entropy

MDPI

Article

A Nuisance-Free Inference Procedure Accounting for the Unknown Missingness with Application to Electronic Health Records

Jiwei Zhao [1,*] and Chi Chen [2]

[1] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53726, USA

[2] Novartis Institutes for Biomedical Research, Shanghai 201203, China; chi-2.chen@novartis.com

* Correspondence: jiwei.zhao@wisc.edu

Received: 24 August 2020; Accepted: 12 October 2020; Published: 14 October 2020

Abstract: We study how to conduct statistical inference in a regression model where the outcome variable is prone to missing values and the missingness mechanism is unknown. The model we consider might be a traditional setting or a modern high-dimensional setting where the sparsity assumption is usually imposed and the regularization technique is popularly used. Motivated by the fact that the missingness mechanism, albeit usually treated as a nuisance, is difficult to specify correctly, we adopt the conditional likelihood approach so that the nuisance can be completely ignored throughout our procedure. We establish the asymptotic theory of the proposed estimator and develop an easy-to-implement algorithm via some data manipulation strategy. In particular, under the high-dimensional setting where regularization is needed, we propose a data perturbation method for the post-selection inference. The proposed methodology is especially appealing when the true missingness mechanism tends to be missing not at random, e.g., patient reported outcomes or real world data such as electronic health records. The performance of the proposed method is evaluated by comprehensive simulation experiments as well as a study of the albumin level in the MIMIC-III database.

Keywords: nuisance; post-selection inference; missingness mechanism; regularization; asymptotic theory; unconventional likelihood

1. Introduction

A major step towards scientific discovery is to identify useful associations from various features and to quantify their uncertainties. This usually warrants building a regression model for an outcome variable and estimating the coefficient associated with each feature as well as the precision of the estimator. Besides the traditional regression with a small dimensionality, with advances in biotechnology, the modern high-dimensional regression usually posits a sparse parameter in the model, and then applies regularization to select the significant features in order to recover the sparsity. In particular, the post-selection inference could be challenging in a regularized regression framework. In this paper, our main interest is to consider a regression model where the outcome variable is prone to missing values. We study both the traditional setting where regularization is not needed and the modern one with regularization.

The missing data issue is an inevitable concern for statistical analysis in various disciplines ranging from biomedical studies to social sciences. In many applications, the occurrence of missing data is usually not the investigator's primary interest but complicates the statistical analysis. The validity of any method devised for missing data heavily depends on the assumption of the missingness mechanism [1]. Unfortunately, those assumptions are largely unknown and difficult, if not infeasible, to be empirically tested. Therefore, one prefers to concentrate on analyzing the regression model for the outcome variable, while treating the mechanism model as a nuisance. A flexible assumption imposed at the minimum level on the mechanism would provide protection against model misspecification at this level.

While it is indeed promising to regard the missingness mechanism as a nuisance with a flexible assumption, a potential issue is the model identifiability problem if the mechanism contains missing-not-at-random cases, i.e., allowing the mechanism to depend on the missing values themselves. In the past few years, researchers have made great progress on this topic by introducing a so-called instrument. This instrument could be a shadow variable [2–7] or an instrumental variable [8,9]. Both approaches are reasonable and are suitable for different applications. In this paper, we adopt the shadow variable approach as it facilitates the interpretability of the regression model for the outcome. The details of the shadow variable approach will be articulated later throughout the paper.

Therefore, we proceed with a semiparametric framework where our primary interest is a parametric regression, e.g., a linear model, where the statistical task is to estimate the parameter of interest and conduct statistical inference (particularly post-selection inference for the setting with regularization). For the nuisance missingness mechanism, we only impose a nonparametric assumption without specifying a concrete form. We encode the shadow variable as Z, which is one component of the covariate $\mathbf{X}$. In general, a shadow variable with a smaller dimensionality allows more flexibility of the missingness mechanism. Therefore, although it could be multidimensional, we only consider univariate Z throughout the paper. With all of these ingredients, we analyze a conditional likelihood approach which will eventually result in a nuisance-free procedure for parameter estimation and statistical inference.

There are at least two extra highlights of our proposed method that are worth mentioning. The first pertains to the algorithm and computation. Although it looks complicated at first sight, we show that, via some data manipulation strategy, the conditional likelihood function can be analytically written as the likelihood of a conventional logistic regression with some prespecified format. Therefore, our objective function can be readily optimized by many existing software packages. This greatly alleviates the computational burden of our procedure. Second, while the variance estimation under the traditional setting is straightforward following the asymptotic approximation, it is challenging for the setting with regularization. To resolve this problem, we present an easy-to-implement data-driven method to estimate the variance of the regularized estimator via a data perturbation technique. It is noted that the current literature on the inference procedure for regularized estimation in the presence of missing values is very scarce. The authors of [10–12] all considered the model selection problem under high dimensionality with missing data; however, none of them studied the post-selection inference in this context.

The remainder of the paper is structured as follows. In Section 2, we first layout our model formulation and introduce the shadow variable and the conditional likelihood. Section 3 details the traditional setting without regularization. We present our algorithm of how to maximize the conditional likelihood function, the theory of how to derive the asymptotic representation of our proposed estimator and how to estimate its variance. In Section 4, we devote ourselves to the modern setting where the sparsity assumption is imposed and the regularization technique is adopted. Both algorithm and theory as well as the variance estimation through the data perturbation technique are presented. In Section 5, we conduct comprehensive simulation studies to examine the finite sample performance of our proposed estimator as well as the comparison to some existing methods. Section 6 is the application of our method to the regression model for the albumin level which suffers from a large amount of missing values in the MIMIC-III study [13]. The paper is concluded with a discussion in Section 7.

2. Methodology

Denote the outcome variable as Y and covariate $\mathbf{X}$. We assume $\mathbf{X} \quad \mathbf{U}^{\mathrm{T}}, Z\ ^{\mathrm{T}}$ where $\mathbf{U}$ is p-dimensional and Z univariate, with detailed interpretation later. We consider the linear model

$$Y \qquad \boldsymbol{\beta}^{\mathrm{T}}\mathbf{U} \quad Z \quad , \tag{1}$$

where $\boldsymbol{\beta}$ is also p-dimensional, and are scalars and the true value of , $_0$, is nonzero, $N\ 0,\ ^2$. We consider the situation that Y has missing values while $\mathbf{X}$ is fully observed. We introduce a binary variable R to indicate missingness: $R \quad 1$ if Y is observed and $R \quad 0$ if missing. To allow the greatest flexibility of the missingness mechanism model, we assume

$$\mathrm{pr}\ R \quad 1 \quad Y, \mathbf{X} \quad \mathrm{pr}\ R \quad 1 \quad Y, \mathbf{U} \quad s\ Y, \mathbf{U}\ , \tag{2}$$

where s merely represents an unknown and unspecified function not depending on Z. We reiterate that, as the assumption (2), in a nonparametric flavor, does not specify a concrete form of s , one does not need to be worrisome of the mechanism model misspecification. Moreover, as it allows the dependence on Y, besides missing-completely-at-random (MCAR) and many scenarios of missing-at-random (MAR), the assumption (2) also contains various situations of missing-not-at-random (MNAR).

We term Z the shadow variable following the works in [5–7,14]. Its existence depends on whether it is sensible that Z and R are conditionally independent (given Y and $\mathbf{U}$) and that Y heavily relies on Z (as $_0 \quad 0$). There are many examples in the literature documenting that the existence of Z is practically reasonable. In application, a surrogate or a proxy of the outcome variable Y, which would not synchronically affect the missingness mechanism, could be a good choice for the shadow variable Z.

We assume independent and identically distributed observations $r_i, y_i, \mathbf{u}_i, z_i$ for $i \quad 1, ..., N$ and the first n subjects are free of missing data. Now we present a s -free procedure via the use of the conditional likelihood. Denote $\mathbf{V} \quad Y, \mathbf{U}^{\mathrm{T}}\ ^{\mathrm{T}}$. We start with

$$\prod_{i\ 1}^{n} p\ \mathbf{v}_i \quad z_i, r_i \quad 1 \qquad \prod_{i\ 1}^{n} \frac{s\ \mathbf{v}_i}{g\ z_i} p\ \mathbf{v}_i \quad z_i\ ,$$

where $g\ z_i \quad \mathrm{pr}\ r_i \quad 1 \quad z_i \quad \mathrm{pr}\ r_i \quad 1 \quad \mathbf{v}\ p\ \mathbf{v} \quad z_i\ d\mathbf{v}$ and p is a generic notation for conditional probability density/mass function. If $\mathbf{V}$ were univariate, we denote as the rank statistic of $v_1, ..., v_n$, then

$$\begin{aligned} &\prod_{i\ 1}^{n} p\ v_i \quad z_i, r_i \quad 1 \qquad p\ v_1, ..., v_n \quad z_1, ..., z_n, r_1 \qquad r_n \quad 1 \\ &p \qquad v_{\ 1}, ..., v_{\ n}, z_1, ..., z_n, r_1 \qquad r_n \quad 1\ p\ v_{\ 1}, ..., v_{\ n} \quad z_1, ..., z_n, r_1 \qquad r_n \quad 1\ . \end{aligned} \tag{3}$$

The conditional likelihood that we use, the first term on the right hand side of (3), is exactly

$$\begin{aligned} &p \qquad v_{\ 1}, ..., v_{\ n}, z_1, ..., z_n, r_1 \qquad r_n \quad 1 \qquad \frac{p\ v_1, ..., v_n \quad z_1, ..., z_n, r_1 \qquad r_n \quad 1}{p\ v_{\ 1}, ..., v_{\ n} \quad z_1, ..., z_n, r_1 \qquad r_n \quad 1} \\ &\frac{\prod_{i\ 1}^{n} p\ v_i \quad z_i, r_i \quad 1}{\Sigma \quad \Omega \prod_{i\ 1}^{n} p\ v \quad _i \quad z_i, r_i \quad 1} \qquad \frac{\prod_{i\ 1}^{n} p\ v_i \quad z_i}{\Sigma \quad \Omega \prod_{i\ 1}^{n} p\ v \quad _i \quad z_i}, \end{aligned} \tag{4}$$

where Ω represents the collection of all one-to-one mappings from $1, ..., n$ to $1, ..., n$. Now (4) is nuisance-free and can be used to estimate the unknown parameters in $p\ v_i \quad z_i$.

Although $\mathbf{V}$ is multidimensional in our case, the idea presented above can still be applied and it leads to

$$\frac{\prod_{i\ 1}^{n} p\ y_i, \mathbf{u}_i \quad z_i, r_i \quad 1}{\Sigma \quad \Omega \prod_{i\ 1}^{n} p\ y \quad _i, \mathbf{u} \quad _i \quad z_i, r_i \quad 1} \qquad \frac{\prod_{i\ 1}^{n} p\ y_i, \mathbf{u}_i \quad z_i}{\Sigma \quad \Omega \prod_{i\ 1}^{n} p\ y \quad _i, \mathbf{u} \quad _i \quad z_i}. \tag{5}$$

Furthermore, to simplify the computation, we adopt the pairwise fashion of (5) following the previous discussion on pairwise pseudo-likelihood in [15], which results

$$\prod_{1\ \ i\ \ j\ \ n} \frac{p\ \ y_i, \mathbf{u}_i\ \ \ z_i\ \ p\ \ y_j, \mathbf{u}_j\ \ \ z_j}{p\ \ y_i, \mathbf{u}_i\ \ \ z_i\ \ p\ \ y_j, \mathbf{u}_j\ \ \ z_j\ \ \ \ \ p\ \ y_i, \mathbf{u}_i\ \ \ z_j\ \ p\ \ y_j, \mathbf{u}_j\ \ \ z_i}.$$

After plugging in model (1) and some algebra, the objective eventually becomes to minimize

$$L\ \ \boldsymbol{\theta}\qquad \begin{matrix}N\\2\end{matrix}^{\ 1} \sum_{1\ \ i\ \ j\ \ N}\ \ _{ij}\ \ \boldsymbol{\theta}\qquad \begin{matrix}N\\2\end{matrix}^{\ 1} \sum_{1\ \ i\ \ j\ \ N} r_i r_j \log\ \ 1\quad W_{ij} \exp\ \ \boldsymbol{\theta}^{\mathrm{T}} \mathbf{d}_{ij}\quad , \tag{6}$$

where $\boldsymbol{\theta}\quad\ \ ,\boldsymbol{\beta}^{\mathrm{T}}\ ^{\mathrm{T}},\quad\ \ /\ ^{2}, \boldsymbol{\beta}\quad\ \boldsymbol{\beta}, \mathbf{d}_{ij}\quad\ \ y_{i\ j} z_{i\ j}, \mathbf{u}_{i\ j}^{\mathrm{T}} z_{i\ j}\ ^{\mathrm{T}}, y_{i\ j}\quad y_i\quad y_j, \mathbf{u}_{i\ j}\quad \mathbf{u}_i\quad \mathbf{u}_j$, $z_{i\ j}\quad z_i\quad z_j$ and $W_{ij}\quad p\ \ z_i\ \ \mathbf{u}_j\ \ p\ \ z_j\ \ \mathbf{u}_i\ \ /\ \ p\ \ z_i\ \ \mathbf{u}_i\ \ p\ \ z_j\ \ \mathbf{u}_j\quad .$

Denote the minimizer of (6) as $\boldsymbol{\theta}$. By checking that

$$\frac{^{2}\ \ _{ij}\ \ \boldsymbol{\theta}}{\boldsymbol{\theta}\ \ \boldsymbol{\theta}^{\mathrm{T}}}\qquad r_i r_j\ \ 1\quad W_{ij} \exp\ \ \boldsymbol{\theta}^{\mathrm{T}} \mathbf{d}_{ij}\quad\ ^{2} W_{ij} \exp\ \ \boldsymbol{\theta}^{\mathrm{T}} \mathbf{d}_{ij}\ \ \mathbf{d}_{ij} \mathbf{d}_{ij}^{\mathrm{T}}$$

is positive definite, $\boldsymbol{\theta}$ uniquely exists. To compute $\boldsymbol{\theta}$, one also needs a model for W_{ij}. Fortunately, this model only depends on fully observed data $\mathbf{x}_i$ and $\mathbf{x}_j$. Essentially any existing parametric, semiparametric, or nonparametric modeling technique for $p\ \ z\quad \mathbf{u}$ can be used, and W_{ij} can be estimated accordingly. Throughout, we denote W_{ij} as an available well-behaved estimator of W_{ij}. Although our procedure stems from $p\ \ y, \mathbf{u}\quad z, r\quad 1$, which only relies on the data $y_i, \mathbf{x}_i$ with $i\quad 1$, it can be seen that, not only the data $y_i, \mathbf{x}_i$ with $i\quad 1$ are used to compute $\boldsymbol{\theta}$, the data $\mathbf{x}_i$ with $i\quad 0$ are also used in the process of estimating W_{ij}. Therefore, all observed data, both from completely-observed subjects and from partially-observed subjects, are utilized in our procedure.

One can notice that, due to the assumption (2) which allows the greatest flexibility of the mechanism model and the adoption of the conditional likelihood, not all parameters $\ ,\boldsymbol{\beta},\ $, and $\ ^{2}$ are estimable. Nevertheless, the parameter $\boldsymbol{\beta}$, which quantifies the association between Y and $\mathbf{U}$ after adjusting for Z and is of primarily scientific interest, can be fully estimable. The remainder of the paper focuses on the estimation and inference of $\boldsymbol{\beta}$, as well as the variable selection procedure based on $\boldsymbol{\beta}$.

Before moving on, we give some comparison with the existing literature to underline the novel contributions we make in this paper. Based on a slightly different but more restrictive missingness mechanism assumption that pr $R\quad 1\quad Y, \mathbf{X}\quad\ a\ \ Y\ \ b\ \ \mathbf{X}$, Refs. [16–18] used the similar idea to analyze non-ignorable missing data for a generalized linear model and a semiparametric proportional likelihood ratio model, respectively. They focused on different aspects of how to use the conditional likelihoods and their consequences such as the partial identifiability issue and the large bias issue. In this paper, we focus on the linear model (1) and we just showed that the parameter $\boldsymbol{\beta}$ is fully identifiable. It can be seen that the method presented in this paper can be applied to different models, but their identifiability problems or some other relevant issues have to be analyzed on a case-by-case basis. For instance, Ref. [19] studied the parameter estimation problem in a logistic regression model with a low dimensionality under assumption (2). They showed that, different from the current paper, all the unknown parameters are identifiable in their context. However, because of the complexity of their objective function, the algorithm studied in [19] is trivial and cannot be extended to a high dimensional setting.

3. Traditional Setting without Regularization

Computation. Directly minimizing $L\ \ \boldsymbol{\theta}$ is feasible; however, it is very computationally involved. From rearranging the terms in $L\ \ \boldsymbol{\theta}$, we realize that it can be rewritten as the negative log-likelihood function of a standard logistic regression model. To be more specific, let k be the index of pair $\ i, j$ with $k \quad 1, ..., K$ and $K \quad {}^{n}_{2}$. Then,

$$L\ \boldsymbol{\theta} \qquad \frac{1}{K}\sum_{k\ \ 1}^{K} \log \ \ 1 \quad \exp \ \ s_k\boldsymbol{\theta}^{\mathrm{T}}\mathbf{t}_k \quad \log W_k \quad , \tag{7}$$

where $s_k \quad \ \ \mathrm{sign}\ z_{i\ j}$, $\mathbf{t}_k \quad \ z_{i\ j}\, y_{i\ j}, \ \ z_{i\ j}\,\mathbf{u}_{i\ j}^{\mathrm{T}}\ {}^{\mathrm{T}}$. Denote $g_k \quad I\ z_{i\ j} \quad 0$, then one can show that the summand in (7), $\log \ \ 1 \quad \exp \ \ s_k\boldsymbol{\theta}^{\mathrm{T}}\mathbf{t}_k \quad \log W_k$, equals,

$$g_k \ \ \boldsymbol{\theta}^{\mathrm{T}}\mathbf{t}_k \quad s_k \log W_k \qquad \log \ \ 1 \quad \exp \ \ \boldsymbol{\theta}^{\mathrm{T}}\mathbf{t}_k \quad s_k \log W_k \qquad ,$$

which is the contribution of the k-th subject to the negative log-likelihood of a logistic regression with g_k as the response, $\boldsymbol{\theta}$ as the coefficient, $\mathbf{t}_k$ as the covariate, and $s_k \log W_k$ as the offset term, but without an intercept. Therefore, $\boldsymbol{\theta}$ can be obtained by fitting the aforementioned logistic regression model. Algorithm 1 describes the steps for data manipulation and model fitting to estimate $\boldsymbol{\theta}$ under this traditional setting.

Algorithm 1 Minimization of (6) without penalization

```
Inputs: $\ y_i, \mathbf{u}_i, z_i \ , \ y_j, \mathbf{u}_j, z_j \ , W_{ij}$, for $i \quad 1, ..., n$ and $j \quad 1, ..., n$
Initialize: $k \quad 0$
for $j \quad 2 : n$ do
    for $i \quad 1 : \ j \quad 1$ do
        $k \quad k \quad 1$
        $y_{i\ j} \quad y_i \quad y_j, \mathbf{u}_{i\ j} \quad \mathbf{u}_i \quad \mathbf{u}_j, z_{i\ j} \quad z_i \quad z_j, W_k \quad W_{ij}$
        $g_k \quad I\ z_{i\ j} \quad 0$
        $s_k \quad \mathrm{sign}\ z_{i\ j}$
        $\mathbf{t}_k \quad z_{i\ j}\, y_{i\ j}, \ \ z_{i\ j}\,\mathbf{u}_{i\ j}^{\mathrm{T}}\ {}^{\mathrm{T}}$
Fit logistic regression with response $\mathbf{g}$, covariate $\mathbf{t}$, offset $\mathbf{s}^{\mathrm{T}} \log \mathbf{W}$, and no intercept.
Outputs: $\boldsymbol{\theta}$
```

Asymptotic Theory. The asymptotic theory of $\boldsymbol{\theta}$ involves a model of $p\ z \quad \mathbf{u}$, which does not contain any missing values, and therefore any statistical model, either parametric, or semiparametric, or nonparametric, can be used. For simplicity, we only discuss the parametric case here, and any further elaborations will be rendered into Section 7. For a parametric model $p\ z \quad \mathbf{u}; \boldsymbol{\eta}$, one can apply the standard maximum likelihood estimate $\boldsymbol{\eta}$. Here, we simply assume

$$\overline{N}\ \ \boldsymbol{\eta} \quad \boldsymbol{\eta}_0 \qquad \mathbf{G}^{\ \ 1}\ \overline{N}\frac{1}{N}\sum_{i\ \ 1}^{N}\frac{}{\boldsymbol{\eta}}\log \ \ p\ z_i \quad \mathbf{u}_i; \boldsymbol{\eta}_0 \qquad o_p\ 1\ , \tag{8}$$

where $\mathbf{G} \quad \mathrm{E}\ \ \frac{2}{\boldsymbol{\eta}\ \boldsymbol{\eta}^{\mathrm{T}}}\log \ \ p\ z \quad \mathbf{u}; \boldsymbol{\eta}_0$, $\mathrm{E}\ \ \frac{2}{\boldsymbol{\eta}\ \boldsymbol{\eta}^{\mathrm{T}}}\log \ \ p\ z \quad \mathbf{u}; \boldsymbol{\eta}_0 \quad {}^{2} \quad \infty$, $\boldsymbol{\eta}_0$ is the true value of $\boldsymbol{\eta}$, and $\ \mathbf{M} \qquad \overline{\mathrm{trace}\ \mathbf{M}\mathbf{M}^{\mathrm{T}}}$ for a matrix $\mathbf{M}$. With this prerequisite, we have the following result for $\boldsymbol{\theta}$, and its proof is provided in Appendix A.

Theorem 1. *Assume (8) as well as* $\mathrm{E}\ \ \frac{2\ \ {}_{ij}\ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0}{\boldsymbol{\theta}\ \boldsymbol{\theta}^{\mathrm{T}}}\ {}^{2} \quad \infty$. *Denote* $\boldsymbol{\theta}_0$ *the true value of* $\boldsymbol{\theta}$. *Then*

$$\overline{N}\ \ \boldsymbol{\theta} \quad \boldsymbol{\theta}_0 \quad {}^{d} \quad N\ \ \mathbf{0}, \mathbf{A}^{\ \ 1}\boldsymbol{\Sigma}\mathbf{A}^{\ \ 1} \ \ ,$$

where $\mathbf{A} \quad E \quad \frac{^2 \ _{ij} \ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0}{\boldsymbol{\theta} \ \boldsymbol{\theta}^{\mathrm{T}}}$, $\boldsymbol{\Sigma} \quad 4E \quad \boldsymbol{\lambda}_{12} \ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0 \ \boldsymbol{\lambda}_{13} \ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0 \ ^{\mathrm{T}}$, $\boldsymbol{\lambda}_{ij} \ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0 \quad \mathbf{B}\mathbf{G} \ ^1\mathbf{M}_{ij} \ \boldsymbol{\eta}_0 \quad \mathbf{N}_{ij} \ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0$, $\mathbf{B} \quad E \quad \frac{^2 \ _{ij} \ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0}{\boldsymbol{\theta} \ \boldsymbol{\eta}^{\mathrm{T}}}$, $\mathbf{M}_{ij} \ \boldsymbol{\eta}_0 \quad \frac{1}{2} \quad \overline{\boldsymbol{\eta}} \log p \ z_i \quad \mathbf{u}_i; \boldsymbol{\eta}_0 \quad \overline{\boldsymbol{\eta}} \log p \ z_j \quad \mathbf{u}_j; \boldsymbol{\eta}_0$, *and* $\mathbf{N}_{ij} \ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0$ $\frac{_{ij} \ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0}{\boldsymbol{\theta}}$.

If one prefers the asymptotic result of $\boldsymbol{\beta}$, we have

Corollary 1. *Let* $\mathbf{C}$ *be a* $p \quad p \quad 1$ *matrix such that* $\mathbf{C}\boldsymbol{\theta} \quad \boldsymbol{\beta}$*, i.e.,*

$$\mathbf{C} \quad \begin{matrix} 0 & 1/ \ _0 & 0 & & 0 \\ 0 & 0 & 1/ \ _0 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & 1/ \ _0 \end{matrix} \quad .$$

Denote $\boldsymbol{\beta}_0$ *the true value of* $\boldsymbol{\beta}$*. Then, following Theorem 1, we have* $\overline{N} \quad \boldsymbol{\beta} \quad \boldsymbol{\beta}_0 \quad ^d \quad N \quad \mathbf{0}, \mathbf{C}\mathbf{A} \ ^1\boldsymbol{\Sigma}\mathbf{A} \ ^1\mathbf{C}^{\mathrm{T}}$.

Variance Estimation. With Theorem 1 and Corollary 1, the variance estimation is straightforward using the plugging in strategy. Note that var $\boldsymbol{\theta} \quad \frac{1}{N}\mathbf{A} \ ^1\boldsymbol{\Sigma}\mathbf{A} \ ^1$, then one would have the estimate var $\boldsymbol{\theta} \quad \frac{1}{N}\mathbf{A} \ ^1\boldsymbol{\Sigma}\mathbf{A} \ ^1$ where $\mathbf{A} \quad {}^N_2 \quad ^1 \sum_1 \ i \ j \ N \frac{^2 \ _{ij} \ \boldsymbol{\theta}, \boldsymbol{\eta}}{\boldsymbol{\theta} \ \boldsymbol{\theta}^{\mathrm{T}}}$, $\boldsymbol{\Sigma} \quad \frac{4}{N \ 1} \sum_i^N \ _1 \ \frac{1}{N \ 1} \sum_j^N \ _{1,j \ i} \quad \mathbf{B}\mathbf{G} \ ^1\mathbf{M}_{ij} \ \boldsymbol{\eta} \quad \mathbf{N}_{ij} \ \boldsymbol{\theta}, \boldsymbol{\eta} \quad ^2$, $\mathbf{B} \quad {}^N_2 \quad ^1 \sum_1 \ i \ j \ N \frac{^2 \ _{ij} \ \boldsymbol{\theta}, \boldsymbol{\eta}}{\boldsymbol{\theta} \ \boldsymbol{\eta}^{\mathrm{T}}}$, and $\mathbf{G}$ $\frac{1}{N} \sum_i^N \ _1 \frac{^2}{\boldsymbol{\eta} \ \boldsymbol{\eta}^{\mathrm{T}}} \log \ p \ z_i \quad \mathbf{u}_i; \boldsymbol{\eta}$.

4. Modern Setting with Regularization

In the past few decades, it has become a standard practice to consider the high-dimensional regression model, where one assumes the parameter $\boldsymbol{\beta}$ is sparse and often uses the regularization technique to recover the sparsity. While it is a prominent problem to analyze this type of model when the data are prone to missing values, the literature is quite scarce primarily because it is cumbersome to rigorously address the missingness under high dimensionality. Therefore, it is valuable to extend the nuisance-free likelihood procedure proposed in Section 3 to the setting with regularization.

Computation. Regularization is a powerful technique to identify the zero elements of a sparse parameter in a regression model. Various penalty functions have been extensively studied, such as LASSO [20], SCAD [21], and MCP [22]. In particular, we study the adaptive LASSO penalty [23] with the objective of minimizing the following function

$$L \quad \boldsymbol{\theta} \qquad L \ \boldsymbol{\theta} \quad \sum_{j \ 1}^{p} \quad _j \quad ^1 \quad _j \ , \tag{9}$$

where $\quad 0$ is the tuning parameter. Following [23], $\ _j$ is a root-N-consistent estimator of $\ _j$; for example, one can use the estimator via minimizing the unregularized objective Function (6). Obviously, the penalty term in (9) does not alter the numerical characteristic of $L \ \boldsymbol{\theta}$ that we presented in Section 3. The $L \quad \boldsymbol{\theta}$ is essentially the regularized log-likelihood of a logistic regression model with the similar format as discussed in (7).

To choose the tuning parameter $\ $, one can follow either the cross-validation method or various information-based criteria. Fortunately, all of these approaches have been extensively studied in the literature. In this paper, we follow the Bayesian information criterion (BIC) to determine $\ $. Specifically, we choose $\ $ to be the minimizer of the following BIC function

$$\mathrm{BIC} \qquad 2L\ \boldsymbol{\theta} \qquad p \ \frac{\log\ n}{n},$$

where p is the number of nonzero elements in $\boldsymbol{\beta}$ and the minimizer of (9) is encoded as $\boldsymbol{\theta} \qquad , \boldsymbol{\beta}^{\mathrm{T}\ \mathrm{T}}$. We summarize the whole computation pipeline as Algorithm 2 below.

Algorithm 2 Minimization of (9) with the ALASSO penalty

```
Inputs:  $y_i, \mathbf{u}_i, z_i \ , \ y_j, \mathbf{u}_j, z_j \ , W_{ij}$, for $i \quad 1, ..., n$ and $j \quad 1, ..., n$
Initialize:  $k \quad 0$
for $j \quad 2 : n$ do
    for $i \quad 1 : \ j \quad 1$ do
        $k \quad k \quad 1$
        $y_{i\ j} \quad y_i \quad y_j, \mathbf{u}_{i\ j} \quad \mathbf{u}_i \quad \mathbf{u}_j, z_{i\ j} \quad z_i \quad z_j, W_k \quad W_{ij}$
        $g_k \quad I\ \ z_{i\ j} \quad 0$
        $s_k \quad \mathrm{sign}\ z_{i\ j}$
        $\mathbf{t}_k \quad z_{i\ j}\, y_{i\ j}, \ z_{i\ j}\, \mathbf{u}_{i\ j}^{\mathrm{T}\ \ \mathrm{T}}$
Fit logistic regression with response $\mathbf{g}$, covariates $\mathbf{t}$, offset $\mathbf{s}^{\mathrm{T}} \log \mathbf{W}$, and no intercept.
Obtain $\boldsymbol{\theta}$.
Fit logistic regression with ALASSO penalty.
Find      which minimizes the BIC.
Outputs:  $\boldsymbol{\theta} \qquad \boldsymbol{\theta}$
```

Asymptotic Theory. Recall that $\boldsymbol{\theta} \qquad , \boldsymbol{\beta}^{\mathrm{T}\ \mathrm{T}}$. Without loss of generality, we assume the first p_0 parameters in $\boldsymbol{\beta}$ are nonzero, where $1 \quad p_0 \quad p$. For simplicity, we denote $\boldsymbol{\theta}_T \qquad , \ _1, ..., \ _{p_0}{}^{\mathrm{T}}$ as the vector of nonzero components and $\boldsymbol{\theta}_{T^c} \qquad _{p_0}\ _1, ..., \ _p{}^{\mathrm{T}}$ as the vector of zeros.

In Theorem 1, we defined $\mathbf{A} \quad \mathrm{E} \ \frac{^2 \ _{ij}\ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0}{\boldsymbol{\theta}\ \boldsymbol{\theta}^{\mathrm{T}}}$, a $p \quad 1 \qquad p \quad 1$ matrix. Now we assume it can be partitioned as $\mathbf{A} \quad \begin{matrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_2^{\mathrm{T}} & \mathbf{A}_3 \end{matrix}$, where $\mathbf{A}_1$ is a $p_0 \quad 1 \qquad p_0 \quad 1$ submatrix corresponding to $\boldsymbol{\theta}_T$. Similarly, we defined $\boldsymbol{\Sigma} \quad 4\mathrm{E} \ \ \boldsymbol{\lambda}_{12}\ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0 \ \boldsymbol{\lambda}_{13}\ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0 \ ^{\mathrm{T}}$, and we also assume it can be partitioned as $\boldsymbol{\Sigma} \quad \begin{matrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_2 \\ \boldsymbol{\Sigma}_2^{\mathrm{T}} & \boldsymbol{\Sigma}_3 \end{matrix}$, where $\boldsymbol{\Sigma}_1$ is a $p_0 \quad 1 \qquad p_0 \quad 1$ submatrix corresponding to $\boldsymbol{\theta}_T$ as well. We denote the minimizer of (9), $\boldsymbol{\theta}$, as $\boldsymbol{\theta} \qquad \boldsymbol{\theta}^{\mathrm{T}}_{\ ,T}, \boldsymbol{\theta}^{\mathrm{T}}_{\ ,T^c}{}^{\mathrm{T}}$, and its true value $\boldsymbol{\theta}_0 \qquad \boldsymbol{\theta}^{\mathrm{T}}_{0,T}, \boldsymbol{\theta}^{\mathrm{T}}_{0,T^c}{}^{\mathrm{T}}$.

Now, we present the oracle property pertaining to $\boldsymbol{\theta}$, which includes the asymptotic normality for the nonzero components and the variable selection consistency. The proof is provided in Appendix B.

Theorem 2. *Assume (8),* $\mathbf{A}_1$ *is positive definite and* $E \ \frac{_{ij}\ \boldsymbol{\theta}_0, \boldsymbol{\eta}_0}{\boldsymbol{\theta}}^{\ 2} \quad \infty$ *for each* $\boldsymbol{\theta}$ *in a neighborhood of* $\boldsymbol{\theta}_0$*. We also assume* $\overline{N} \qquad 0$ *and* $N \qquad \infty$*. Then,*

$$\overline{N} \quad \boldsymbol{\theta}_{\ ,T} \quad \boldsymbol{\theta}_{0,T} \quad \overset{d}{\ } \quad N \ \ \mathbf{0}, \mathbf{A}_1^{\ 1} \boldsymbol{\Sigma}_1 \mathbf{A}_1^{\ 1} \ \ .$$

In addition, let $T_N \qquad j \qquad 1, ..., p \ : \ \ _{j,} \qquad 0$ *and* $T \qquad j \qquad 1, ..., p \ : \ \ _{j,0} \qquad 0$ *, then*

$$\lim_{N \quad \infty} pr\ \ T_N \qquad T \qquad 1.$$

Variance Estimation. Although the above theory provides a rigorous justification for the asymptotic property of $\boldsymbol{\theta}$, in practice, however, it does not guide the standard error estimation. Here, we propose a data perturbation approach for the variance estimation. Specifically, following [24], we generate a set of independent and identically distributed positive random variables $\Xi \qquad {}_i, i \quad 1, ..., N$ with E ${}_i \quad$ 1 and var ${}_i \quad$ 1, e.g., the standard exponential distribution. Since it is based on a U-statistic structure, we perturb our objective function by adding ${}_{ij} \quad {}_i \; {}_j$ to each of its pairwise terms. We first obtain the estimator $\boldsymbol{\theta}$ by minimizing the perturbed version of (6):

$$L \quad \boldsymbol{\theta} \qquad \begin{matrix} N \\ 2 \end{matrix}^{\;1} \sum_{1 \;\; i \;\; j \;\; N} \; {}_{ij} \; {}_{ij} \; \boldsymbol{\theta} \; .$$

Then, we obtain the estimator $\boldsymbol{\theta}$ by minimizing the perturbed version of (9):

$$L \quad \boldsymbol{\theta} \qquad \begin{matrix} N \\ 2 \end{matrix}^{\;1} \sum_{1 \;\; i \;\; j \;\; N} \; {}_{ij} \; {}_{ij} \; \boldsymbol{\theta} \qquad \sum_{j \;\; 1}^{p} \frac{\quad}{{}_j} \quad {}_j \; ,$$

where the optimal is also computed by the BIC. We repeat this data perturbation scheme a large number of times, say, M.

Following the theory in [25,26], under some regularity conditions, one can first show that $\overline{N} \;\; \boldsymbol{\theta} \;\; {}_{,T} \quad \boldsymbol{\theta}_{0,T}$ converges in distribution to $N \;\; \mathbf{0}, \mathbf{A}_1 \;{}^{1} \boldsymbol{\Sigma}_1 \mathbf{A}_1 \;{}^{1}$, the same limiting distribution of $\overline{N} \;\; \boldsymbol{\theta} \qquad \boldsymbol{\theta}_0$. Furthermore, one can also show pr $\quad \boldsymbol{\theta} \;\; {}_{,T^c} \qquad 0 \qquad$ 1, where pr is the probability measure generated by the original data and the perturbation data Ξ. In addition, one can show that the distribution of $\overline{N} \;\; \boldsymbol{\theta} \;\; {}_{,T} \quad \boldsymbol{\theta} \;\; {}_{,T}$ conditional on the data can be used to approximate the unconditional distribution of $\overline{N} \;\; \boldsymbol{\theta} \;\; {}_{,T} \quad \boldsymbol{\theta}_{0,T}$ and that pr $\quad \boldsymbol{\theta} \;\; {}_{,T^c} \qquad 0 \qquad\qquad$ 1.

To achieve a confidence interval for ${}_j$, the j-th coordinate in $\boldsymbol{\theta}$, the lower and upper bounds can be formed by ${}_{,j, \;\; /2}$ and ${}_{,j,1 \quad /2}$, respectively, where ${}_{,j,q}$ represents the q-th quantile of ${}_{,j,m}, m \quad 1, ..., M$.

5. Simulation Studies

We conduct comprehensive simulation studies to evaluate the finite sample performance of our proposed estimators and also compare with some currently existing methods. We first present the results under the model without regularization, then with regularization.

5.1. Scenarios without Regularization

For the proposed estimator studied in Section 3, we generate $R_i, Y_i, \mathbf{U}_i^{\mathrm{T}}, Z_i \;\; , i \qquad 1, \ldots, N,$ independent and identically distributed copies of $R, Y, \mathbf{U}^{\mathrm{T}}, Z$, as follows. We first generate the random vector $\mathbf{U} \qquad U_1, \ldots, U_p \;\; {}^{\mathrm{T}}$ with $U_i \qquad N \;\; 0.5, 1$ and $p \quad 4$, and then generate $Z \qquad {}_z \quad \boldsymbol{\eta}^{\mathrm{T}} \mathbf{U} \qquad {}_z$ with ${}_z \quad 0.5, \boldsymbol{\eta} \qquad 0.5, 1, \quad 1, 1.5 \;\; {}^{\mathrm{T}}, \;\; {}_z \qquad N \;\; 0, 1$. Afterwards, the outcome variable Y is generated following the model (1) with $\qquad 1, \boldsymbol{\beta} \qquad 0.5, 1, \quad 1, 1.5 \;\; {}^{\mathrm{T}}, \qquad 0.5$, and $\qquad N \;\; 0, 1$, and the missingness indicator R is generated following pr $R \quad 1 \quad Y, \mathbf{U} \qquad I \;\; Y \qquad 2.5, U_1 \qquad 2, U_2 \qquad 2, U_3 \qquad 2, U_4 \qquad 2$ which results in around 40% missing values. We examine two situations with sample size $N \quad 500$ and $N \quad 1000$ respectively. Besides the estimator studied in Section 3 (`Proposed`), we also implement the estimator using all simulated data (`FullData`) and the estimator using completely observed subjects only (`CC`). Based on 1000 simulation replicates, for each of the three estimators, we summarize the sample bias, sample standard deviation, estimated standard error, and coverage probability of 95% confidence intervals in Table 1.

Table 1. In Section 5.1, sample bias (Bias), sample standard deviation (SD), estimated standard error (SE), and coverage probability (CP) of 95% confidence interval of the estimator of FullData (using all simulated data), CC (using only completely observed subjects), and of the proposed estimator studied in Section 3.

N	Parameter	Method	Bias	SD	SE	CP
500		FullData	0.0026	0.0444	0.0450	0.9540
		CC	0.0329	0.0564	0.0560	0.9100
		Proposed	0.0174	0.0829	0.0789	0.9450
	1	FullData	0.0022	0.0489	0.0503	0.9510
		CC	0.0376	0.0670	0.0699	0.9300
		Proposed	0.0164	0.1644	0.1607	0.9400
	2	FullData	0.0017	0.0657	0.0635	0.9310
		CC	0.0649	0.0851	0.0835	0.8680
		Proposed	0.0399	0.2305	0.2239	0.9360
	3	FullData	0.0022	0.0616	0.0635	0.9540
		CC	0.0778	0.0871	0.0867	0.8430
		Proposed	0.0462	0.2323	0.2298	0.9410
	4	FullData	0.0045	0.0792	0.0810	0.9530
		CC	0.0988	0.1007	0.1043	0.8550
		Proposed	0.0672	0.3081	0.3047	0.9380
1000		FullData	0.0012	0.0317	0.0317	0.9540
		CC	0.0348	0.0396	0.0393	0.8510
		Proposed	0.0068	0.0573	0.0555	0.9350
	1	FullData	0.0011	0.0367	0.0355	0.9370
		CC	0.0399	0.0490	0.0494	0.8840
		Proposed	0.0154	0.1154	0.1138	0.9460
	2	Full Data	0.0020	0.0448	0.0448	0.9500
		CC	0.0649	0.0577	0.0588	0.8110
		Proposed	0.0153	0.1531	0.1591	0.9590
	3	Full Data	0.0015	0.0458	0.0449	0.9460
		CC	0.0779	0.0605	0.0611	0.7490
		Proposed	0.0135	0.1598	0.1634	0.9480
	4	Full Data	0.0009	0.0564	0.0571	0.9540
		CC	0.0949	0.0720	0.0734	0.7550
		Proposed	0.0242	0.2091	0.2167	0.9430

Furthermore, we consider a similar simulation setting where the generation is the same as above except for a logistic missingness mechanism model with logit pr R 1 $Y, \mathbf{U}$ 3 $2Y$ $0.5U_1$ U_2 U_3 $1.5U_4$, which also results in around 40% missing values. We replicate the results, shown in Table 2.

We can reach the following conclusions from Tables 1 and 2. For the estimator `Proposed`, although its bias is slightly larger than the benchmark `FullData`, it is still very close to zero. The sample standard deviation and the estimated standard error are rather close to each other. The sample coverage probability of the estimated 95% confidence interval is also very close to the nominal level. This observation well matches our theoretical justification in Theorem 1. On the contrary, the estimator `CC` is clearly biased, resulting in empirical coverage far from the nominal level, and therefore is not recommended to use in practice. It is also clear that, compared to the benchmark `FullData`, the estimator `Proposed` has estimation efficiency loss to some extent. This is because the proposed method uses the conditional likelihood approach and it completely eliminates the effect of the nuisance.

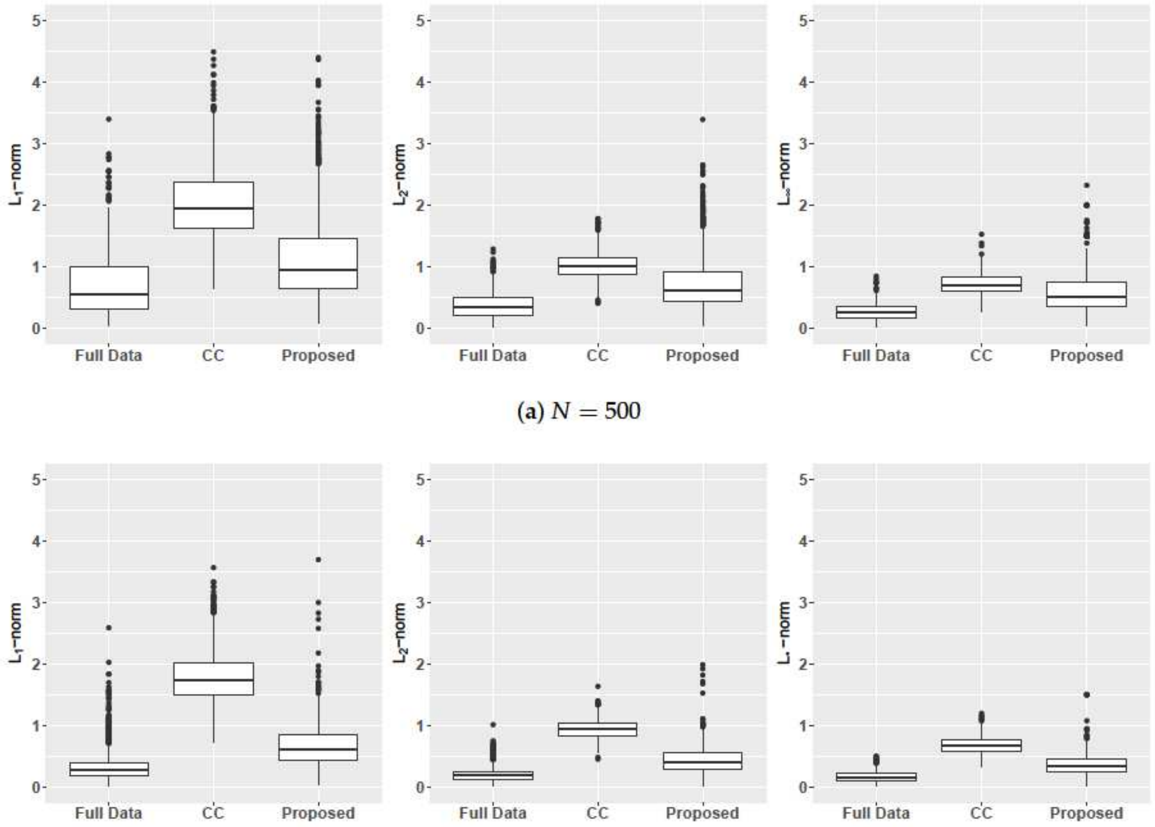
Table 2. In Section 5.1, sample bias (Bias), sample standard deviation (SD), estimated standard error (SE), and coverage probability (CP) of 95% confidence interval of the estimator of FullData (using all simulated data), CC (using only completely observed subjects), and of the proposed estimator studied in Section 3, with a logistic missingness mechanism model.

N	Parameter	Method	Bias	SD	SE	CP
500		FullData	0.0011	0.0464	0.0451	0.9410
		CC	0.0306	0.0567	0.0567	0.9200
		Proposed	0.0100	0.0822	0.0787	0.9380
	1	FullData	0.0004	0.0509	0.0503	0.9520
		CC	0.0440	0.0636	0.0637	0.8930
		Proposed	0.0146	0.1308	0.1236	0.9420
	2	FullData	0.0013	0.0639	0.0637	0.9520
		CC	0.0871	0.0828	0.0821	0.8190
		Proposed	0.0173	0.1824	0.1753	0.9430
	3	FullData	0.0030	0.0655	0.0636	0.9400
		CC	0.0876	0.0847	0.0821	0.8030
		Proposed	0.0214	0.1840	0.1756	0.9440
	4	FullData	0.0023	0.0845	0.0812	0.9390
		CC	0.1307	0.1083	0.1061	0.7560
		Proposed	0.0331	0.2533	0.2384	0.9360
1000		FullData	0.0004	0.0315	0.0317	0.9490
		CC	0.0286	0.0396	0.0398	0.8950
		Proposed	0.0060	0.0568	0.0555	0.9390
	1	FullData	0.0007	0.0362	0.0354	0.9420
		CC	0.0442	0.0451	0.0447	0.8410
		Proposed	0.0079	0.0910	0.0859	0.9290
	2	FullData	0.0004	0.0450	0.0448	0.9390
		CC	0.0879	0.0571	0.0576	0.6640
		Proposed	0.0044	0.1277	0.1220	0.9420
	3	FullData	0.0009	0.0450	0.0448	0.9450
		CC	0.0880	0.0588	0.0577	0.6660
		Proposed	0.0114	0.1309	0.1222	0.9380
	4	FullData	0.0005	0.0576	0.0572	0.9510
		CC	0.1342	0.0755	0.0745	0.5740
		Proposed	0.0191	0.1757	0.1661	0.9370

5.2. Scenarios with Regularization

For the estimator studied in Section 4, the independent and identically distributed samples are generated as follows. The variable $\mathbf{U}$ $U_1, \ldots, U_p$ $^{\mathrm{T}}$ is generated from MVN $\mathbf{0}, \boldsymbol{\Sigma}_u$ with $\boldsymbol{\Sigma}_u$ $0.5^{i\ \ j}$ $_{1\ \ i,j\ \ p}$ and p 8. Then, the shadow variable Z is generated following Z $_z$ $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{U}$ $_z$ with $_z$ 0, $\boldsymbol{\eta}$ 0.5, 0.5, 1, 1, 0.5, 0.5, 1, 1 $^{\mathrm{T}}$ and $_z$ N 0, 1 . The outcome variable Y is generated from model (1) with 0, $\boldsymbol{\beta}$ 3, 1.5, 0, 0, 2, 0, 0, 0 $^{\mathrm{T}}$, 3, N 0, ^{2} and 3. The distribution of the missingness indicator follows from logit pr R 1 $Y, \mathbf{U}$ 5 5Y 0.2U_1 0.2U_7, which results in about 45% missing values. Similar to Section 5.1, we also examine two situations with sample size N 500 and N 1000 respectively, and we implement three estimators `FullData`, `CC`, and `Proposed`. When the estimator `Proposed` is implemented, we perform M 500 perturbations in order to obtain the confidence interval for the unknown parameter. The results summarized below are based on 1000 simulation replicates.

Figure 1 shows the L_1, L_2, and L_∞ norms of the bias for the three different estimators. As sample size increases, there is no doubt that the estimation bias is getting smaller for any method. It is also clear that the bias of the `Proposed` estimator is larger than the benchmark `FullData`, but much smaller than the method `CC`.



(a) $N = 500$

(b) $N = 1000$

Figure 1. In Section 5.2, L_1 (1st column), L_2 (2nd column), and L_∞ (3rd column) norms of the estimation bias of the estimator of FullData (using all simulated data), CC (using only completely observed subjects), and of the proposed estimator studied in Section 4.

We present the statistical inference results in Table 3 for $N = 500$ and Table 4 for $N = 1000$, respectively, including sample bias, sample standard deviation, estimated standard error, coverage probability, and length of 95% confidence interval for the three different methods. For the nonzero β's as well as $\widetilde{\gamma}$, similar to Section 5.1, the method `CC` clearly prompts coverage probability far from the nominal level hence is not reliable. For the method `Proposed`, its estimation bias is quite close to zero, and its sample standard deviation and estimated standard error are quite close to each other. The coverage probability of the confidence interval converges to the nominal level 95% as the sample size gets larger. For the noisy zero β's, the coverage probabilities in the three methods are all close to 1, reflecting the variable selection consistency in the oracle property, even for the `CC` method. Furthermore, a very nice finite sample property of our proposed estimator is that it produces the confidence interval with the shortest length, which can be clearly seen from both Tables 3 and 4.

Table 3. In Section 5.2, with sample size N 500, sample bias (Bias), sample standard deviation (SD), estimated standard error (SE), coverage probability (CP), and length (Length) of 95% confidence interval of the estimator of FullData (using all simulated data), CC (using only completely observed subjects) and of the proposed estimator studied in Section 4.

Parameter		Method	Bias	SD	SE	CP	Length
		FullData	0.0001	0.0120	0.0132	0.9480	0.0515
		CC	0.0729	0.0180	0.0183	0.0370	0.0716
		Proposed	0.0423	0.0500	0.0498	0.8200	0.1926
True Nonzero	1	FullData	0.0021	0.1686	0.1649	0.9400	0.6415
		CC	0.6547	0.2207	0.2114	0.1460	0.8233
		Proposed	0.0354	0.4698	0.4746	0.9320	1.8513
	2	Full Data	0.0275	0.1692	0.1791	0.9440	0.6952
		CC	0.3501	0.2227	0.2174	0.6180	0.8471
		Proposed	0.2654	0.5843	0.5609	0.8940	1.9237
	5	Full Data	0.0172	0.1576	0.1756	0.9650	0.6826
		CC	0.4478	0.2172	0.2161	0.4370	0.8418
		Proposed	0.1251	0.4037	0.4611	0.9330	1.8063
True Zero	3	FullData	0.0085	0.1567	0.1890	0.9960	0.7184
		CC	0.0063	0.2067	0.2304	0.9890	0.8890
		Proposed	0.0109	0.0988	0.1690	1.0000	0.4398
	4	Full Data	0.0019	0.1581	0.1900	0.9940	0.7206
		CC	0.0017	0.2097	0.2307	0.9900	0.8914
		Proposed	0.0126	0.1112	0.1447	1.0000	0.3668
	6	Full Data	0.0045	0.1212	0.1606	0.9980	0.6146
		CC	0.0053	0.1749	0.1953	0.9900	0.7560
		Proposed	0.0034	0.0664	0.1160	1.0000	0.2555
	7	Full Data	0.0014	0.1351	0.1839	0.9980	0.7063
		CC	0.0055	0.1870	0.2245	0.9950	0.8717
		Proposed	0.0024	0.0386	0.1115	1.0000	0.2538
	8	Full Data	0.0072	0.1295	0.1748	0.9990	0.6653
		CC	0.0062	0.1795	0.2125	0.9940	0.8251
		Proposed	0.0016	0.0741	0.1066	1.0000	0.2284

Table 4. In Section 5.2, with sample size N 1000, sample bias (Bias), sample standard derivation (SD), estimated standard error (SE), coverage probability (CP), and length (Length) of 95% confidence interval of the estimator of FullData (using all simulated data), CC (using only completely observed subjects) and of the proposed estimator studied in Section 4.

Parameter		Method	Bias	SD	SE	CP	Length
		FullData	0.0005	0.0073	0.0088	0.9690	0.0344
		CC	0.0730	0.0126	0.0130	0.0000	0.0507
		Proposed	0.0213	0.0311	0.0334	0.8700	0.1293
True Nonzero	1	FullData	0.0005	0.1186	0.1170	0.9300	0.4547
		CC	0.6655	0.1568	0.1507	0.0090	0.5864
		Proposed	0.0211	0.2911	0.2969	0.9300	1.1631
	2	Full Data	0.0321	0.1175	0.1249	0.9550	0.4861
		CC	0.3387	0.1477	0.1534	0.3960	0.5972
		Proposed	0.0979	0.2907	0.3383	0.9230	1.3115
	5	Full Data	0.0225	0.1051	0.1206	0.9590	0.4698
		CC	0.4485	0.1478	0.1534	0.1770	0.5964
		Proposed	0.0621	0.2351	0.2526	0.9290	0.9871

Table 4. *Cont.*

Parameter		Method	Bias	SD	SE	CP	Length
True Zero	3	FullData	0.0007	0.0621	0.1162	1.0000	0.4253
		CC	0.0023	0.1414	0.1614	0.9920	0.6180
		Proposed	0.0044	0.0581	0.0910	1.0000	0.2091
	4	Full Data	0.0020	0.0632	0.1170	1.0000	0.4271
		CC	0.0005	0.1333	0.1608	0.9930	0.6207
		Proposed	0.0063	0.0584	0.0887	1.0000	0.2107
	6	Full Data	0.0013	0.0571	0.1010	1.0000	0.3670
		CC	0.0034	0.1159	0.1378	0.9950	0.5313
		Proposed	0.0012	0.0281	0.0688	1.0000	0.1430
	7	Full Data	0.0028	0.0599	0.1144	1.0000	0.4231
		CC	0.0033	0.1243	0.1584	0.9970	0.6131
		Proposed	0.0016	0.0288	0.0698	1.0000	0.1421
	8	Full Data	0.0039	0.0589	0.1080	1.0000	0.3970
		CC	0.0028	0.1256	0.1497	0.9940	0.5752
		Proposed	0.0000	0.0333	0.0644	1.0000	0.1314

6. Real Data Application

The Medical Information Mart for Intensive Care III (MIMIC-III) is an openly available electronic health records (EHR) database, developed by the MIT Lab for Computational Physiology [13], comprising de-identified health-related data associated with intensive care unit patients with rich information including demographics, vital signs, laboratory test, medications, and more.

Our initial motivation for this data analysis is to understand the missingness mechanism for some laboratory test biomarkers in this EHR system. As for the EHR database, since the data are collected in a non-prescheduled fashion, i.e., only available when the patient seeks care or the physician orders care, the visiting process could be potentially informative about the patients' risk categories. Therefore, it is very plausible that the data are missing not at random, or a mix of missing not at random and missing at random [27,28]. When we first conducted the data cleaning process briefly, an interesting phenomenon we observe is that, compared to most biomarkers which usually have 3% missing values, the albumin level in the blood sample, a very indicative biomarker associated with different types of diseases [29], has around 30% missingness.

To further understand this phenomenon, we concentrate on a subset of the data with sample size N 1359 in which 421 samples have missing values in the albumin level but all other variables are complete. We aim to apply the proposed method to the study of the albumin level (Y). The calcium level in the blood sample, free of missing data, has been shown in the biomedical literature that it has high correlation with the albumin level [30–32]; therefore, we adopt the calcium level as the shadow variable Z. Seventeen other variables comprise the vector $\mathbf{U}$, which are either demographics (age and gender), chart events (respiratory rate, glucose, heart rate, systolic blood pressure, diastolic blood pressure, and temperature), other laboratory tests (urea nitrogen, platelets, magnesium, hematocrit, red blood cell, white blood cell, and peripheral capillary oxygen saturation (SpO2)), or aggregated metrics (simplified acute physiology score (SAPS-II) and sequential organ failure assessment score (SOFA)).

We implement the proposed estimator studied in Section 4 to achieve both variable selection and post-selection inference. We also compare it with the `CC` method which naively fits the regularized linear regression with the ALASSO penalty. For each of the methods, we apply the data perturbation scheme presented in Section 4 with M 500 for standard error estimation. The results are summarized in Table 5. The solution path of the `Proposed` method, as the tuning parameter varies, is also provided in Figure 2.

Table 5. In Section 6, the parameter estimate (Estimate), standard error (SE), and confidence interval (CI) of the estimator of CC (using only completely observed subjects) and of the proposed estimator studied in Section 4 in the MIMIC−III study.

Effect	CC			Proposed		
	Estimate	SE	CI	Estimate	SE	CI
Calcium(shadow)	0.7707	0.0691	[0.6532, 0.9153]	1.5271	0.1796	[1.1815, 1.8835]
Red Blood Cell	0.6491	0.0514	[0.5337, 0.7257]	0.7545	0.1631	[0.3594, 1.0109]
Magnesium	0.0000	0.0686	[−0.2073, 0.0000]	0.2731	0.2452	[0.0000, 0.6609]
SOFA	−0.2720	0.0268	[−0.3135, −0.2099]	−0.1852	0.1040	[−0.3467, 0.0000]
Temperature	−0.0360	0.0351	[−0.0883, 0.0659]	0.0000	0.0964	[0.0000, 0.3132]
White Blood Cell	−0.0245	0.0123	[−0.0416, 0.0000]	0.0000	0.0025	[0.0000, 0.0000]
Age	0.0000	0.0008	[0.0000, 0.0000]	0.0000	0.0017	[0.0000. 0.0000]
Gender	0.0000	0.0240	[−0.0477, 0.0662]	0.0000	0.1320	[−0.4025, 0.0000]
Respiratory Rate	0.0000	0.0034	[−0.0141, 0.0000]	0.0000	0.0008	[0.0000, 0.0000]
Glucose	0.0000	0.0000	[0.0000, 0.0000]	0.0000	0.0005	[0.0000, 0.0000]
Heart Rate	0.0000	0.0025	[−0.0091, 0.0000]	0.0000	0.0004	[0.0000, 0.0000]
Systolic BP	0.0000	0.0045	[−0.0139, 0.0000]	0.0000	0.0000	[0.0000, 0.0000]
Diastolic BP	0.0000	0.0072	[0.0000, 0.0223]	0.0000	0.0000	[0.0000, 0.0000]
Urea Nitrogen	0.0000	0.0004	[0.0000, 0.0000]	0.0000	0.0000	[0.0000, 0.0000]
Platelets	0.0000	0.0000	[0.0000, 0.0000]	0.0000	0.0000	[0.0000, 0.0000]
Hematocrit	0.0000	0.0027	[0.0000, 0.0000]	0.0000	0.0000	[0.0000, 0.0000]
SpO2	0.0000	0.0145	[−0.0479, 0.0000]	0.0000	0.0162	[0.0000, 0.0000]
SAPS-II	0.0000	0.0106	[−0.0051, 0.0269]	0.0000	0.0000	[0.0000, 0.0000]

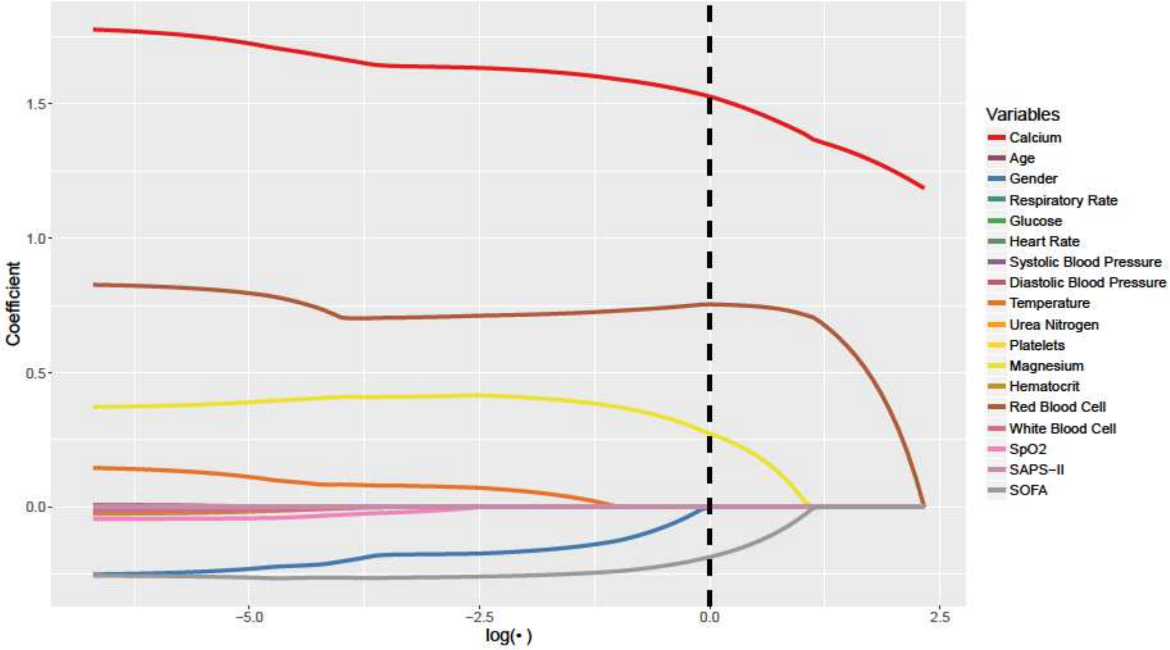


Figure 2. In Section 6, as tuning parameter λ varies, the solution path of the proposed estimator in the MIMIC-III study. The optimal λ, λ^*, equals 1.0030 and $\log\lambda^* = 0.0030$.

In general, both methods achieve the goal of variable selection and post-selection inference by leveraging the regularization technique coupled with the data perturbation strategy, and identify many variables as noise with zero coefficients. In particular, the `Proposed` method provides larger effects for the calcium level (the shadow variable) and the red blood cell count, whereas a smaller effect for the aggregated SOFA score. The `Proposed` method simplifies the body temperature and the white blood cell count as nonsignificant variables, which are identified as nonzero but with a very small effect using the `CC` method. It is also worthwhile to mention that the `Proposed` method signifies the magnesium level with a quite significant coefficient, which was extensively investigated in the scientific literature [33–35].

7. Discussion

In this paper, we provide a systematic approach for parameter estimation and statistical inference in both traditional linear model where the regularization is not needed and the modern regularized regression setting, when the outcome variable is prone to missing values and the missingness mechanism can be arbitrarily flexible. A pivotal condition rooted in our procedure is the shadow variable Z, which overcomes the model identifiability problem and enables the nuisance-free conditional likelihood process.

Certainly any method would have its own limitations and could be potentially improved. One needs a model $p\ z\quad \mathbf{u}$ to implement the proposed estimator in Sections 3 and 4. As its modeling does not involve any missing data, we simply use the parametric maximum likelihood estimation in our algorithm as well as in the theoretical justification. Indeed, any statistical or machine learning method can be used for modeling $p\ z\quad \mathbf{u}$. For instance, if one would like to consider a semiparametric model [36], e.g.,

$$p\ z\quad \mathbf{u};\boldsymbol{\eta},F\qquad \frac{\exp\ \boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}z\ f\ z}{\exp\ \boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}z\ dF\ z},$$

where $\boldsymbol{\eta}\qquad {}_{1},\ldots,\ {}_{p}\ {}^{\mathrm{T}}$ is a vector of unknown parameters and $f\ z$ is the density of an unknown baseline distribution function F with respect to some dominating measure . With this model fitted, W_{ij} can be simplified to $W_{ij}\quad \exp\quad z_i\ {}_j\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}_i\ {}_j$. Therefore, a similar conditional likelihood approach can be used to estimate $\boldsymbol{\eta}$ without estimating the nonparametric component $f\ z$.

Author Contributions: Conceptualization, J.Z.; Experiment, J.Z. and C.C.; Writing, J.Z. and C.C.; Supervision, J.Z. All authors have read and agreed to the published version of the manuscript

Funding: Jiwei Zhao is supported by the National Science Foundation award 1953526.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Theorem 1

Proof. Note that $\boldsymbol{\theta}$ is obtained by setting estimating equation $\frac{L\ \boldsymbol{\theta},\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad 0$, which is equivalent to

$$\frac{L\ \boldsymbol{\theta},\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad \frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}\qquad \frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad \frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}}\qquad \frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}}\quad 0. \tag{A1}$$

Specifically,

$$\frac{L\ \boldsymbol{\theta},\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad \frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad \frac{{}^2L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}\ \boldsymbol{\theta}^{\mathrm{T}}}\ \boldsymbol{\theta}\quad \boldsymbol{\theta}_0\qquad o_p\ N^{\ \frac{1}{2}}\ , \tag{A2}$$

by Taylor expansion. Similarly,

$$\frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad \frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}}\quad \frac{{}^2L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}\ \boldsymbol{\eta}^{\mathrm{T}}}\ \boldsymbol{\eta}\quad \boldsymbol{\eta}_0\qquad o_p\ N^{\ \frac{1}{2}}\ . \tag{A3}$$

With (A2) and (A3) plugging into (A1), we can obtain the following equation,

$$\overline{N}\frac{{}^2L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}\ \boldsymbol{\theta}^{\mathrm{T}}}\ \boldsymbol{\theta}\quad \boldsymbol{\theta}_0\qquad \overline{N}\frac{{}^2L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}\ \boldsymbol{\eta}^{\mathrm{T}}}\ \boldsymbol{\eta}\quad \boldsymbol{\eta}_0\qquad \overline{N}\frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}}\quad o_p\ 1\qquad 0. \tag{A4}$$

As $\overline{N}\ \boldsymbol{\eta}\quad \boldsymbol{\eta}_0\qquad \mathbf{G}^{\ 1}\ \ \overline{N}\frac{1}{N}\sum_i^N{}_1\ \frac{}{\boldsymbol{\eta}}\log\ \ p\ z_i\quad \mathbf{u}_i;\boldsymbol{\eta}_0\qquad o_p\ 1$ from the asymptotic property of $\boldsymbol{\eta}$, (A4) is equivalent to

$$\overline{N}\frac{^{2}L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}\ \boldsymbol{\theta}^{\mathrm{T}}}\quad \boldsymbol{\theta}\quad \boldsymbol{\theta}_0\qquad \frac{^{2}L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}\ \boldsymbol{\eta}^{\mathrm{T}}}\qquad \mathbf{G}\ ^{1}\ \overline{N}\frac{1}{N}\sum_{i\ 1}^{N}\frac{}{\boldsymbol{\eta}}\log\ \ p\ z_i\quad \mathbf{u}_i;\boldsymbol{\eta}_0$$
$$\overline{N}\frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}}\qquad o_p\ 1\qquad 0.$$

Thus,

$$\overline{N}\ \ \boldsymbol{\theta}\quad \boldsymbol{\theta}_0$$
$$\frac{^{2}L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}\ \boldsymbol{\theta}^{\mathrm{T}}}\ ^{1}\qquad \frac{^{2}L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}\ \boldsymbol{\eta}^{\mathrm{T}}}\qquad \mathbf{G}\ ^{1}\ \overline{N}\frac{1}{N}\sum_{i\ 1}^{N}\frac{}{\boldsymbol{\eta}}\log\ \ p\ z_i\quad \mathbf{u}_i;\boldsymbol{\eta}_0\qquad\qquad \overline{N}\frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}}$$
$$o_p\ 1$$
$$\mathbf{A}\ ^{1}\quad \mathbf{B}\qquad \mathbf{G}\ ^{1}\ \overline{N}\frac{1}{N}\sum_{i\ 1}^{N}\frac{}{\boldsymbol{\eta}}\log\ \ p\ z_i\quad \mathbf{u}_i;\boldsymbol{\eta}_0\qquad\qquad \overline{N}\frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}}\qquad o_p\ 1\ , \tag{A5}$$

where $\frac{^{2}L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}\ \boldsymbol{\theta}^{\mathrm{T}}}\ ^{p}\ \mathbf{A}\quad \mathrm{E}\quad \frac{^{2}\ _{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}\ \boldsymbol{\theta}^{\mathrm{T}}}$, and $\frac{^{2}L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}\ \boldsymbol{\eta}^{\mathrm{T}}}\ ^{p}\ \mathbf{B}\quad \mathrm{E}\quad \frac{^{2}\ _{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}\ \boldsymbol{\eta}^{\mathrm{T}}}$. In addition, we need to form a projection of $\frac{1}{N}\sum_{i\ 1}^{N}\frac{}{\boldsymbol{\eta}}\log\ \ p\ z_i\quad \mathbf{u}_i;\boldsymbol{\eta}_0$ in (A5) through

$$\frac{1}{N}\sum_{i\ 1}^{N}\frac{}{\boldsymbol{\eta}}\log\ \ p\ z_i\quad \mathbf{u}_i;\boldsymbol{\eta}_0\qquad \frac{N}{2}\ ^{1}\sum_{1\ i\ j\ N}\frac{1}{2}\ \frac{}{\boldsymbol{\eta}}\log\ \ p\ z_i\quad \mathbf{u}_i;\boldsymbol{\eta}_0\qquad \frac{}{\boldsymbol{\eta}}\log\ \ p\ z_j\quad \mathbf{u}_j;\boldsymbol{\eta}_0\quad ,$$

and

$$\frac{L\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}}\qquad \frac{N}{2}\ ^{1}\sum_{1\ i\ j\ N}\frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}}.$$

To sum up, (A5) can be formed as

$$\overline{N}\ \ \boldsymbol{\theta}\quad \boldsymbol{\theta}_0\qquad \mathbf{A}\ ^{1}\ \overline{N}\ \frac{N}{2}\ ^{1}\sum_{1\ i\ j\ N}\ \mathbf{BG}\ ^{1}\mathbf{M}_{ij}\ \boldsymbol{\eta}_0\qquad \mathbf{N}_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0\qquad o_p\ 1\ ,$$

where $\mathbf{M}_{ij}\ \boldsymbol{\eta}_0\quad \frac{1}{2}\ \frac{}{\boldsymbol{\eta}}\log\ \ p\ z_i\quad \mathbf{u}_i;\boldsymbol{\eta}_0\qquad \frac{}{\boldsymbol{\eta}}\log\ \ p\ z_j\quad \mathbf{u}_j;\boldsymbol{\eta}_0$ and $\mathbf{N}_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0\qquad \frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}_0}{\boldsymbol{\theta}}$. □

Appendix B. Proof of Theorem 2

Proof. Define function

$$q_{ij}\ \boldsymbol{\theta}\qquad _{ij}\ \ \boldsymbol{\theta}_0\quad \frac{\boldsymbol{\theta}}{\overline{N}},\boldsymbol{\eta}\qquad _{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}\qquad \frac{\boldsymbol{\theta}}{\overline{N}}\ ^{\mathrm{T}}\ \frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad O_p\ \ \frac{1}{N}\ \ , \tag{A6}$$

and we can form a U-statistic based on $q_{ij}\ \ \boldsymbol{\theta}$ as

$$Q_N\ \boldsymbol{\theta}\qquad \frac{2}{N\ N\quad 1}\sum_{1\ i\ j\ N}q_{ij}\ \boldsymbol{\theta}$$
$$L\quad \boldsymbol{\theta}_0\quad \frac{\boldsymbol{\theta}}{\overline{N}}\qquad L\ \boldsymbol{\theta}_0\qquad \frac{1}{\overline{N}}\ \frac{2}{N\ N\quad 1}\boldsymbol{\theta}^{\mathrm{T}}\sum_{1\ i\ j\ N}\frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}.$$

The variance of $Q_N\ \ \boldsymbol{\theta}$ is bounded by var $Q_N\ \boldsymbol{\theta}\qquad \frac{2}{N}$var $q_{ij}\ \ \boldsymbol{\theta}$, from Corollary 3.2 of [37]. Meanwhile, $\frac{2}{N}$var $q_{ij}\ \boldsymbol{\theta}\qquad \frac{2}{N}\ \ \mathrm{E}\ \ q_{ij}\ \boldsymbol{\theta}\ ^{2}\qquad \mathrm{E}\ \ q_{ij}\ \boldsymbol{\theta}\quad ^{2}\qquad \frac{2}{N}\mathrm{E}\ \ q_{ij}\ \boldsymbol{\theta}\ ^{2}$, as $\mathrm{E}\ \ q_{ij}\ \boldsymbol{\theta}\quad ^{2}\qquad 0$. As $_{ij}\ \boldsymbol{\theta},\boldsymbol{\eta}$ is convex, that is, differentiable at $\boldsymbol{\theta}_0$, we can conclude

$$_{ij}\quad \boldsymbol{\theta}_0\quad \frac{\boldsymbol{\theta}}{\overline{N}},\boldsymbol{\eta}\qquad _{ij}\quad \boldsymbol{\theta}_0,\boldsymbol{\eta}\qquad \frac{\boldsymbol{\theta}}{\overline{N}}^{\mathrm{T}}\ \frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}, \tag{A7}$$

from which we can obtain $q_{ij}\ \boldsymbol{\theta}\quad 0$. Similarly,

$$_{ij}\quad \boldsymbol{\theta}_0\quad \frac{\boldsymbol{\theta}}{\overline{N}},\boldsymbol{\eta}\qquad _{ij}\quad \boldsymbol{\theta}_0,\boldsymbol{\eta}\qquad \frac{\boldsymbol{\theta}}{\overline{N}}^{\mathrm{T}}\ \frac{_{ij}\ \boldsymbol{\theta}_0\ \frac{\boldsymbol{\theta}}{\overline{N}},\boldsymbol{\eta}}{\boldsymbol{\theta}}. \tag{A8}$$

From (A6)–(A8), we can conclude

$$0\quad q_{ij}\ \boldsymbol{\theta}\quad \frac{\boldsymbol{\theta}}{\overline{N}}^{\mathrm{T}}\ \frac{_{ij}\ \boldsymbol{\theta}_0\ \frac{\boldsymbol{\theta}}{\overline{N}},\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad \frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad .$$

Therefore, we can bound

$$\frac{2}{N}\mathrm{E}\ \ q_{ij}\ \boldsymbol{\theta}^{\ 2}\qquad \frac{2}{N}\ \frac{1}{\overline{N}}^{2}\ \mathrm{E}\ \ \boldsymbol{\theta}^{\mathrm{T}}\ \ \overline{\boldsymbol{\theta}}\ _{ij}\ \ \boldsymbol{\theta}_0\ \ \frac{\boldsymbol{\theta}}{\overline{N}},\boldsymbol{\eta}\qquad \frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}^{\ 2}\ .$$

The term $\boldsymbol{\theta}^{\mathrm{T}}\ \ \overline{\boldsymbol{\theta}}\ _{ij}\ \ \boldsymbol{\theta}_0\ \ \frac{\boldsymbol{\theta}}{\overline{N}},\boldsymbol{\eta}\quad \frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad ^{p}\ 0$ as $N\quad\infty$. Thus, var $N\ \ Q_N\ \boldsymbol{\theta}\quad ^{p}\ 0$ and consequently

$$N\ \ Q_N\ \boldsymbol{\theta}\qquad N\ \ \mathrm{E}\ \ Q_N\ \boldsymbol{\theta}\quad ^{p}\ 0. \tag{A9}$$

Meanwhile, $\mathrm{E}\ \ Q_N\ \boldsymbol{\theta}\qquad \mathrm{E}\quad _{ij}\ \ \boldsymbol{\theta}_0\ \ \frac{\boldsymbol{\theta}}{\overline{N}},\boldsymbol{\eta}\qquad \mathrm{E}\quad _{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}$. Eventually from (A9) we have

$$N\quad L\quad \boldsymbol{\theta}_0\quad \frac{\boldsymbol{\theta}}{\overline{N}}\qquad L\quad \boldsymbol{\theta}_0\qquad \boldsymbol{\theta}^{\mathrm{T}}\ \ \overline{N}\frac{2}{N\ \ N\quad 1}\ \sum_{1\ \ i\ \ j\ \ N}\ \frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}$$
$$N\ \ \mathrm{E}\quad _{ij}\quad \boldsymbol{\theta}_0\quad \frac{\boldsymbol{\theta}}{\overline{N}},\boldsymbol{\eta}\qquad \mathrm{E}\quad _{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}\quad ^{p}\ 0. \tag{A10}$$

The third term on the left side of (A10) has convergence properties

$$N\ \ \mathrm{E}\quad _{ij}\quad \boldsymbol{\theta}_0\quad \frac{\boldsymbol{\theta}}{\overline{N}},\boldsymbol{\eta}\qquad \mathrm{E}\quad _{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}$$
$$N\ \ \mathrm{E}\quad _{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}\qquad \frac{\boldsymbol{\theta}}{\overline{N}}^{\mathrm{T}}\ \frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad \frac{1}{2}\quad \frac{\boldsymbol{\theta}}{\overline{N}}^{\mathrm{T}}\ \frac{^{2}\ _{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}\ \ \boldsymbol{\theta}^{\mathrm{T}}}\ \frac{\boldsymbol{\theta}}{\overline{N}}\quad o_p\ \ \frac{1}{N}$$
$$\mathrm{E}\quad _{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}$$
$$^{p}\quad \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\theta}.$$

By CLT for U-statistics,

$$\overline{N}\quad \frac{2}{N\ \ N\quad 1}\ \sum_{1\ \ i\ \ j\ \ N}\ \frac{_{ij}\ \boldsymbol{\theta}_0,\boldsymbol{\eta}}{\boldsymbol{\theta}}\quad ^{d}\ \ N\ \ 0,\boldsymbol{\Sigma}\ \ .$$

Using Slutsky's theorem, we can simplify (A10) as

$$N\quad L\quad \boldsymbol{\theta}_0\quad \frac{\boldsymbol{\theta}}{\overline{N}}\qquad L\ \ \boldsymbol{\theta}_0\quad ^{d}\ \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\theta}\quad \boldsymbol{\theta}^{\mathrm{T}}\mathbf{W},$$

where $\mathbf{W}\quad N\ \ 0,\boldsymbol{\Sigma}$. Based on convexity [38], for every compact set $\mathbf{K}\quad \mathbb{R}^{p\quad 1}$, we have

$$N\quad L\quad \boldsymbol{\theta}_0\quad \frac{\boldsymbol{\theta}}{\overline{N}}, \boldsymbol{\eta}\qquad L\ \boldsymbol{\theta}_0, \boldsymbol{\eta}\quad : \boldsymbol{\theta}\quad \mathbf{K}\quad ^{d}\quad \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\theta}\quad \boldsymbol{\theta}^{\mathrm{T}}\mathbf{W} : \boldsymbol{\theta}\quad \mathbf{K}\quad . \tag{A11}$$

Now we develop large sample properties on the penalty term in objective function with adaptive LASSO penalty. We modify the penalty term as

$$N\sum_{j\ \ 1}^{p}\frac{\quad}{j}\qquad _{j,0}\quad \frac{j}{\overline{N}}\qquad N\sum_{j\ \ 1}^{p}\frac{\quad}{j}\qquad _{j,0}\ .$$

From Theorem 1, we have already obtained $\overline{N}\quad _{j}\quad _{j,0}\quad O_p\ 1$. Meanwhile, $N\quad \infty$ and $\overline{N}\quad 0$. If $_{j,0}\quad 0$, then $\overline{N}\ /\ \ _{j}\ \ ^{p}\ 0$ and $\overline{N}\ _{j,0}\quad _{j}\quad \overline{N}\ _{j,0}\quad \mathrm{sign}\quad _{j,0}\quad _{j}$. Eventually

$$N\frac{\quad}{j}\qquad _{j,0}\quad \frac{j}{\overline{N}}\qquad _{j,0}\qquad \overline{N}\frac{\quad}{j}\qquad \overline{N}\ _{j,0}\quad _{j}\qquad \overline{N}\ _{j,0}\quad ^{p}\ 0.$$

If $_{j,0}\quad 0$, then $\overline{N}\ /\quad _{j}\quad N\ /\quad \overline{N}\quad _{j}\quad ^{p}\ \infty$, consequently

$$N\frac{\quad}{j}\qquad _{j,0}\quad \frac{j}{\overline{N}}\qquad _{j,0}\qquad \overline{N}\frac{\quad}{j}\quad _{j}\quad ^{p}\quad \begin{matrix} 0, & \text{if}\ \ _{j}\quad 0, \\ \infty, & \text{if}\ \ _{j}\quad 0. \end{matrix}$$

Therefore, we can summarize

$$N\sum_{j\ \ 1}^{p}\frac{\quad}{j}\qquad _{j,0}\quad \frac{j}{\overline{N}}\qquad _{j,0}\quad ^{p}\quad \begin{matrix} 0, & \text{if}\ \boldsymbol{\beta}\qquad _{1}, \ldots,\ \ _{p_0}, 0, \ldots, 0\ , \\ \infty, & \text{otherwise}. \end{matrix}$$

We have infinity in the limit function, so we cannot use standard argumentation relating to uniform convergence in probability on compacts [39]. However, we can apply slightly more complicated epi-convergence. Thus, based on the works in [23,40,41], we have

$$N\quad L\quad \boldsymbol{\theta}_0\quad \frac{\boldsymbol{\theta}}{\overline{N}}\qquad L\quad \boldsymbol{\theta}_0\qquad N\sum_{j\ \ 1}^{p}\frac{\quad}{j}\qquad _{j,0}\quad \frac{j}{\overline{N}}\qquad _{j,0}\quad ^{e\ \ d}\quad V\ \ \boldsymbol{\theta}\ , \tag{A12}$$

and

$$V\ \ \boldsymbol{\theta}\qquad \begin{matrix} \frac{1}{2}\boldsymbol{\theta}_T^{\mathrm{T}}\mathbf{A}_1\boldsymbol{\theta}_T\quad \boldsymbol{\theta}_T^{\mathrm{T}}\mathbf{W}_T, & \text{if}\ \boldsymbol{\theta}\qquad ,\ \ _{1}, \ldots,\ \ _{p_0}, 0, \ldots, 0\ , \\ \infty, & \text{otherwise}. \end{matrix}$$

and $\mathbf{W}_T\quad N\ \ \mathbf{0}, \boldsymbol{\Sigma}_1$. Specifically, the left side of (A12) is minimized if $\boldsymbol{\theta}\qquad \overline{N}\quad \boldsymbol{\theta}\qquad \boldsymbol{\theta}_0$ and $V\ \ \boldsymbol{\theta}$ has a unique minimizer $\quad \mathbf{A}_1\ ^{1}\mathbf{W}_T\ \ ^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}\quad ^{\mathrm{T}}$ by setting $\frac{V\ \ \boldsymbol{\theta}}{\boldsymbol{\theta}}\quad 0$. Therefore, convergence of minimizers [40] can be concluded from (A12):

$$\overline{N}\quad \boldsymbol{\theta}\quad _{,T}\quad \boldsymbol{\theta}_{0,T}\quad ^{d}\quad \mathbf{A}_1\ ^{1}\mathbf{W}_T\ \text{ and }\ \overline{N}\quad \boldsymbol{\theta}\quad _{,T^c}\quad \boldsymbol{\theta}_{0,T^c}\quad ^{d}\quad \mathbf{0}. \tag{A13}$$

For $j\quad T$,

$$\mathrm{pr}\quad j\ /\ T_N\qquad \mathrm{pr}\quad _{j,}\qquad 0\qquad 0.$$

Thus, $\mathrm{pr}\quad T\qquad T_N\qquad 1$. In addition, $\boldsymbol{\theta}$ minimizes the convex objective function $L\quad \boldsymbol{\theta}$ so that $\mathbf{0}\qquad L\quad \boldsymbol{\theta}$. As $L\quad \boldsymbol{\theta}$ might be nondifferentiable and gradient of $L\quad \boldsymbol{\theta}$ does not exist for some $\boldsymbol{\theta}$, we use $L \quad \boldsymbol{\theta}$ to represent an arbitrary selection of the subgradient of $L \quad \boldsymbol{\theta}$. By taking the subgradient of the objective function with adaptive LASSO penalty, we can obtain

$$L \quad \boldsymbol{\theta} \qquad L\ \boldsymbol{\theta} \qquad \sum_{j\ \ 1}^{p} \frac{\quad}{j} \quad {}_{j,} \qquad .$$

For $j \ / \ T$, pr $\ j \quad T_N$ can be upper bounded by

$$\text{pr} \qquad {}_j L\ \boldsymbol{\theta} \qquad \frac{\quad}{j}\text{sign} \quad {}_{j,} \qquad 0 \qquad \text{pr} \qquad \overline{N} \quad {}_j L\ \boldsymbol{\theta} \qquad \overline{N}\frac{\quad}{j} \qquad , \tag{A14}$$

where ${}_j$ is the j-th coordinate of subgradient and $\overline{N} \ / \quad {}_j \ \ ^{p} \ \infty$ as $j \ / \ T$.

We can expand the subgradient $\overline{N} \ L\ \boldsymbol{\theta}$ as

$$\overline{N} \ L\ \boldsymbol{\theta} \qquad \overline{N} \quad L\ \boldsymbol{\theta} \qquad L\ \boldsymbol{\theta}_0 \quad \mathbf{A} \quad \boldsymbol{\theta} \quad \boldsymbol{\theta}_0 \qquad \overline{N} \ L\ \boldsymbol{\theta}_0 \qquad \overline{N}\mathbf{A} \quad \boldsymbol{\theta} \quad \boldsymbol{\theta}_0 \ , \tag{A15}$$

where $\overline{N} \ L\ \boldsymbol{\theta}_0$ is bounded in probability, $\overline{N}\mathbf{A} \quad \boldsymbol{\theta} \quad \boldsymbol{\theta}_0 \quad ^{D} \quad \overline{N}\mathbf{W}$ which is bounded in probability as well. By Theorem 1 of the work in [42],

$$\sup_{\boldsymbol{\theta} \quad \boldsymbol{\theta}_0 \quad M/ \quad \overline{N}} \quad L\ \boldsymbol{\theta} \qquad L\ \boldsymbol{\theta}_0 \quad \mathbf{A} \quad \boldsymbol{\theta} \quad \boldsymbol{\theta}_0 \qquad o_p \ \frac{1}{\overline{N}} \quad .$$

Therefore, $\overline{N} \quad L\ \boldsymbol{\theta} \qquad L\ \boldsymbol{\theta}_0 \quad \mathbf{A} \quad \boldsymbol{\theta} \quad \boldsymbol{\theta}_0 \quad ^{p} \ 0$. Finally, $\overline{N} \quad {}_j L\ \boldsymbol{\theta}$ is bounded and the right side of (A14) converges to 0, which proves pr $\ j \quad T_N \quad 0$ for $j \ / \ T$. □

References

1. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2002.
2. Shao, J.; Zhao, J. Estimation in longitudinal studies with nonignorable dropout. *Stat. Its Interface* **2013**, *6*, 303–313. [CrossRef]
3. Wang, S.; Shao, J.; Kim, J.K. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Stat. Sin.* **2014**, *24*, 1097–1116. [CrossRef]
4. Zhao, J.; Shao, J. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *J. Am. Stat. Assoc.* **2015**, *110*, 1577–1590. [CrossRef]
5. Miao, W.; Tchetgen Tchetgen, E.J. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika* **2016**, *103*, 475–482. [CrossRef]
6. Zhao, J.; Ma, Y. Optimal pseudolikelihood estimation in the analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **2018**, *105*, 479–486. [CrossRef]
7. Miao, W.; Liu, L.; Tchetgen Tchetgen, E.; Geng, Z. Identification, Doubly Robust Estimation, and Semiparametric Efficiency Theory of Nonignorable Missing Data With a Shadow Variable. *arXiv* **2019**, arXiv:1509.02556.
8. Tchetgen Tchetgen, E.J.; Wirth, K.E. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics* **2017**, *73*, 1123–1131. [CrossRef]
9. Sun, B.; Liu, L.; Miao, W.; Wirth, K.; Robins, J.; Tchetgen Tchetgen, E.J. Semiparametric estimation with data missing not at random using an instrumental variable. *Stat. Sin.* **2018**, *28*, 1965–1983.
10. Zhao, J.; Yang, Y.; Ning, Y. Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data. *Stat. Sin.* **2018**, *28*, 2125–2148. [CrossRef]
11. Jiang, W.; Bogdan, M.; Josse, J.; Miasojedow, B.; Rockova, V.; Group, T. Adaptive Bayesian SLOPE–High-dimensional Model Selection with Missing Values. *arXiv* **2019**, arXiv:1909.06631.

12. Jiang, W.; Josse, J.; Lavielle, M.; Group, T. Logistic regression with missing covariates—Parameter estimation, model selection and prediction within a joint-modeling framework. *Comput. Stat. Data Anal.* **2020**, *145*, 106907. [CrossRef]
13. Johnson, A.E.; Pollard, T.J.; Shen, L.; Li-wei, H.L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.A.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef] [PubMed]
14. Zhao, J.; Ma, Y. A versatile estimation procedure without estimating the nonignorable missingness mechanism. *arXiv* **2019**, arXiv:1907.03682.
15. Liang, K.Y.; Qin, J. Regression analysis under non-standard situations: A pairwise pseudolikelihood approach. *J. R. Stat. Soc. Ser. B* **2000**, *62*, 773–786. [CrossRef]
16. Zhao, J.; Shao, J. Approximate conditional likelihood for generalized linear models with general missing data mechanism. *J. Syst. Sci. Complex.* **2017**, *30*, 139–153. [CrossRef]
17. Zhao, J. Reducing bias for maximum approximate conditional likelihood estimator with general missing data mechanism. *J. Nonparametr. Stat.* **2017**, *29*, 577–593. [CrossRef]
18. Yang, Y.; Zhao, J.; Wilding, G.; Kluczynski, M.; Bisson, L. Stability enhanced variable selection for a semiparametric model with flexible missingness mechanism and its application to the ChAMP study. *J. Appl. Stat.* **2020**, *47*, 827–843. [CrossRef]
19. Zhao, J.; Chen, C. Estimators based on unconventional likelihoods with nonignorable missing data and its application to a children's mental health study. *J. Nonparametric Stat.* **2019**, *31*, 911–931. [CrossRef]
20. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
21. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
22. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
23. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
24. Cai, T.; Tian, L.; Wei, L. Semiparametric Box–Cox power transformation models for censored survival observations. *Biometrika* **2005**, *92*, 619–632. [CrossRef]
25. Kosorok, M.R. *Introduction to Empirical Processes and Semiparametric Inference*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007.
26. Minnier, J.; Tian, L.; Cai, T. A perturbation method for inference on regularized regression estimates. *J. Am. Stat. Assoc.* **2011**, *106*, 1371–1382. [CrossRef]
27. Hu, Z.; Melton, G.B.; Arsoniadis, E.G.; Wang, Y.; Kwaan, M.R.; Simon, G.J. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *J. Biomed. Inform.* **2017**, *68*, 112–120. [CrossRef]
28. Li, J.; Wang, M.; Steinbach, M.S.; Kumar, V.; Simon, G.J. Don't Do Imputation: Dealing with Informative Missing Values in EHR Data Analysis. In Proceedings of the 2018 IEEE International Conference on Big Knowledge (ICBK), Singapore, 17–18 November 2018; pp. 415–422.
29. Phillips, A.; Shaper, A.G.; Whincup, P. Association between serum albumin and mortality from cardiovascular disease, cancer, and other causes. *Lancet* **1989**, *334*, 1434–1436. [CrossRef]
30. Katz, S.; Klotz, I.M. Interactions of calcium with serum albumin. *Arch. Biochem. Biophys.* **1953**, *44*, 351–361. [CrossRef]
31. Butler, S.; Payne, R.; Gunn, I.; Burns, J.; Paterson, C. Correlation between serum ionised calcium and serum albumin concentrations in two hospital populations. *Br. Med. J.* **1984**, *289*, 948–950. [CrossRef]
32. Hossain, A.; Mostafa, G.; Mannan, K.; Prosad Deb, K.; Hossain, M. Correlation Between Serum Albumin Level and Ionized Calcium in Idiopathic Nephrotic Syndrome in Children. *Urol. Nephrol. Open Access. J.* **2015**, *3*, 70–71. [CrossRef]
33. Kroll, M.; Elin, R. Relationships between magnesium and protein concentrations in serum. *Clin. Chem.* **1985**, *31*, 244–246. [CrossRef]
34. Huijgen, H.J.; Soesan, M.; Sanders, R.; Mairuhu, W.M.; Kesecioglu, J.; Sanders, G.T. Magnesium levels in critically ill patients: What should we measure? *Am. J. Clin. Pathol.* **2000**, *114*, 688–695. [CrossRef]

35. Djagbletey, R.; Phillips, B.; Boni, F.; Owoo, C.; Owusu-Darkwa, E.; deGraft Johnson, P.K.G.; Yawson, A.E. Relationship between serum total magnesium and serum potassium in emergency surgical patients in a tertiary hospital in Ghana. *Ghana Med. J.* **2016**, *50*, 78–83. [CrossRef]
36. Luo, X.; Tsai, W.Y. A proportional likelihood ratio model. *Biometrika* **2011**, *99*, 211–222. [CrossRef]
37. Shao, J. *Mathematical Statistics*; Springer Texts in Statistics; Springer: Berlin/Heidelberg, Germany, 2003.
38. Arcones, M.A. Weak convergence of convex stochastic processes. *Stat. Probab. Lett.* **1998**, *37*, 171–182. [CrossRef]
39. Rejchel, W. Model selection consistency of U-statistics with convex loss and weighted lasso penalty. *J. Nonparametric Stat.* **2017**, *29*, 768–791. [CrossRef]
40. Geyer, C.J. On the asymptotics of constrained *M*-estimation. *Ann. Stat.* **1994**, *22*, 1993–2010. [CrossRef]
41. Pflug, G.C. Asymptotic stochastic programs. *Math. Oper. Res.* **1995**, *20*, 769–789. [CrossRef]
42. Niemiro, W. Least empirical risk procedures in statistical inference. *Appl. Math.* **1993**, *22*, 55–67. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.