = 1 | X_i) > c > 0,$$

for a fixed positive constant  $c$ , and the observed data are independent realizations of the random vector  $(R, X^T)^T$ .

Assume that  $f_\alpha(r|x) = Pr(R = r | X = x)$  is a parametric regression model, and  $\alpha$  is a vector of parameters. We are interested in estimating  $\alpha$  in the regression model  $f_\alpha(r|x)$  for MAR data.

### 3 | METHODS

To investigate the estimation method, we start with a semiparametric likelihood model for a simple monotone missing data pattern, then we extend the model to deal with nonmonotone missing data patterns.

#### 3.1 | Simple monotone missing data pattern

In a simple case, we assume that there are only two variables in the vector  $X$ , that is,  $m = 2$ , one of them is fully observed and the other one has missing values. We note that in this case  $K = 1$  and we have a simple monotone missing data pattern. Using the notation in Section 2, for  $i \in V$  we observe  $X_i$ , and for  $i \in \bar{V}^1$  we have  $(U_i^1, Z_i^1)$  where  $U_i^1$  represents the missing part of  $X_i$  and  $Z_i^1$  is the observed part of  $X_i$ . We will omit the superscript 1 without confusion as there is only one missing group. Therefore,  $U$  denotes the variable with missing values and  $Z$  is the fully observed variable. The regression model for the missing data probability  $f_\alpha(r|x)$  can be written as  $f_\alpha(r|u, z)$ .

Let  $G(u|z)$  be the conditional distribution of  $U$  given  $Z = z$ . To estimate  $\alpha$  we consider the following likelihood function.

$$L(\alpha, G) = \prod_{i \in V} f_\alpha(r_i | u_i, z_i) dG(u_i | z_i) \prod_{i \in \bar{V}} \int f_\alpha(r_i | u, z_i) dG(u | z_i), \quad (1)$$

which was originally proposed for regression models in a two-phase study.<sup>9</sup> To avoid parametric assumption for the conditional distribution  $G(u|z)$  we model it using a piecewise empirical distribution.<sup>11</sup> Let's define a categorical variable  $H = h(Z)$  such that  $H$  has a few categories, for example, we let the function  $h(\cdot)$  divide the observed values of  $Z$  into a few categories such that there is approximately the same number of observations in each category. Let

$$g_u^h = Pr\{U = u | H = h\}, \quad \text{for } u \in \mathcal{U}_h, \quad h \in \mathcal{H}, \quad (2)$$

be the probability mass assigned to the pair of data  $(u, h)$ , where  $\mathcal{H}$  and  $\mathcal{U}_h$  are the sample space of  $H$  and  $U$  given  $H = h$ , respectively, with  $\mathcal{U}_h = \{u_i : i \in V \text{ and } H_i = h\}$  according to the empirical distribution principle. The log likelihood in (1) is then

$$l(\alpha, g) = \sum_{i \in V} [\log\{f_\alpha(r_i | u_i, z_i)\} + \log(g_{u_i}^{h_i})] + \sum_{i \in \bar{V}} \log \left\{ \sum_{u \in \mathcal{U}_{h_i}} f_\alpha(r_i | u, z_i) g_u^{h_i} \right\}, \quad (3)$$

where  $\sum_{u \in \mathcal{U}_h} g_u^h = 1$  for each  $h \in \mathcal{H}$ . We note that although the method requires dividing the continuous covariate  $Z$  into a few categories such that we can estimate the covariate distribution  $G(u|z)$  using an empirical distribution for each category separately the original continuous covariate  $Z$  is still used in the regression model  $f_\alpha(r|u, z)$  (see Equation (3)). The system of score functions is obtained as follows.

$$S_\alpha = \frac{\partial l(\alpha, g)}{\partial \alpha} = \sum_{i \in V} S_\alpha(r_i | u_i, z_i) + \sum_{i \in \bar{V}} \sum_{u \in \mathcal{U}_{h_i}} W_i^u S_\alpha(r_i | u, z_i), \quad (4)$$

$$S_{g_u^h} = \frac{\partial l(\alpha, g)}{\partial g_u^h} = \frac{n_{uh}^V + \sum_{i \in \bar{V}_h} W_i^u}{g_u^h} - \frac{n_{u'h}^V + \sum_{i \in \bar{V}_h} W_i^{u'}}{g_{u'}^h}, \quad \text{for } u \in \mathcal{U}_h, \quad h \in \mathcal{H}, \quad (5)$$

where  $S_\alpha(r_i | u_i, z_i) = \partial \log\{f_\alpha(r_i | u_i, z_i)\} / \partial \alpha$  is the score function of  $\alpha$  of a complete observation,  $n_{uh}^V$  is the number of observations in  $V$  with  $(U, H) = (u, h)$ ,  $\bar{V}_h$  is the subset of indices  $i$  in  $\bar{V}$  with  $h_i = h$ ,  $u'$  is a specified value of  $U$  in  $\mathcal{U}_h$ , and the weight

$$W_i^u = \frac{f_\alpha(r_i | u, z_i) g_u^{h_i}}{\sum_{u \in \mathcal{U}_{h_i}} f_\alpha(r_i | u, z_i) g_u^{h_i}}. \quad (6)$$

To compute semiparametric maximum likelihood estimates (SPMLE's) of  $\alpha$  and  $g_u^h$  we solve the system of score equations  $S_\alpha = 0$  and  $S_{g_u^h} = 0$  simultaneously. First, from  $S_{g_u^h} = 0$  we get

$$g_u^h = \frac{n_{uh}^V + \sum_{i \in \bar{V}_h} W_i^u}{n_h^{V \cup \bar{V}}}. \quad (7)$$

Giving initial estimates  $\alpha^{(0)}$  and  $g_u^{h(0)}$ , an EM algorithm<sup>9</sup> for obtaining these estimates is as follows. (i) Computing  $W_i^{u^{(t+1)}}$  in (6) at current  $\alpha^{(t)}$  and  $g_u^{h(t)}$ , (ii) Computing  $\alpha^{(t+1)}$  from  $S_\alpha = 0$  at given  $W_i^{u^{(t+1)}}$ , and (iii) Updating  $g_u^{h(t+1)}$  in (7) at given  $W_i^{u^{(t+1)}}$ . We repeat the above steps iteratively until convergence to get the SPMLE's  $\hat{\alpha}$  and  $\hat{g}_u^h$ .

We emphasize the following facts of the semiparametric model. (i) The support  $\mathcal{U}_h$  for the empirical distribution in (2) is defined based on the complete observations which is reliable when  $Pr(U|Z) = Pr(U|Z, R = 1)$ , that is, the data are MAR. (ii) The updating Equations (6) and (7) indicate that the covariate distribution is estimated not only using the complete observations but also the incomplete observations and the logistic regression model. (iii) The model can deal with cases where both  $U$  and  $Z$  are vectors of variables. (iv) The categorical variable  $H$  is used in modeling the conditional distribution  $G(u|z)$ . When the correlation  $Corr(U, Z)$  is high we may let  $H$  have more categories to capture the strong association. In finite samples in order to use all the incomplete observations in the analysis the method requires that for each category of  $H$  in the missing group there is at least one observation in  $V$  with a common  $H$ . In our simulations the number of categories for each variable in  $Z$  from 2 to 5 produce reasonable good results (see the numerical studies in Sections 4 and 5).

### 3.2 | Nonmonotone missing data pattern

Direct extension of the above SPMLE to nonmonotone missing data patterns is complex and computationally intensive. This section introduces the pseudo-likelihood model,<sup>10</sup> also referred to as composite likelihoods,<sup>12</sup> to estimate the missing data probability for nonmonotone missing data.

Using the notation in Section 2, in a general nonmonotone missing data pattern, there are  $K < 2^m$  missing data groups. Let  $G(u^k|z^k)$  be the conditional distribution of  $U^k$  given  $Z^k = z^k$ ,  $k = 1, \dots, K$ . We consider the following likelihood functions.

$$L^k(\alpha, G) = \prod_{i \in V} f_\alpha(r_i|u_i^k, z_i^k) dG(u_i^k|z_i^k) \prod_{k=1}^K \prod_{i \in \bar{V}^k} \int f_\alpha(r_i|u^k, z_i^k) dG(u^k|z_i^k), \quad \text{for } k = 1, \dots, K. \quad (8)$$

We define a categorical variable  $H^k = h(Z^k)$  such that  $H^k$  has a few categories, and we assume that  $g_{u^k}^{h^k}$  is the probability mass assigned to the pair of data  $(u^k, h^k)$ , that is

$$g_{u^k}^{h^k} = Pr\{U^k = u^k | H^k = h^k\}, \quad \text{for } u^k \in \mathcal{U}_{h^k}, \quad h^k \in \mathcal{H}^k, \quad (9)$$

where  $\mathcal{H}^k$  and  $\mathcal{U}_{h^k}$  are the sample space of  $H^k$  and  $U^k$  given  $h(Z^k) = h^k$ , respectively, with  $\mathcal{U}_{h^k} = \{u_i^k : i \in V \text{ and } H_i^k = h^k\}$ . Then from (8), we obtain the pseudo-log-likelihoods of  $(\alpha, g)$  as

$$l^k(\alpha, g) = \sum_{i \in V} [\log\{f_\alpha(r_i|u_i^k, z_i^k)\} + \log(g_{u_i^k}^{h_i^k})] + \sum_{k=1}^K \sum_{i \in \bar{V}^k} \log \left\{ \sum_{u^k \in \mathcal{U}_{h_i^k}} f_\alpha(r_i|u^k, z_i^k) g_{u^k}^{h_i^k} \right\}, \quad \text{for } k = 1, 2, \dots, K, \quad (10)$$

where  $\sum_{u^k \in \mathcal{U}_{h^k}} g_{u^k}^{h^k} = 1$  for each  $h^k \in \mathcal{H}^k$ . The system of score functions is as follows.

$$S_\alpha = \sum_{i \in V} S_\alpha(r_i|u_i^k, z_i^k) + \sum_{k=1}^K \sum_{i \in \bar{V}^k} \sum_{u^k \in \mathcal{U}_{h_i^k}} W_i^{u^k} S_\alpha(r_i|u^k, z_i^k), \quad (11)$$

$$S_{g_{u^k}^{h^k}} = \frac{n_{u^k h^k}^V + \sum_{i \in \bar{V}_{h^k}^k} W_i^{u^k}}{g_{u^k}^{h^k}} - \frac{n_{u^{k'} h^k}^V + \sum_{i \in \bar{V}_{h^k}^k} W_i^{u^{k'}}$$

where  $n_{u^k h^k}^V$  is the number of observations in  $V$  with  $(U^k, H^k) = (u^k, h^k)$ ,  $\bar{V}_{h^k}^k$  is the subset of indices  $i$  in  $\bar{V}^k$  with  $h_i^k = h^k$ ,  $u^{k'}$  is a specified value of  $U^k$  in  $\mathcal{U}_{h^k}$ , and

$$W_i^{u^k} = \frac{f_\alpha(r_i | u^k, z_i^k) g_{u^k}^{h_i^k}}{\sum_{u^k \in \mathcal{U}_{h_i^k}^k} f_\alpha(r_i | u^k, z_i^k) g_{u^k}^{h_i^k}}. \tag{13}$$

The SPMLE's for  $\alpha$  and  $g_{u^k}^{h^k}$  can be obtained by solving the score equations  $S_\beta = 0$  and  $S_{g_{u^k}^{h^k}} = 0$ ,  $k = 1, \dots, K$  according to the EM algorithm in Section 3.1 with updating equations for  $W_i^{u^k}$  and  $g_{u^k}^{h^k}$  replaced with (13) and (14), respectively.

$$g_{u^k}^{h^k} = \frac{n_{u^k h^k}^V + \sum_{i \in \bar{V}_{h^k}^k} W_i^{u^k}}{n_{h^k}^{V \cup \bar{V}^k}}, \text{ for } u^k \in \mathcal{U}_{h^k}, \quad h^k \in \mathcal{H}^k, \quad k = 1, \dots, K. \tag{14}$$

We note that the regular full likelihood methods<sup>8,13,14</sup> require estimating the distribution of the multivariate covariates which is often a high-dimensional object. The above pseudo-likelihoods replace the high-dimensional object with a low-dimensional empirical model for the conditional distribution of covariates which significantly reduce the computation complexity.

### 3.3 | Asymptotic variance estimation

In certain semiparametric settings profile log likelihoods behave like regular log likelihoods.<sup>15</sup> Methods for estimating the asymptotic variances of the SPMLE's from profile log likelihoods or the profile score functions have been investigated for regression analysis with missing data,<sup>9,10</sup> where they consider modeling the profile log likelihood as a quadratic function of the regression parameters at a close neighborhood of the SPMLE's or using the inverse of the negative Hessian matrix through numerical differentiation of the profile score functions.

Our semiparametric models using the pseudo-log-likelihoods in (3) and (10) to approximate the log-likelihoods, which produce the score function  $S_\alpha$ . We know that the score function is a function of the parameter of interest  $\alpha$  and the nuisance parameter  $g$ , so we write it as  $S_\alpha(\alpha, g)$ . The profile score for  $\alpha$ ,  $S_\alpha\{\alpha, g(\alpha)\}$  can be easily evaluated using the EM algorithm for any given  $\alpha$ . When  $f_\alpha(r|x)$  is the true model we would expect that the profile score for  $\alpha$ ,  $S_\alpha\{\alpha, g(\alpha)\}$  behaves like a regular score function.<sup>15</sup> Let  $I_\alpha\{\alpha, g(\alpha)\}$  denote the profile information matrix for  $\alpha$ , then under certain regularity conditions we have

$$E[S_\alpha\{\alpha, g(\alpha)\} S_\alpha\{\alpha, g(\alpha)\}^T] = E[I_\alpha\{\alpha, g(\alpha)\}].$$

Therefore, we can estimate the asymptotic variances of the SPMLE's by estimating  $E[S_\alpha\{\hat{\alpha}, g(\hat{\alpha})\} S_\alpha\{\hat{\alpha}, g(\hat{\alpha})\}^T]$ , which can be computed directly from the score functions in (4) or (11) and does not require numerical differentiation or further iteration with the EM algorithm.

## 4 | SIMULATIONS

In this section, we use simulation study to examine the finite sample performance of the proposed SPMLE's and the estimator of the asymptotic variance. For comparison, we also compute the maximum likelihood estimates (MLE's) based on the full data.

We consider two simulations with different missing data models. In simulation one, we generate data from the following logistic regression model

$$\text{logit}\{Pr(X_4 = 1|X_1, X_2, X_3)\} = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3, \quad (15)$$

where  $X_1, X_2$  and  $X_3$  are standard normal, and both  $X_2$  and  $X_3$  are correlated with  $X_1$  with a common correlation equal to 0.20. Let  $\beta_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}) = (-1, 2, 1, 1)$ . Variables  $X_1$  and  $X_4$  are fully observed, but  $X_2$  and  $X_3$  have data MAR. We divide the data into two groups randomly, in one group we let  $X_3$  fully observed but  $X_2$  partially observed, and in the other group we let  $X_2$  fully observed but  $X_3$  partially observed. Let  $X = (X_1, X_2, X_3, X_4)$ ,  $R_2$  and  $R_3$  be the missing data indicator for  $X_2$  and  $X_3$ , respectively. We generate  $R_2$  and  $R_3$  from the following logistic regression models, respectively.

$$\begin{aligned} \text{logit}\{Pr(R_2 = 1|X)\} &= \beta_{20} + \beta_{21}X_1 + \beta_{22}X_3 + \beta_{23}X_4, \\ \text{logit}\{Pr(R_3 = 1|X)\} &= \beta_{30} + \beta_{31}X_1 + \beta_{32}X_2 + \beta_{33}X_4. \end{aligned}$$

Let  $\beta_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23})$  and  $\beta_3 = (\beta_{30}, \beta_{31}, \beta_{32}, \beta_{33})$ . We consider the following three settings. (i) The missingness only depends on the fully observed variables with  $\beta_2 = \beta_3 = (-0.5, 0.5, 0, -0.5)$ , (ii) the missingness depends on both the full observed and partially observed variables with  $\beta_2 = \beta_3 = (-0.5, 0.5, 0.5, -0.5)$ , and (iii) the missingness only depends on the partially observed variable with  $\beta_2 = \beta_3 = (-0.5, 0, 0.5, 0)$ . We let the sample size  $n = 1500$ . In each setting there are about 500 complete observations.

In the second simulation, we consider the multinomial missing data model.<sup>6</sup> We generate  $(X_1, X_2, X_3)$  from the truncated normal distributions with  $X_1 \sim N(\mu = 0, \sigma = 0.5)$ ,  $X_2 \sim N(\mu = X_1 + X_1^2, \sigma = 0.5)$  and  $X_3 \sim N(\mu = X_2 + 0.8X_1X_2, \sigma = 0.5)$  on the support  $(X_1, X_2, X_3) \in [0, 2]^3$ . Then  $X_4$  is generated from the logistic regression model in (15) at  $\beta_1 = (-2.5, 0.7, 0.8, 1)$ . For the two pairs of variables  $(X_1, X_4)$  and  $(X_2, X_3)$ , we assume that the variables in the same pair are either missing or observed together and the missingness follows a multinomial distribution. We generate  $R_1$  from a multinomial distribution with

$$\begin{aligned} \text{logit}\{Pr(R_1 = 2|X)\} &= \gamma_{20} + \gamma_{21}X_1 + \gamma_{22}X_4, \\ \text{logit}\{Pr(R_1 = 3|X)\} &= \gamma_{30} + \gamma_{31}X_2 + \gamma_{32}X_3, \end{aligned}$$

and  $Pr(R_1 = 1|X) = 1 - Pr(R_1 = 2|X) - Pr(R_1 = 3|X)$ . We observe  $(X_1, X_4)$  for  $R_1 = 2$  or  $(X_2, X_3)$  for  $R_1 = 3$ . If  $R_1 = 1$  we have a complete observation, otherwise we have an incomplete observation. Let  $\gamma_2 = (\gamma_{20}, \gamma_{21}, \gamma_{22})$  and  $\gamma_3 = (\gamma_{30}, \gamma_{31}, \gamma_{32})$ . We set  $\gamma_3 = (-1.2, 0.3, 0.3)$ , and  $\gamma_2 = (-0.8, 0.2, 0)$ ,  $(-0.8, 0.2, -0.2)$ , or  $(-0.8, 0.2, 0.5)$ . Let sample size  $n = 2000$ . There are about 300 to 400 complete observations in each case.

We estimate the missing data probability using the following logistic regression model.

$$\text{logit}\{Pr(R = 1|X)\} = \alpha_0 + \alpha_1X_1 + \alpha_2X_2 + \alpha_3X_3 + \alpha_4X_4,$$

where  $R = 1$  if  $X$  is fully observed and 0 otherwise. In simulation one  $R = 1$  if both  $R_2$  and  $R_3$  equal 1 and in simulation two  $R = 1$  if  $R_1$  equals 1, and 0 otherwise. The values of the true  $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$  used in computing the biases and the mean squared errors are estimated from large samples with sample size  $n = 1\,000\,000$ . To compute the SPMLE's, we define  $H^k$ ,  $k = 1$  or 2 such that it divides each continuous variable into two and three categories for simulation one and simulation two, respectively.

Table 1 shows the simulation results for the SPMLE's and the full data MLE's based on 1000 replications. We see that for the two simulations in all the different settings (i) both the SPMLE's and the full data MLE's have small biases, the biases of the SPMLE's are slightly bigger than those of the full data MLE's when the missingness depends on the partially observed variables, (ii) in simulation one the roots of mean squared errors (RMSE's) of the SPMLE's are very close to those of the full data MLE's, while in simulation two the RMSE's for the SPMLE's are consistently smaller than those of the full data MLE's and (iii) the empirical standard deviations (*s.d.*'s) are close to the average of the corresponding *s.e.*'s computed from the inverse of the estimated  $ES_\alpha\{\hat{\alpha}, g(\hat{\alpha})\}S_\alpha\{\hat{\alpha}, g(\hat{\alpha})\}^T$  for the coefficients of the fully observed covariates, while the *s.e.*'s for the coefficients of the partially observed covariates are slightly bigger than the corresponding *s.d.*'s. In general the simulation results indicate that the performance of the proposed method is acceptable for estimating the missing data probability for MAR data in practice.

For the purpose of comparison in simulation one, we have also computed the IPW estimates of  $E(X_2)$  and  $E(X_3)$  using two different weights, one from the proposed SPMLE and the other one from the MLE based on the fully observed

**TABLE 1** Simulation results

	Bias					s.d.					RMSE					95%CP					s.e.						
	$\alpha_0$	$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$	$\alpha_0$	$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$	$\alpha_0$	$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$	$\alpha_0$	$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$	$\alpha_0$	$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$		
Simulation one																											
(i) The missingness only depends on the fully observed variables, and $\alpha^* = (-0.5046, -0.4961, 0.4981, -0.0010, 0.0042)$																											
Full	2	5	0	5	1	72	62	59	60	144	72	62	59	60	144	94	96	95	95	95	71	64	60	60	60	147	
SPMLE	0	6	3	10	5	70	62	52	53	140	70	62	52	54	140	95	96	98	97	97	71	64	60	60	60	147	
(ii) The missingness depends on full observed and partially observed variables, and $\alpha^* = (-0.4843, -0.4898, 0.4865, 0.2435, 0.2455)$																											
Full	1	8	2	2	5	70	67	64	64	141	70	67	64	64	141	95	95	94	93	96	72	65	62	62	62	146	
SPMLE	7	18	17	13	30	71	67	58	59	142	72	70	60	60	145	96	93	96	95	96	72	65	62	62	62	146	
(iii) The missingness only depends on the partially observed variable, and $\alpha^* = (-0.4868, -0.0034, -0.0022, 0.2466, 0.2441)$																											
Full	4	3	2	1	4	68	58	60	58	136	68	58	60	58	136	96	95	95	96	95	70	60	59	59	59	139	
SPMLE	18	6	20	22	28	69	59	55	51	134	71	59	59	56	137	95	95	96	96	95	71	60	59	59	59	140	
Simulation two																											
(i) $\gamma_2 = (-0.8, 0.2, 0)$ , and $\alpha^* = (-0.0982, -0.1917, -0.3078, -0.3009, -0.0027)$																											
Full	6	14	8	13	1	115	250	171	122	106	115	250	171	122	106	94	95	95	95	96	111	250	174	122	109		
SPMLE	2	36	1	2	11	110	222	151	105	91	110	225	151	105	92	96	97	97	98	112	247	174	122	109			
(ii) $\gamma_2 = (-0.8, 0.2, -0.2)$ , and $\alpha^* = (-0.1064, -0.2097, -0.2874, -0.2936, 0.1804)$																											
Full	7	17	7	0	4	110	252	166	120	106	110	253	166	120	107	95	94	96	96	110	247	172	121	108			
SPMLE	1	3	9	7	25	107	218	150	106	95	107	218	150	106	99	96	97	98	97	111	244	171	121	108			
(iii) $\gamma_2 = (-0.8, 0.2, 0.5)$ , and $\alpha^* = (-0.0473, -0.6289, -0.2229, -0.3409, -0.3271)$																											
Full	5	20	9	1	11	117	264	190	133	121	117	265	190	133	121	95	95	94	94	115	262	183	128	120			
SPMLE	24	41	22	26	61	113	215	157	113	98	115	219	159	116	115	96	98	98	96	117	260	182	128	119			

Note: Entries of absolute value of bias (bias), empirical standard deviation (s.d.), square root of MSE (RMSE) and standard error (s.e.) are multiplied by 1000, and coverage rates of 95% confidence interval (95%CP) are multiplied by 100. The  $\alpha^*$ 's are the true values of  $\alpha$  computed from the large samples.

Abbreviation: SPMLE, semiparametric maximum likelihood estimate.

	MLE	SPMLE (1)	SPMLE (2)
	<i>Est.(s.e.)</i>	<i>Est.(s.e.)</i>	<i>Est.(s.e.)</i>
(Intercept)	0.513 (0.081)	0.320 (0.084)	0.370 (0.084)
Sigdz	0.157 (0.087)	0.585 (0.095)	0.503 (0.094)
Age	-0.518 (0.041)	-0.604 (0.044)	-0.594 (0.044)
Cholesterol	-	-0.646 (0.037)	-0.628 (0.037)
log(1+cad.dur)	0.279 (0.037)	0.369 (0.039)	0.351 (0.039)
Gender	0.036 (0.085)	0.511 (0.096)	0.420 (0.094)

Abbreviations: MLE, maximum likelihood estimate; SPMLE, semiparametric maximum likelihood estimate.

**TABLE 2** Cardiac catheterization coronary artery disease data

variables only.<sup>8</sup> We observe that the IPW estimates using weights from MLE are significantly biased with  $|\text{bias}| \approx 0.1$  and  $s.d. \approx 0.04$  in setting (ii) and (iii), while the IPW estimates with weights based on SPMLE are not biased and more efficient with  $|\text{bias}| < 0.01$  and  $s.d. \approx 0.03$  in all the settings as we expected.

## 5 | EXAMPLE

We consider the Duke cardiac catheterization coronary artery disease diagnostic dataset (available at Vanderbilt Biostatistics Wiki), which has been analyzed under the assumption that the data are MAR.<sup>16</sup> There are five variables in the dataset: gender, age, duration of symptoms of coronary artery disease (cad.dur), cholesterol level, and an indicator of significant coronary disease of cardiac cath (sigdz). Among these variables cholesterol level is missing for 1246 patients and the rest variables are fully observed for all the 3504 patients. To estimate the missing data probability we consider the logistic regression model in (16), where continuous variables are standardized.<sup>16</sup>

$$\begin{aligned} & \text{logit}\{Pr(R = 1 | \text{sigdz, age, cholesterol, cad.dur, gender})\} \\ & = \alpha_0 + \alpha_1 \text{sigdz} + \alpha_2 \text{age} + \alpha_3 \text{cholesterol} + \alpha_4 \log(1 + \text{cad.dur}) + \alpha_5 \text{gender}. \end{aligned} \quad (16)$$

Using the notation in Section 2 this dataset has one missing data group, that is,  $K = 1$ , and  $U = \text{cholesterol}$ ,  $Z = (\text{sigdz, age, cad.dur, gender})$ . We compute two SPMLE's of  $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$ . In SPMLE (1), we let  $H = h(Z)$  have 16 categories where we divide each continuous variable into two categories such that there is approximately the same number of observations in each category, and in SPMLE (2) we let  $H = h(Z)$  have 36 categories where we divide each continuous variable into three categories. For the purpose of comparison, we also compute the MLE's based on the fully observed the variables. The results reported in Table 2 indicate that (i) SPMLE (1) and SPMLE (2) are very close and SPMLE (2) is slightly more efficient compared with SPMLE (1), (ii) the partially observed variable, cholesterol level, is a significant predictor for the missing data probability, and (iii) the SPMLE's of the sigdz effect and the gender effect are significantly different from those of the MLE's. In the simple monotone missing data pattern the missingness does not depend on the variables with missing values given the fully observed variables if the data are MAR.<sup>6</sup> From the results of our analysis, we see that the variable with missing values, cholesterol level, has significant effect on the missing data mechanism given other fully observed variables in the model. Therefore, the example data are not MAR.

## 6 | DISCUSSION

This research proposes a semiparametric likelihood method for statistical inference for missing data mechanisms with arbitrary nonmonotone missing data patterns. It can estimate the effects of both the fully observed variables and the partially observed variables in a parametric regression model for the missing data mechanism, which can be highly significant especially for nonmonotone missing data patterns as discussed by many researchers.<sup>8</sup> The method uses a piecewise empirical distribution to model the conditional distribution of the missing variables given the observed variables for each missing data pattern in the likelihood, which is easily implemented in the EM algorithm with closed form expressions for both E-step and M-step. The empirical distributions depend on complete observations to estimate



the conditional distributions of the missing variables given the observed variables for each missing data pattern. It may force some restrictions on the MAR assumption. An alternative more flexible product piecewise empirical distribution for the covariates, which uses all the observed data to model the joint distribution of the covariates for all the missing data patterns simultaneously, is under investigation.

We estimate the asymptotic variance through the profile score function based on the profile likelihood principle. As we noted that pseudo-likelihoods also belong to composite likelihoods.<sup>12</sup> The Godambe information<sup>12,17</sup> and bootstrap method can also be used for estimating the asymptotic variances, which may require further investigations for recommending better methods for practice.

As far as I know there is no method available for directly estimating logistic regression model of missing data mechanism through splines or other machine learning methods for nonmonotone MAR data. Investigations on parametric and nonparametric models for missing data mechanism based on kernel methods, splines and other machine learning methods are valuable.

## ACKNOWLEDGEMENTS

The author would like to thank the editor and the two referees for providing detailed comments on the article which are very helpful to improve the presentation of the article. This research was partially supported by grant from the Natural Sciences and Engineering Research Council of Canada (YZ).

## DATA AVAILABILITY STATEMENT

Dataset is available at Vanderbilt Biostatistics Wiki.

## ORCID

Yang Zhao  <https://orcid.org/0000-0002-9768-7683>

## REFERENCES

1. Zhao LP, Lipsitz S. Designs and analysis of two-stage studies. *Stat Med*. 1992;11:769-782.
2. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics 3rd. New York, NY: John Wiley and Sons; 2019.
3. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592.
4. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89:846-866.
5. Tsiatis A. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. New York, NY: Springer-Verlag New York Inc.; 2006.
6. Sun B, Tchetgen EJT. On inverse probability weighting for nonmonotone missing at random data. *J Am Stat Assoc*. 2018;113:369-379.
7. Robins JM, Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. *Stat Med*. 1997;16:39-56.
8. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: a comparative review. *J Am Stat Assoc*. 2005;100:332-346.
9. Zhao Y, Lawless JF, McLeish DL. Likelihood methods for regression models with expensive variables missing by design. *Biom J*. 2009;51:123-136.
10. Zhao Y. Semiparametric model for regression analysis with nonmonotone missing data. *Stat Methods Appl*. 2020. <https://doi.org/10.1007/s10260-020-00530-w> [Epub ahead of print].
11. Zhao Y. Regression analysis with covariates missing at random: a piece-wise nonparametric model for missing covariates. *Commun Stat Theory Methods*. 2009;38:3736-3744.
12. Varin C, Reid N, Firth D. An overview of composite likelihood methods. *Stat Sin*. 2011;21:5-42.
13. Ibrahim JG. Incomplete data in generalized linear models. *J Am Stat Assoc*. 1990;85:765-769.
14. Lipsitz SR, Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. *Biometrika*. 1996;83(4):916-922.
15. Murphy SA, van der Vaart AW. On profile likelihood. *J Am Stat Assoc*. 2000;95:449-465.
16. Tomita H, Fujisawa H, Henmi M. A bias-corrected estimator in multiple imputation for missing data. *Stat Med*. 2018;37:3373-3386.
17. Godambe V. An optimum property of regular maximum likelihood estimation. *Ann Math Stat*. 1960;31:1208-1211.
18. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*. 1952;47:663-685.

**How to cite this article:** Zhao Y. Statistical inference for missing data mechanisms. *Statistics in Medicine*. 2020;39:4325-4333. <https://doi.org/10.1002/sim.8727>