Statistica Sinica Preprint No: SS-2016-0312	
Title	Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data
Manuscript ID	SS-2016-0312
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0312
Complete List of Authors	Jiwei Zhao Yang Yang and Yang Ning
Corresponding Author	Jiwei Zhao
E-mail	zhaoj@buffalo.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

Statistica Sinica

PENALIZED PAIRWISE PSEUDO LIKELIHOOD FOR VARIABLE SELECTION WITH NONIGNORABLE MISSING DATA

Jiwei Zhao[1], Yang Yang[1] and Yang Ning[2]

[1]*State University of New York at Buffalo,* [2]*Cornell University*

Abstract: The regularization approach for variable selection was well developed for a completely observed data set in the past two decades. In the presence of missing values, this approach needs to be tailored to different missing data mechanisms. In this paper, we focus on a flexible and generally applicable missing data mechanism. That contains both ignorable and nonignorable missing data mechanism assumptions. We show how the regularization approach for variable selection can be adapted to the situation under this missing data mechanism. The computational and theoretical properties for variable selection consistency are established. The proposed method is further illustrated by comprehensive simulation studies and data analyses.

Key words and phrases: Variable selection, Regularization, Missing data mechanism, Nonignorable missing data, Penalized pairwise pseudo likelihood, Selection consistency.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

1. Introduction

Variable selection is an important topic in regression analysis. In the past two decades, researchers investigated a series of regularization approaches for variable selection and developed both theoretical and computational properties. The two mainstream techniques are the LASSO (Least Absolute Shrinkage and Selection Operator; Tibshirani (1996)), or the L_1-penalization, and the nonconvex penalizations such as SCAD (Smoothly Clipped Absolute Deviation; Fan and Li (2001)) and MCP (Minimax Concave Penalty; Zhang (2010)). The LASSO owns its popularity largely due to its computational convenience, but it induces estimation bias for parameters with large absolute values. Using a nonconvex penalty, one has to minimize a nonconvex function, which raises extra computational challenges, but the intrinsic estimation bias of the LASSO can be eliminated, and corrected. With completely observed data, it is shown that, under regularity conditions, both of the two mainstream techniques achieve variable selection consistency properties; however, when missing values are present the appropriate regularization approaches are relatively limited in the literature. In this paper, we explore how to use both techniques for variable selection in the presence of missing data.

In the missing data literature, one often defines an indicator R, with

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

$R = 1$ for a completely observed subject, and $R = 0$ otherwise. The probability distribution function of R conditional on all the data, termed the missing data mechanism (Little and Rubin, 2002), should be incorporated in the analysis compensating for the effect of missing data. There are various missing data mechanism assumptions. Briefly, if it only depends on the completely observed data, the mechanism is called missing at random (MAR); otherwise, it is called missing not at random, or nonignorable. The likelihood-based methods, usually under the MAR assumption, can be derived for variable selection. Most of the currently existing literature falls in this category. For example, Ibrahim et al. (2008) developed likelihood methods for the computation of model selection criteria based on the output of the EM algorithm. They derived a class of information criteria for missing data problems. Garcia et al. (2010) considered the regularization approach using SCAD or adaptive LASSO and adopted the EM technique to formulate the observed likelihood for variable selection in a low-dimensional setting.

In general, likelihood-based methods need to specify a parametric distribution of the missing data mechanism. One has to be cautious about this type of assumption. First, it is well known that a parametric assumption is very sensitive and may easily induce a misspecified model. If this happens,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

it prompts biased estimation and inaccurate selection results. Second, although MAR occurs in some applications, in many situations there is a suspicion that the missing data mechanism is nonignorable (Ibrahim et al., 1999). For nonignorable missing data, applying methods derived under the MAR assumption may result in serious estimation bias and incorrect conclusions. Third, the situation with nonignorable missing data is generally more challenging to deal with. One notorious feature of nonignorable missingness is the identifiability issue (Robins and Ritov, 1997). In theory, one has to first carefully study the model identification conditions before doing any statistical analyses. Refer to Kim and Shao (2013) for the most recent developments of nonignorable missing data.

Due to the complexity of the missing data mechanism assumptions, in reality one can carry out the sensitivity analysis to validate the analysis results. The other preferred and ideal remedy is to impose an assumption as flexible and generally applicable as possible. This type of assumption usually does not specify a parametric model, and is often called an unspecified missing data mechanism. The works of Liang and Qin (2000); Tang et al. (2003); Shao and Zhao (2013); Zhao and Shao (2015); Fang et al. (2017); Zhao (2017) follow this direction.

In this paper, our motivation is to conduct variable selection with miss-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

ing data, and more interestingly with nonignorable missing data. Under the high dimensional setting, we consider the generalized linear model (GLM; McCullagh and Nelder (1989)), which can be applied to either continuous or categorical data. We impose an unspecified missing data mechanism assumption, which is flexible and generally applicable and robust for the potential model misspecification. Besides MAR cases, it contains many nonignorable scenarios. Under this assumption, although not all parameters are identifiable, a pseudo-likelihood function that produces an estimator of a dispersion-scaled version of the original parameter is developed. We show that the variable selection can be carried out by penalizing the aforementioned pseudo-likelihood through an estimable dispersion-scaled parameter. Due to the messy missing data and the flexible mechanism assumption, we may not fully retrieve all the information contained in the original data, hence are not able to estimate all the unknown parameters. However, the key idea is that our regularization procedure can still be carried out for the purpose of variable selection, based on the pseudo-likelihood function and the estimable dispersion-scaled parameter.

We propose algorithms to efficiently optimize the penalized pseudo likelihood for both the LASSO and nonconvex penalties. This is not a trivial task due to the complicated U-statistic structure in the pseudo-likelihood

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

function. For the LASSO penalty, we find that, the objective function can be transformed to the penalized likelihood function for a standard penalized logistic regression model without the intercept term after some data manipulation. More importantly, for the nonconvex penalties, we develop an iterative algorithm based on the trick we used in the LASSO and the local linear approximation (LLA; Zou and Li (2008); Fan et al. (2014)).

We show that, under the high dimensional setting, variable selection consistency can be achieved with some mild regularity conditions. The challenges in this are from the complicated pairwise U-statistic structure in the pseudo-likelihood, and from the nonconvex penalties.

There is scarce literature on high dimensional problems with missing data. Loh and Wainwright (2012) considered a linear model with covariates that may have missing values, and studied the theoretical properties of the estimators using a regularization approach via the LASSO. More recently, Ning et al. (2017) showed some results on parameter estimation in a similar context, but our paper is distinctively different from theirs.

The remainder of the paper is organized as follows. In Section 2, we provide a brief review of the regularization approach in the case of no missing data, and then introduce our proposed penalized pseudo-likelihood. The algorithms designed for both the LASSO and nonconvex penalties are pre-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

sented in Section 3. Section 4 contains the theoretical results on variable selection consistency, Section 5 includes the numerical results illustrating the finite sample performance of our proposed method and its comparison with some existing methods. In Section 6, we conclude our paper with a discussion. Technical details, some extra simulation studies and two data analyses are in the online supplementary material.

2. Methodology

2.1 Brief Review in the Case of No Missing Data

Assume that we have a collection of independent observations $\{y_i, \boldsymbol{x}_i\}, i = 1, \ldots, N$, where $(y_i, \boldsymbol{x}_i)$'s are identically distributed realizations of $(Y, \boldsymbol{X})$. We let Y denote the scalar response variable, and $\boldsymbol{X}$ be a p-dimensional covariate variable. Assume that, with a canonical link, the conditional distribution of Y given $\boldsymbol{X}$ belongs to a generalized linear model (GLM; McCullagh and Nelder (1989)) with the density

$$p(Y|\boldsymbol{X};\boldsymbol{\theta}) = \exp[\phi^{-1}\{Y\eta - b(\eta)\} + c(y;\phi)], \tag{2.1}$$

where b and c are known functions, $\eta = \alpha + \boldsymbol{\beta}^T\boldsymbol{X}$, $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^T, \phi)^T$, and ϕ represents the positive dispersion parameter.

To carry out variable selection through the regularization approach, we

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

obtain the minimizer of the penalized likelihood function

$$-\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})+\sum_{j=1}^{p}p_\lambda(|\beta_j|), \tag{2.2}$$

where $p_\lambda(t)$ represents a penalty function, and $\lambda \geq 0$ is the tuning parameter. Here the penalty term is only applied to $\boldsymbol{\beta}$. The variable selection can be achieved without estimating the dispersion parameter ϕ.

In the LASSO penalty, $p_\lambda(t) = \lambda|t|$, and $p'_\lambda(t) = \lambda$ for $t > 0$. Due to the convexity of the LASSO, the coordinate descent algorithm Friedman et al. (2010) is very efficient in minimizing (2.2). Theoretically, a strong irrepresentable condition is necessary for the LASSO to be selection consistent (Zhao and Yu, 2006).

In Fan and Li (2001), the authors advocated penalty functions that provide estimators with three properties: sparsity, unbiasedness and continuity. Clearly, the L_1-penalty does not satisfy the unbiasedness property. In this class of nonconvex penalty functions, two frequently used representatives are the SCAD and the MCP. The SCAD (Fan and Li, 2001) is

$$p'_\lambda(t) = \lambda\mathbb{1}(t \leq \lambda) + \frac{(a\lambda - t)_+}{a-1}\mathbb{1}(t > \lambda),$$

for some $a > 2$ and $t > 0$, where $\mathbb{1}(\cdot)$ is the indicator function; the MCP (Zhang, 2010) is

$$p'_\lambda(t) = \frac{(a\lambda - t)_+}{a},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

for some $a > 0$ and $t > 0$. Numerous papers have been devoted to study the statistical properties of the resulting estimators, for instance, Fan and Lv (2010, 2011) and the references therein. Computation for this approach is much more involved, because the resulting optimization problem (2.2) is nonconvex and may have multiple local minimizers. Fan and Li (2001) proposed the local quadratic approximation algorithm as a unified method for optimizing the nonconvex penalized likelihood, while Zou and Li (2008) worked out a local linear approximation (LLA) algorithm that turns a nonconvex penalization problem into a series of reweighed L_1-penalization problems. Both of them are relevant to the majorization-minimization (MM) principle (Hunter and Lange, 2004).

2.2 Variable Selection with Missing Data

We use the variable R indicating whether the data from each subject are completely observed. Without loss of generality, we take the first n subjects as fully observed, $r_i = 1$, $i = 1, \ldots, n$, and the remaining $N - n$ subjects may contain missing components, $r_i = 0, i = n + 1, \ldots, N$.

The foremost difficulty dealing with missing data is the assumption on the missing data mechanism, $\Pr(R = 1|Y, \boldsymbol{X})$. With a parametric model for it, the likelihood based methods can be developed for model selection,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

especially for the MAR case Ibrahim et al. (2008); Garcia et al. (2010). However, there are important limitations to adopting this approach. Since the underlying truth of the missing data mechanism is unknown and its assumption is unverifiable, one looks for an assumption that is robust and as flexible and generally applicable as possible.

We impose the general assumption

$$\Pr(R = 1|Y, \boldsymbol{X}) = s(Y)t(\boldsymbol{X}), \tag{2.3}$$

where s and t are some functions, not necessarily known or specified. We do not impose any concrete form on them. We assume $0 < \Pr(R = 1) < 1$ throughout.

Our assumption is very flexible and it includes many specific scenarios commonly seen in the missing data literature. Thus with Y having missing values and $\boldsymbol{X}$ fully observed is a special case of (2.3) if s=constant; the nonignorable nonresponse assumption in Tang et al. (2003) is also a special case of (2.3) if t=constant. Situations included in (2.3) can also allow the covariate $\boldsymbol{X}$ to have missing values, and both response Y and covariate $\boldsymbol{X}$ to have missing values.

Chan (2013) considered the problem of nuisance parameter elimination in a proportional likelihood ratio model under this assumption and Zhao and Shao (2017) studied identifiability in a GLM with non-canonical

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

link under this assumption. Both of them only considered the classic low-dimensional statistical models. More recently, Ning et al. (2017) studied the parameter estimation problem and an associated inference procedure under this assumption in a high-dimensional setting. Here, we focus on the variable selection problem in a high-dimensional GLM and there are challenges due to the high dimensionality, compared to Chan (2013) and Zhao and Shao (2017). In a high-dimensional framework, Ning et al. (2017) mainly addressed the estimation problem that needs different analytic tools than we do.

Because of the complexity of the missing data structure and the presence of unknown functions s and t, we propose a pseudo likelihood function. Note that

$$p(Y|\boldsymbol{X}, R=1) = \frac{\Pr(R=1|Y,\boldsymbol{X})}{w(\boldsymbol{X})} p(Y|\boldsymbol{X}), \tag{2.4}$$

where $w(\boldsymbol{X}) = \int \Pr(R=1|Y,\boldsymbol{X}) p(Y|\boldsymbol{X}) dY = \Pr(R=1|\boldsymbol{X})$. Under (2.3), $\Pr(R=1|Y,\boldsymbol{X})/w(\boldsymbol{X})$ in (2.4) is a multiplier of an $\boldsymbol{X}$-only function $s(\boldsymbol{X})/w(\boldsymbol{X})$, and a Y-only function $t(Y)$. Restricting attention to completely observed subjects with subscripts ranging from $\{1,\ldots,n\}$, decomposing $\{y_1,\ldots,y_n\}$ as rank statistics and order statistics, and conditioning on the order statistics $\{y_{(1)},\ldots,y_{(n)}\}$, we have the conditional likelihood for

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

$\boldsymbol{\theta}$ as

$$p(y_1, \ldots, y_n | r_1 = \ldots = r_n = 1, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, y_{(1)}, \ldots, y_{(n)}).$$

After some derivations, it can be shown that this is

$$\frac{\Pi_{i=1}^n p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})}{\sum_c \Pi_{i=1}^n p(y_{(i)}|\boldsymbol{x}_i;\boldsymbol{\theta})}, \tag{2.5}$$

where the summation in the denominator covers all possible permutations of $\{1, \ldots, n\}$.

A nice feature of this method is that all s, t and w functions are all canceled out through conditioning. This idea was first outlined in Kalbfleisch (1978), but in practice it encounters a computational burden with an order of $n!$ (Liang and Qin, 2000). To reduce the computational burden, Liang and Qin (2000) advocated the pairwise pseudo likelihood

$$\prod_{1\le i<j\le n} \frac{p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})p(y_j|\boldsymbol{x}_j;\boldsymbol{\theta})}{p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})p(y_j|\boldsymbol{x}_j;\boldsymbol{\theta}) + p(y_i|\boldsymbol{x}_j;\boldsymbol{\theta})p(y_j|\boldsymbol{x}_i;\boldsymbol{\theta})}. \tag{2.6}$$

Under the GLM assumption, the negative part of the log-version of (2.6), after adding a normalizing constant, can be written as

$$\mathcal{L}(\boldsymbol{\gamma}) = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \log\{1 + \exp(-y_{i\backslash j}\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\gamma})\}, \tag{2.7}$$

where $y_{i\backslash j} = y_i - y_j$, $\boldsymbol{x}_{i\backslash j} = \boldsymbol{x}_i - \boldsymbol{x}_j$ and $\boldsymbol{\gamma} = \boldsymbol{\beta}/\phi$. To perform variable selection, we propose to minimize the penalized pairwise pseudo likelihood

$$\mathcal{L}(\boldsymbol{\gamma}) + \sum_{j=1}^p p_\lambda(|\gamma_j|) = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \log\{1 + \exp(-y_{i\backslash j}\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\gamma})\} + \sum_{j=1}^p p_\lambda(|\gamma_j|) \tag{2.8}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

and we denote the minimizer as $\widehat{\boldsymbol{\gamma}}$. It can be seen that, the unpenalized component $\mathcal{L}(\boldsymbol{\gamma})$ is a U-statistic, where even the original function b in the definition of GLM disappears. We cannot estimate the whole unknown parameter $\boldsymbol{\theta}$ itself, but we can estimate a dispersion-scaled parameter $\boldsymbol{\gamma} = \boldsymbol{\beta}/\phi$ and we carry out variable selection through this dispersion-scaled parameter.

We turn to the computational and theoretical properties of variable selection through the regularization approach (2.8).

3. Computational Algorithms

The unpenalized component $\mathcal{L}(\boldsymbol{\gamma})$ in (2.8) is a U-statistic and it is not trivial to optimize. In this Section, we propose tractable and efficient algorithms to minimize (2.8) for the LASSO and nonconvex penalties. We also discuss how to choose the regularization tuning parameter λ.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

3.1 Algorithm for the LASSO

The unpenalized component $\mathcal{L}(\boldsymbol{\gamma})$ in (2.8) can be written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}) &= \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \log\{1+\exp(-y_{i\backslash j}\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\gamma})\} \\ &= \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \log\{1+\exp(-\operatorname{sign}(y_{i\backslash j})|y_{i\backslash j}|\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\gamma})\} \\ &= \frac{2m}{n(n-1)} \cdot \frac{1}{m}\sum_{k=1}^{m} \log\{1+\exp(w_k\boldsymbol{v}_k^T\boldsymbol{\gamma})\} + \left\{1-\frac{2m}{n(n-1)}\right\}\log(2),\end{aligned}$$

where we let m denote the number of terms in the summation across $1 \le i < j \le n$ such that $y_{i\backslash j} \neq 0$. For example, when Y is continuous, $m = n(n-1)/2$; when Y is binary, $m = n_0 n_1$, where n_0 is the total number of 0's and n_1 is the total number of 1's, and $n_0 + n_1 = n$. Also, we let $\operatorname{sign}(\cdot)$ denote the sign function, and we define $w_k = -\operatorname{sign}(y_{i\backslash j})$ and $\boldsymbol{v}_k = \boldsymbol{x}_{i\backslash j}|y_{i\backslash j}|$ for $k = 1, \ldots, m$.

It can be seen that, the essential component $\frac{1}{m}\sum_{k=1}^{m}\log\{1+\exp(w_k\boldsymbol{v}_k^T\boldsymbol{\gamma})\}$ in $\mathcal{L}(\boldsymbol{\gamma})$ can be treated as the negative log-likelihood function of a regular logistic regression with response u_k, covariate $\boldsymbol{v}_k$, without the intercept term, where

$$u_k = \begin{cases} 1 & \text{if } y_{i\backslash j} > 0 \\ 0 & \text{if } y_{i\backslash j} < 0. \end{cases}$$

Therefore, to minimize (2.8) with the LASSO penalty, after the aforementioned data manipulation, it can be carried out directly as a regular

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

penalized logistic regression forcing the intercept to zero, with m subjects where the k-th subject has response u_k and covariate $\boldsymbol{v}_k$. In R, this procedure can be implemented using the package `glmnet` Friedman et al. (2010).

3.2 Algorithm for Nonconvex Penalties

With nonconvex penalties such as the SCAD and the MCP, we adopt a similar data manipulation technique as for the LASSO, and the LLA algorithm (Zou and Li, 2008; Fan et al., 2014). The LLA algorithm transforms a concave regularization problem into a series of weighted L_1-penalization problems by taking advantage of the nonconvex structure of the penalty functions and the MM principle. Moreover, the MM principle provides a guarantee on the convergence of the LLA algorithm to a stationary point of the nonconvex penalization problem. In Fan et al. (2014), the authors showed that, as long as the problem is localizable and the oracle estimator is well behaved, one can obtain the oracle estimator by using the one-step LLA. In addition, once the oracle estimator is obtained, the LLA algorithm produces the same estimator in its following iterations. Here, we summarize the details of the LLA algorithm as follows

1. Initialize $\widehat{\boldsymbol{\gamma}}^{(0)} = (\hat{\gamma}_1^{(0)}, \ldots, \hat{\gamma}_p^{(0)})^T$ and compute the adaptive weight

$$\widehat{\boldsymbol{\omega}}^{(0)} = (\hat{\omega}_1^{(0)}, \ldots, \hat{\omega}_p^{(0)})^T = (p'_\lambda(|\hat{\gamma}_1^{(0)}|), \ldots, p'_\lambda(|\hat{\gamma}_p^{(0)}|))^T.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

2. For $m = 1, 2, \ldots$, repeat the LLA iteration till convergence

2.a Obtain $\widehat{\boldsymbol{\gamma}}^{(m)}$ by solving the optimization problem

$$\widehat{\boldsymbol{\gamma}}^{(m)} = \arg\min_{\boldsymbol{\gamma}} \left\{ \mathcal{L}(\boldsymbol{\gamma}) + \sum_{j=1}^{p} \hat{\omega}_j^{(m-1)} |\gamma_j| \right\}, \tag{3.9}$$

2.b Update the adaptive weight vector $\widehat{\boldsymbol{\omega}}^{(m)}$ with $\hat{\omega}_j^{(m)} = p'_\lambda(|\hat{\gamma}_j^{(m)}|)$.

In our numerical studies, the initial $\widehat{\boldsymbol{\gamma}}^{(0)}$ is chosen as the LASSO solution. In R, the major step (3.9) is implemented using the package `glmnet`.

3.3 Tuning Parameter Selection

How to select the regularization parameter λ is of paramount importance in penalized likelihood estimation since λ governs the complexity of the selected model. A large value of λ tends to choose a simple model, whereas a small value of λ inclines to a complex model. The trade-off between the model complexity and the prediction accuracy yields an optimal choice of λ. This is frequently done by using a K-fold cross-validation. Specifically, we denote the data set indexed by $\{1, \ldots, n\}$ as T, and cross validation training and test sets by $T \backslash T^{(\kappa)}$ and $T^{(\kappa)}$, for $\kappa = 1, \ldots, K$. Each time, for fixed λ and κ, we find the minimizer $\widehat{\boldsymbol{\gamma}}^{(-\kappa)}(\lambda)$ of $\mathcal{L}(\boldsymbol{\gamma}) + \sum_{j=1}^{p} p_\lambda(|\gamma_j|)$ using the training set $T \backslash T^{(\kappa)}$. Finally, we choose λ to be the minimizer of the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

cross validation function

$$\mathrm{CV}(\lambda) = \sum_{\kappa=1}^{K} \mathcal{L}^{(\kappa)}(\widehat{\boldsymbol{\gamma}}^{(-\kappa)}(\lambda)),$$

where $\mathcal{L}^{(\kappa)}(\cdot)$ represents the evaluation of $\mathcal{L}(\cdot)$ using the test set $T^{(\kappa)}$.

Alternatively one can select λ by an information criterion, for example, the generalized information criterion for high-dimensional penalized likelihood proposed by Fan and Tang (2013). They showed that the criterion with a uniform choice of the model complexity penalty identifies the true model with probability tending to 1 when the dimensionality grows at most exponentially fast with the sample size.

Although cross validation is computationally more expensive, it is less parsimonious and can often yield more satisfactory performance in practice. In this paper, we select the tuning parameter λ by the K-fold cross validation with $K = 5$.

4. Theoretical Results

We present the theoretical conditions and properties of our method for variable selection in the presence of missing data. For interpretation simplicity, we only show the results for a family of nonconvex penalties, including the SCAD and the MCP. Parallel results for the LASSO can be similarly developed, and hence are skipped. The assumptions for the LASSO to

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

be selection consistent are stronger (Zhao and Yu, 2006). The results we present hold when the number of covariates can grow at most exponentially fast with the sample size.

4.1 Notations

We need some notation. For positive sequences a_n and b_n, we write $a_n \lesssim b_n$, if $a_n/b_n = O(1)$. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For a vector $\boldsymbol{v} = (v_1, \ldots, v_p)^T \in R^p$, we take $\text{supp}(\boldsymbol{v}) = \{i : v_i \neq 0\}$, $|\text{supp}(\boldsymbol{v})| = \text{card}\{\text{supp}(\boldsymbol{v})\} = \|\boldsymbol{v}\|_0$, and $|A|$ is the cardinality of a set A. For $1 \leq q < \infty$, the L_q-norm is $\|\boldsymbol{v}\|_q = (\sum_{i=1}^p |v_i|^q)^{1/q}$. Let $\|\boldsymbol{v}\|_\infty = \max_{1\leq i\leq p} |v_i|$ be the L_∞-norm and $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^T$ be the Kronecker product. For two vectors $\boldsymbol{v}, \boldsymbol{u} \in R^p$, we write $\boldsymbol{v} \circ \boldsymbol{u} = (v_1u_1, \ldots, v_pu_p)^T$ as the Hadamard product. For an $n \times p$ matrix $\mathbf{M}$, its matrix L_1-norm is $\|\mathbf{M}\|_{L_1} = \max_{1\leq j\leq p} \sum_{i=1}^n |M_{ij}|$, the spectral norm is $\|\mathbf{M}\|_2 = \sqrt{\lambda_{\max}(\mathbf{M}^T\mathbf{M})}$, the matrix L_∞-norm is $\|\mathbf{M}\|_{L_\infty} = \max_{1\leq i\leq n} \sum_{j=1}^p |M_{ij}|$, the elementwise L_1-norm is $\|\mathbf{M}\|_1 = \sum_{i=1}^n \sum_{j=1}^p |M_{ij}|$, and the elementwise supreme norm is $\|\mathbf{M}\|_\infty = \max_{i,j}\{|M_{ij}|\}$. If $\mathbf{M}$ is squared and symmetric, we write $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ as the minimal and maximal eigenvalues of $\mathbf{M}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

The pairwise pseudo likelihood in (2.8) can be written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}) &= \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \log\{1+\exp(-y_{i\backslash j}\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\gamma})\} \\ &= \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \{\psi(y_{i\backslash j}\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\gamma}) - y_{i\backslash j}\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\gamma}\},\end{aligned}$$

and its first and second order gradients are

$$\nabla\mathcal{L}(\boldsymbol{\gamma}) = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \{\psi'(y_{i\backslash j}\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\gamma})y_{i\backslash j}\boldsymbol{x}_{i\backslash j} - y_{i\backslash j}\boldsymbol{x}_{i\backslash j}\},$$

$$\nabla^2\mathcal{L}(\boldsymbol{\gamma}) = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \{\psi''(y_{i\backslash j}\boldsymbol{x}_{i\backslash j}^T\boldsymbol{\gamma})y_{i\backslash j}^2\boldsymbol{x}_{i\backslash j}^{\otimes 2}\},$$

where we have $\psi(t) = \log(1+e^t)$, and hence $\psi'(t) = \frac{e^t}{1+e^t}$, $\psi''(t) = \frac{e^t}{(1+e^t)^2}$, $\psi'''(t) = \frac{e^t(1-e^t)}{(1+e^t)^3}$. After some algebra, it can be verified that the derivative functions are bounded, with $|\psi''(t)| \le 0.25$ and $|\psi'''(t)| \le 0.1$.

Throughout, we denote the penalty function as $\mathcal{P}_\lambda(\boldsymbol{\gamma}) = \sum_{j=1}^p p_\lambda(|\gamma_j|)$. We take $q_\lambda(t) = p_\lambda(t) - \lambda|t|$, $\mathcal{Q}_\lambda(\boldsymbol{\gamma}) = \mathcal{P}_\lambda(\boldsymbol{\gamma}) - \lambda\|\boldsymbol{\gamma}\|_1$, and $\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}) = \mathcal{L}(\boldsymbol{\gamma}) + \mathcal{Q}_\lambda(\boldsymbol{\gamma}) = \mathcal{L}(\boldsymbol{\gamma}) + \mathcal{P}_\lambda(\boldsymbol{\gamma}) - \lambda\|\boldsymbol{\gamma}\|_1$. Therefore, the penalized objective function in (2.8) can be written as

$$\mathcal{L}(\boldsymbol{\gamma}) + \mathcal{P}_\lambda(\boldsymbol{\gamma}) = \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}) + \lambda\|\boldsymbol{\gamma}\|_1. \tag{4.10}$$

We let $\boldsymbol{\theta}^*$ be the true value of parameter $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}^* = (\gamma_1^*, \ldots, \gamma_p^*)^T$ the true value of $\boldsymbol{\gamma}$. We define $S = \{j : \gamma_j^* \neq 0\} = \{j : \beta_j^* \neq 0\}$, and its complement $\bar{S} = \{j : \gamma_j^* = 0\} = \{j : \beta_j^* = 0\}$, where $s^* = |S| < n$. For

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

any vector $\boldsymbol{\xi} \in R^p$, $\boldsymbol{\xi}_S = \{v_j : j \in S\} \in R^{s^*}$. For the $p \times p$ Hessian matrix $\nabla^2 \mathcal{L}(\boldsymbol{\gamma})$, we write $\nabla^2_{SS} \mathcal{L}(\boldsymbol{\gamma})$ as the corresponding $s^* \times s^*$ sub-matrix with restrictions to the coordinates in S. Finally, the oracle estimator is defined as

$$\widehat{\boldsymbol{\gamma}}_{\mathrm{O}} = \underset{\mathrm{supp}(\boldsymbol{\gamma}) \subset S, \boldsymbol{\gamma} \in R^p}{\arg\min} \mathcal{L}(\boldsymbol{\gamma}).$$

4.2 Assumptions

In this subsection, we present the main assumptions that are necessary to derive our theoretical results. Our first assumption is on how to control the tail behavior of Y given $\boldsymbol{X}$.

Assumption 1. Assume that $\|\boldsymbol{X}\|_\infty < M < \infty$, $|\boldsymbol{X}^T \boldsymbol{\gamma}^*| < B < \infty$, and Y given $\boldsymbol{X}$ satisfies for any $\delta > 0$, $\Pr(|Y| \geq \delta | \boldsymbol{X}) \leq c_1 \exp(-c_2 \delta)$, where c_1 and c_2 are positive constants.

This assumption is similar to the Assumption 3.7 in Ning et al. (2017). As they verified, the sub-exponential tail assumption is satisfied for most commonly used GLMs in practice, for example, linear regression with Gaussian noise and logistic regression.

We need some conditions on the extreme sparse eigenvalues of a matrix $\mathbf{M}$.

Definition 1. Let s be a positive integer. The largest and smallest s-sparse

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

eigenvalues of a p-dimensional squared matrix $\mathbf{M}$ are

$$\rho_+(\mathbf{M}, s) = \sup\{\boldsymbol{v}^T\mathbf{M}\boldsymbol{v} : \|\boldsymbol{v}\|_0 \leq s, \|\boldsymbol{v}\|_2 = 1\},$$

$$\rho_-(\mathbf{M}, s) = \inf\{\boldsymbol{v}^T\mathbf{M}\boldsymbol{v} : \|\boldsymbol{v}\|_0 \leq s, \|\boldsymbol{v}\|_2 = 1\}.$$

Since we frequently use $\rho_+(\nabla^2\mathcal{L}(\boldsymbol{\gamma}^*), s)$ and $\rho_-(\nabla^2\mathcal{L}(\boldsymbol{\gamma}^*), s)$ in the following derivation, we abbreviate them as $\rho_+(s) = \rho_+(\nabla^2\mathcal{L}(\boldsymbol{\gamma}^*), s)$ and $\rho_-(s) = \rho_-(\nabla^2\mathcal{L}(\boldsymbol{\gamma}^*), s)$.

Assumption 2. There exists positive constants ρ_* and ρ^*, such that

$$\rho_* \leq \rho_-(s) \leq \rho_+(s) \leq \rho^*.$$

The sparse eigenvalue conditions are usually proposed to bound the estimation error in high dimensional problems. Similar concepts, although in slightly different forms, have been defined and studied in Bickel et al. (2009); Ning et al. (2017); Yang et al. (2014). In the Appendix, we verify that Assumption 2 is satisfied with probability at least $1 - C_1p^2\exp(-C_2n/s^2)$ for the most commonly used GLMs, linear regression with Gaussian noise and logistic regression. The definitions of C_1 and C_2 are in the Appendix.

The theory presented in this Section applies not only to the SCAD and the MCP penalties, but also to a broad class of nonconvex penalties. We rely on regularity conditions for penalty functions.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

Assumption 3. For $p_\lambda(t)$ or $q_\lambda(t)$ or their first order derivatives

(a) $q_\lambda(-t) = q_\lambda(t)$ for any t, and $q_\lambda(0) = 0$;

(b) for $t' > t$, there exist two constants $\zeta_- \geq 0$ and $\zeta_+ \geq 0$ such that

$$-\zeta_- \leq \frac{q'_\lambda(t') - q'_\lambda(t)}{t' - t} \leq -\zeta_+ \leq 0;$$

(c) $|q'_\lambda(t)| \leq \lambda$ for any t, and $q'_\lambda(0) = 0$;

(d) $q'_\lambda(t)$ has bounded difference with respect to λ: $|q'_{\lambda_1}(t) - q'_{\lambda_2}(t)| \leq |\lambda_1 - \lambda_2|$ for any t;

(e) There exist $c_7 \in [0, 1]$ and $c_8 \in (0, \infty)$ such that $p'_\lambda(t) \geq c_7\lambda$ for $t \in (0, c_8\lambda]$;

(f) $p'_\lambda(t) = 0$ once $|t| > \nu > c_9\sqrt{\frac{\log p}{n}}$ for some positive constant c_9.

The assumptions presented here are similar to those of Wang et al. (2014); Yang et al. (2014). In (b), ζ_- and ζ_+ are two parameters that control the concavity of $q_\lambda(t)$. Taking $t' \to t$ in (b), we have $q''_\lambda(t) \in [-\zeta_-, -\zeta_+]$, which suggests that larger ζ_- and ζ_+ allow $q_\lambda(t)$ to be more concave. For example, in SCAD we have $\zeta_- = 1/(a-1)$ with some $a > 2$ and $\zeta_+ = 0$, and in MCP we have $\zeta_- = 1/a$ with some $a > 0$ and $\zeta_+ = 0$. In Wang et al. (2014), they found that all these conditions hold for both SCAD and MCP.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

In some of our following derivations, we also need a relation between the concavity parameter ζ_- and $\rho_-(\nabla^2\mathcal{L}, 2s^*)$, the smallest $(2s^*)$-sparse eigenvalue of the Hessian matrix $\nabla^2\mathcal{L}$.

Assumption 4. The concavity parameter ζ_- defined in the conditions for the penalty function satisfies

$$\zeta_- \leq c_{10}\rho_-(\nabla^2\mathcal{L}, 2s^*),$$

with some constant $c_{10} < 1$.

Since in fact $\zeta_+ \leq \zeta_-$ and $\rho_-(\nabla^2\mathcal{L}, 2s^*) \leq \rho_+(\nabla^2\mathcal{L}, 2s^*)$, this restriction implies that $\zeta_+ \leq c_{10}\rho_+(\nabla^2\mathcal{L}, 2s^*)$. Theoretically, for each penalty, these two restrictions are satisfied by going through the verification of the Assumption 2 and appropriately choosing the t and ρ_*, ρ^* values.

4.3 Main Results

Our main objective in this subsection is to show that the estimator from our proposed method, $\widehat{\boldsymbol{\gamma}}$, has the same support as the true value $\boldsymbol{\gamma}^*$, also as $\boldsymbol{\beta}^*$, thus variable selection consistency holds. A sequence of results are presented. The first result shows that the true value of $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}^*$, minimizes $\mathrm{E}(\mathcal{L}(\boldsymbol{\gamma}))$, with $\mathcal{L}(\boldsymbol{\gamma})$ as in (2.7). This result provides the intuition as to why $\mathcal{L}(\boldsymbol{\gamma})$ is a legitimate loss function.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

Lemma 1. *We have* $\mathrm{E}(\nabla\mathcal{L}(\boldsymbol{\gamma}^*)) = 0$ *and* $\boldsymbol{\gamma}^*$ *is a global minimizer of* $\mathrm{E}(\mathcal{L}(\boldsymbol{\gamma}))$*, where* $\mathrm{E}(\cdot)$ *is the expectation under the true parameter* $\boldsymbol{\theta}^*$*.*

Now $\nabla\mathcal{L}(\boldsymbol{\gamma})$ has a second-order U-statistic structure. Our second result concerns the concentration inequality for U-statistics with a sub-exponential kernel function. We only present the result for second-order U-statistics. In Ning et al. (2017) and Yang et al. (2014), the authors have a more general concentration inequality.

Lemma 2. *Let* $X_1, \ldots, X_n$ *be independent random variables. Consider the U-statistics of order 2,*

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i_1 < i_2 \le n} u(X_{i_1}, X_{i_2}),$$

where $\mathrm{E}\{u(X_{i_1}, X_{i_2})\} = 0$ *for all* $i_1 < i_2$*. If there exist constants* L_1 *and* L_2 *such that*

$$\Pr(|u(X_{i_1}, X_{i_2})| \ge x) \le L_1 \exp(-L_2 x),$$

for all $i_1 < i_2$ *and all* $x \ge 0$*, then*

$$\Pr(|U_n| \ge x) \le 2\exp\left[-\min\left\{\frac{L_2^2 x^2}{8L_1^2}, \frac{L_2 x}{4L_1}\right\} k\right],$$

where $k = \lfloor n/2 \rfloor$ *is the largest integer less than* $n/2$*.*

The proofs of these results are in the literature, see Ning et al. (2017); Yang et al. (2014), so they are omitted. The next result controls the magnitude of $\|\nabla\mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

Lemma 3. *Given Assumption 1, we have*

$$\|\nabla\mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty \leq C_3\sqrt{\log p/n},$$

with probability at least $1-\delta_1$*, where* $\delta_1 = 2p\exp\left[-\min\left\{C_4\log p, C_5 n^{1/2}(\log p)^{1/2}\right\}\right]$*,* C_3 *is a positive constant,* C_4 *and* C_5 *are constants detailed in the Appendix.*

Based on the magnitude of $\|\nabla\mathcal{L}(\boldsymbol{\gamma}^*)\|_\infty$, we can provide a bound for the difference between the truth $\boldsymbol{\gamma}^*$ and the oracle estimator $\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}$, as follows.

Lemma 4. *Given Assumption 1 and that* $\|\nabla^2_{SS}\mathcal{L}(\boldsymbol{\gamma}^*)^{-1}\|_{L_\infty} < C$*,* $\log(n)(s^*)^2\sqrt{\frac{\log p}{n}} = o(1)$*, we have*

$$\|\widehat{\boldsymbol{\gamma}}_{\mathrm{O}} - \boldsymbol{\gamma}^*\|_\infty < 2CC_3\sqrt{\frac{\log s^*}{n}},$$

with probability at least $1-\delta_2$*, where* $\delta_2 = 2s^*\exp\left[-\min\left\{C_4\log s^*, C_5 n^{1/2}(\log s^*)^{1/2}\right\}\right] + c_{12}p^{-1} + c_1 n^{-1}$*.*

We present a characteristic of our surrogate loss function that, for the coordinates in $\bar{S}$, the cardinality of the support set of $\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2$ is bounded by the true number of "important" variables, which bounds the false positive magnitude.

Lemma 5. *Given Assumption 3, if* $\boldsymbol{\gamma}_1$ *and* $\boldsymbol{\gamma}_2$ *are two* p*-dimensional sparse vectors that satisfy* $\|(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2)_{\bar{S}}\|_0 \leq s^*$*, then the surrogate loss function satisfies the restricted strong convexity*

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_2) \geq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_1) + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_1)^T(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) + \frac{\rho_-(\nabla^2\mathcal{L}, 2s^*) - \zeta_-}{2}\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2^2,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

and the restricted strong smoothness

$$\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_2) \leq \widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_1) + \nabla\widetilde{\mathcal{L}}_\lambda(\boldsymbol{\gamma}_1)^T(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1) + \frac{\rho_+(\nabla^2\mathcal{L}, 2s^*) - \zeta_+}{2}\|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1\|_2^2.$$

Finally, we present our variable selection consistency result. We achieve this goal by showing the support set of our proposed estimator and that of the oracle estimator are the same as that of the true parameter.

Theorem 1. *If Assumptions 1, 2, 3, 4 hold,* $\|\nabla^2_{SS}\mathcal{L}(\boldsymbol{\gamma}^*)^{-1}\|_{L_\infty} < C$*, where* C *is a positive constant specified in Lemma 4,* $\log(n)(s^*)^2\sqrt{\frac{\log p}{n}} = o(1)$*, and the weakest signal strength satisfies* $\min_{j\in S}|\gamma_j^*| > 2\nu > 2\lambda$*, where* $\lambda \asymp \sqrt{\log p/n}$*. Then, when* n *is sufficiently large, we have* $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}}_{\mathrm{O}}$*, and hence*

$$supp(\widehat{\boldsymbol{\gamma}}) = supp(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}}) = supp(\boldsymbol{\gamma}^*),$$

with probability at least $1 - \delta_1 - \delta_2 - \delta_3$*, where* δ_1 *is defined in Lemma 3,* δ_2 *is defined in Lemma 4,* $\delta_3 = C_1p^2\exp(-C_2n/(s^*)^2)$ *comes from the Assumption 2.*

Remark 1. In Theorem 1, the lower bound of the high probability comes from those in Assumption 2, Lemma 3 and Lemma 4. To be more specific, using Lemma 4, in equation (1) in the proof, we show that $|(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}})_j| > \nu$ with probability at least $1 - \delta_2$; using Lemma 3, in equation (2) in the proof, we have $\|\nabla\mathcal{L}(\widehat{\boldsymbol{\gamma}}_{\mathrm{O}})\|_\infty \leq C_3\sqrt{\log p/n}$ with probability at least $1 - \delta_1$; based on

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

the Assumption 2, we establish $\|\widehat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma}^*\|_2 \leq c_{14}\rho_*^{-1}\sqrt{s^*}\lambda$ in equation (3) with probability at least $1-\delta_3$. The final lower bound of the high probability comes from the combination of the three and the fact that $P(A \bigcap B \bigcap C) \geq P(A) + P(B \bigcap C) - 1 \geq P(A) + P(B) + P(C) - 2 \geq 1 - \delta_1 - \delta_2 - \delta_3$ where A, B and C are three arbitrary events, and $P(A) \geq 1 - \delta_1$, $P(B) \geq 1 - \delta_2$, $P(C) \geq 1 - \delta_3$.

Remark 2. With respect to the high dimensional set-up, we allow both $\log p$, the logarithm of the dimensionality, and s^*, the number of nonzero components in the original parameter $\boldsymbol{\beta}$, to grow with n. From Theorem 1 and its proof, the condition $\log p$ and s^* need to be satisfied is that $\log(n)(s^*)^2\sqrt{\frac{\log p}{n}} = o(1)$. This implies that if $s^* = o(n^{\varsigma})$ for some $0 < \varsigma < 1/4$, then $\log p = o(n^{1-4\varsigma}/(\log n)^2)$. Here we follow the most recent literature for the definition of high dimensionality. For example, in Fan and Lv (2011), the high dimensionality refers to $\log p = O(n^{\alpha})$, for some $0 < \alpha < 1$. Here we have $\log p = o(n^{1-4\varsigma}/(\log n)^2)$. In the high-dimensional GLM that we consider, the number of covariates p can grow at most exponentially fast with n, the sample size of the completely observed subjects.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

5. Simulation Studies

The objective of our simulation studies is two-fold. First, we evaluate the finite sample performance of our proposed method by examining two commonly used models: linear regression and logistic regression, and three representative penalty functions: LASSO, SCAD and MCP. Second, we compare our proposed method to two existing methods: one assuming that there is no missing data, and the other assuming the missing data mechanism is MAR.

In all of our eight simulation settings (S1)–(S8), we generated the covariate $\boldsymbol{X}$ from p-dimensional $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{ij} = \rho^{|i-j|}$, and we considered $\rho \in \{0, 0.5\}$. Here $b(\eta) = \eta^2/2$ corresponds to linear regression and $b(\eta) = \log(1 + e^{\eta})$ corresponds to logistic regression.

Our simulation settings (S1)–(S4) were as follows:

(S1): $b(\eta) = \eta^2/2$, with $\eta = \alpha + \boldsymbol{\beta}^T \boldsymbol{X}$, $\alpha = 0$, $\boldsymbol{\beta} = (3, 1.5, 0.5, 0, \ldots, 0)^T$, the dispersion parameter $\phi = 1$, $s^* = 3$, $p = 8$ and $N = 200$. The missing data mechanism $\Pr(R = 1|Y, \boldsymbol{X}) = I_{\{Y>\gamma_1\}} I_{\{X_1>\gamma_2\}}$ with $\gamma_1 = -3.3$, $\gamma_2 = -0.4$ for $\rho = 0$ and $\gamma_1 = -3.8$, $\gamma_2 = -0.3$ for $\rho = 0.5$.

(S2): same as (S1) except that $p = 200$, $\gamma_1 = -2.8$, $\gamma_2 = -0.4$ for $\rho = 0$ and $\gamma_1 = -4.1$, $\gamma_2 = -0.3$ for $\rho = 0.5$.

(S3): $b(\eta) = \log(1+e^{\eta})$, with $\eta = \alpha + \boldsymbol{\beta}^T \boldsymbol{X}$, $\alpha = 0$, $\boldsymbol{\beta} = (2, -2, 1, -1, 0, \ldots, 0)^T$,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

$s^* = 4$, $p = 8$ and $N = 500$. The missing data mechanism $\Pr(R = 1|Y, \boldsymbol{X}) = I_{\{X_1>\gamma\}} \cdot (2Y + 3)/5$ with $\gamma = -0.7$ for either $\rho = 0$ or $\rho = 0.5$.

(S4): same as (S3) except that $p = 500$.

The purpose of the different choices of γ values was to guarantee that, in each setting, the observed proportion was about 60% to 65%. We report the results based on 100 replications in each setting, with false positive (FP) as the one with true zero value but falsely estimated as nonzero; and false negative (FN) as the one with true nonzero value but falsely estimated as zero. We counted the number of false positives (#FP) and the number of false negatives (#FN) and report them in a boxplot in each setting in Figures 1–4, respectively. We also list the mean and standard deviation (SD) of #FP and #FN for each setting in Tables 1–2 for linear regression and logistic regression, respectively.

Some conclusions can be reached from simulation studies (S1)–(S4). First, in almost all scenarios, our proposed method outperforms the method assuming MAR in terms of smaller FP and FN mean/median values. Second, in most scenarios, the method with no missing data, treated as a gold standard, outperforms our proposed method. Third, the nonconvex penalties almost always perform better than the LASSO penalty in terms of variable selection, which is consistent with the previous literature.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

Under the assumption (2.3), the method assuming MAR produces biased estimators and hence worse results for variable selection, while the proposed estimator satisfies the variable selection consistency property and hence better (than the MAR method) variable selection performance is expected. Our numerical findings in (S1)–(S4) match well with the theory.

6. Discussion

One can observe that the proposed method only uses the information contained in the completely observed samples. In applications, there may exist many partially observed samples, for example, the covariate $\boldsymbol{X}$ values are always available. It is difficult to directly get these partially observed samples involved in the current proposed approach. Some imputation techniques, for example Chen and Wang (2013); Long and Johnson (2015); Liu et al. (2016), may be helpful and this warrants further study. How to conduct high dimensional statistical inference, especially the post-selection inference, is interesting but challenging when the data contain missing values. This is beyond the scope of our paper and certainly warrants further investigation.

Finally, we provide some practical guidance on using the proposed method. In reality, the missing data mechanism assumption is unverifi-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

able and its underlying truth is unknown. Our assumption (2.3) is more flexible than a single parametric assumption, and hence more generally applicable. From our data analyses in Section 6, the proposed method and the method assuming MAR will always have some agreement and some disagreement. Although we cannot reach a definite conclusion in practice, our proposed approach and analysis may provide some insight on the data, especially when the MAR assumption is suspect.

Supplementary Materials

The online supplementary material contains more detailed derivations of the results presented in this paper, some additional simulation results, and two data analyses.

Acknowledgements

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award Number UL1TR001412. The authors thank the editor, an associate editor and two anonymous referees for their constructive comments and insightful suggestions, which have led to a significantly improved paper.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

References

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Chan, K. C. G. (2013). Nuisance parameter elimination for proportional likelihood ratio models with nonignorable missingness and random truncation. *Biometrika 100*(1), 269–276.

Chen, Q. and S. Wang (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. *Statistics in Medicine 32*(21), 3646–3659.

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

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*(1), 101–148.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory 57*(8), 5467–5484.

Fan, J., L. Xue, and H. Zou (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics 42*(3), 819–849.

Fan, Y. and C. Y. Tang (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(3), 531–552.

Fang, F., J. Zhao, and J. Shao (2017). Imputation-based adjusted score equations in generalized

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

linear models with nonignorable missing covariate values. *Statistica Sinica*, to appear.

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

Garcia, R. I., J. G. Ibrahim, and H. Zhu (2010). Variable selection for regression models with missing data. *Statistica Sinica 20*(1), 149–165.

Hunter, D. R. and K. Lange (2004). A tutorial on mm algorithms. *The American Statistician 58*(1), 30–37.

Ibrahim, J. G., S. R. Lipsitz, and M.-H. Chen (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(1), 173–190.

Ibrahim, J. G., H. Zhu, and N. Tang (2008). Model selection criteria for missing-data problems using the em algorithm. *Journal of the American Statistical Association*, 1648–1658.

Kalbfleisch, J. D. (1978). Likelihood methods and nonparametric tests. *Journal of the American Statistical Association 73*(361), 167–170.

Kim, J. K. and J. Shao (2013). *Statistical Methods for Handling Incomplete Data*. CRC Press.

Liang, K.-Y. and J. Qin (2000). Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 773–786.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (2 ed.). Wiley.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

Liu, Y., Y. Wang, Y. Feng, and M. W. Melanie (2016). Variable selection and prediction with incomplete high-dimensional data. *The Annals of Applied Statistics 10*, 418–450.

Loh, P.-L. and M. J. Wainwright (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics 40*(3), 1637–1664.

Long, Q. and B. A. Johnson (2015). Variable selection in the presence of missing data: resampling and imputation. *Biostatistics 16*(3), 596–610.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2 ed.). Chapman & Hall/CRC.

Ning, Y., T. Zhao, and H. Liu (2017). A likelihood ratio framework for high-dimensional semiparametric regression. *The Annals of Statistics 45*(6), 2299–2327.

Robins, J. M. and Y. Ritov (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine 16*(3), 285–319.

Shao, J. and J. Zhao (2013). Estimation in longitudinal studies with nonignorable dropout. *Statistics and Its Interface 6*, 303–313.

Tang, G., R. J. Little, and T. E. Raghunathan (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika 90*(4), 747–764.

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

Wang, Z., H. Liu, and T. Zhang (2014). Optimal computational and statistical rates of conver-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

gence for sparse nonconvex learning problems. *The Annals of Statistics 42*(6), 2164–2201.

Yang, Z., Y. Ning, and H. Liu (2014). On semiparametric exponential family graphical models. *arXiv preprint arXiv:1412.8697*.

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

Zhao, J. (2017). Reducing bias for maximum approximate conditional likelihood estimator with general missing data mechanism. *Journal of Nonparametric Statistics 29*, 577–593.

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

Zhao, J. and J. Shao (2017). Approximate conditional likelihood for generalized linear models with general missing data mechanism. *Journal of System Science and Complexity 30*(1), 139–153.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics 36*(4), 1509–1566.

Jiwei Zhao

Department of Biostatistics,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

State University of New York at Buffalo,

Buffalo, NY 14214, U.S.A.

E-mail: zhaoj@buffalo.edu

Yang Yang

Department of Biostatistics,

State University of New York at Buffalo,

Buffalo, NY 14214, U.S.A.

E-mail: yyang39@buffalo.edu

Yang Ning

Department of Statistical Science,

Cornell University,

Ithaca, NY 14853, U.S.A.

E-mail: yn265@cornell.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

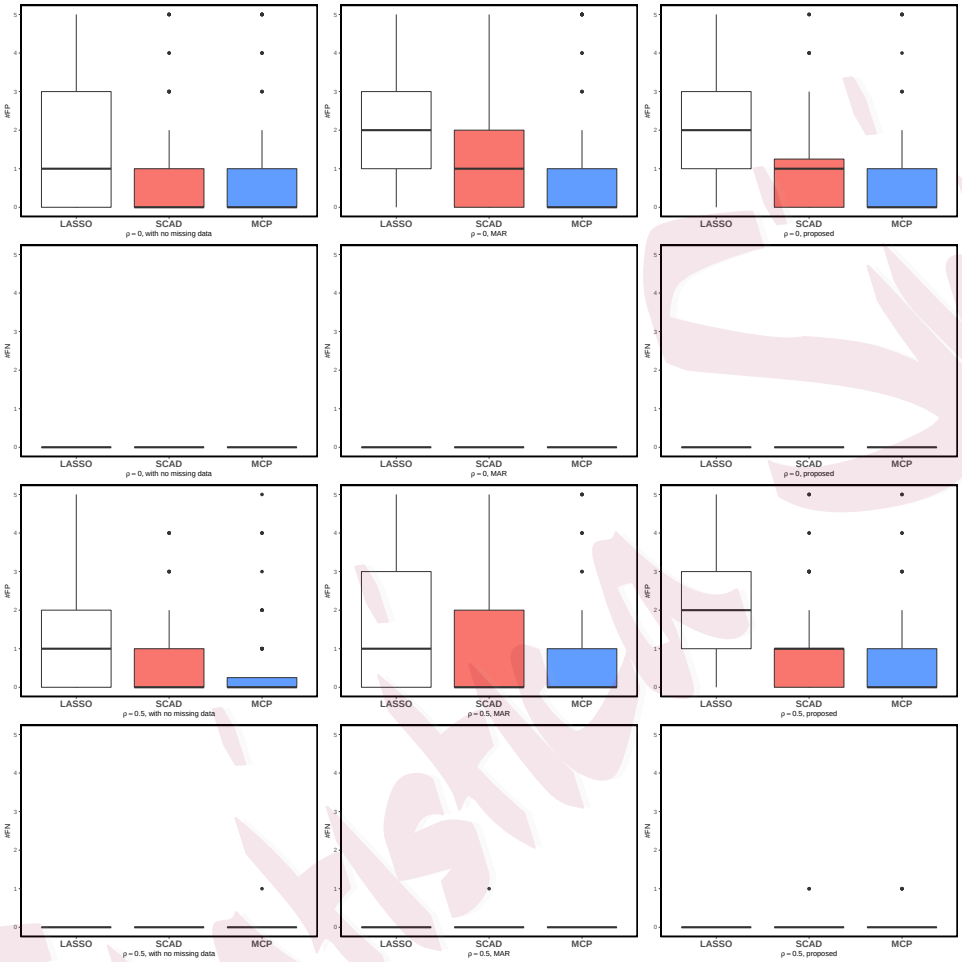

Figure 1: Boxplots of #FP and #FN in simulation setting (S1). The three columns represent the methods with no missing data, MAR and proposed, respectively. The first and third rows show #FP while the second and fourth rows show #FN. The first two rows are for the case with $\rho = 0$ and the last two rows are for the case with $\rho = 0.5$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

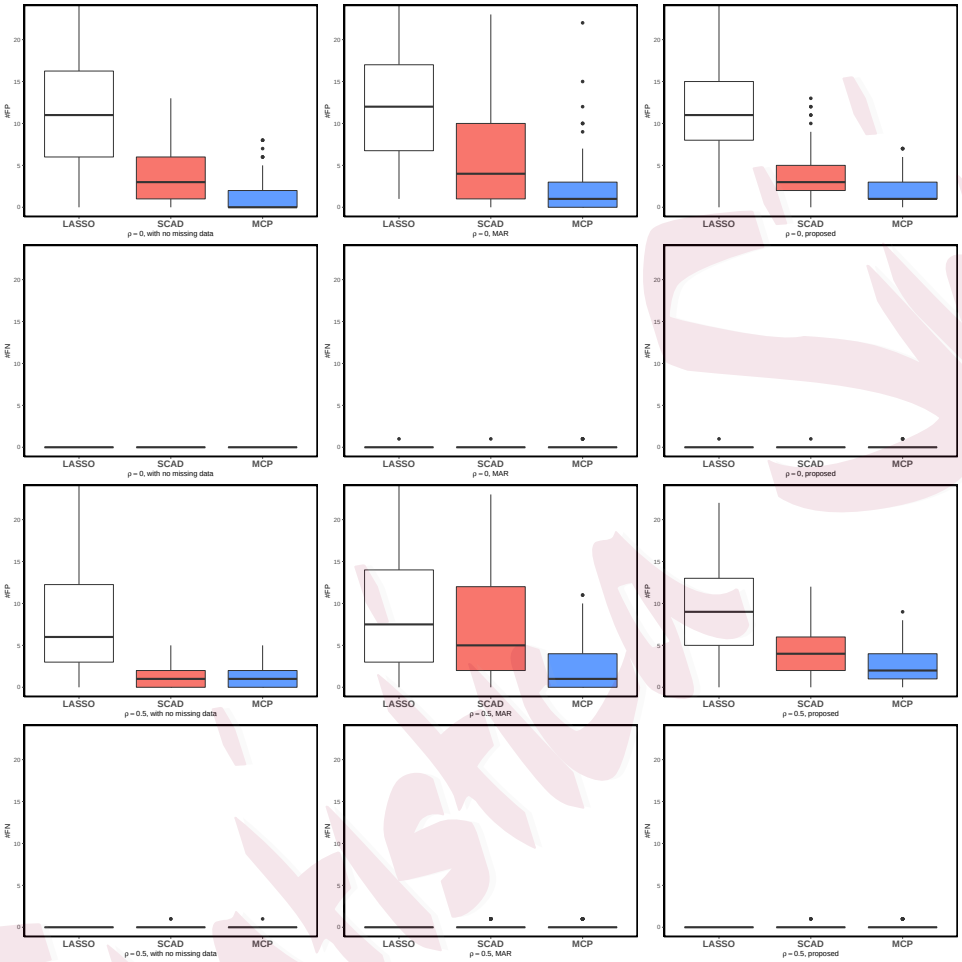

Figure 2: Boxplots of #FP and #FN in simulation setting (S2). The three columns represent the methods with no missing data, MAR and proposed, respectively. The first and third rows show #FP while the second and fourth rows show #FN. The first two rows are for the case with $\rho = 0$ and the last two rows are for the case with $\rho = 0.5$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

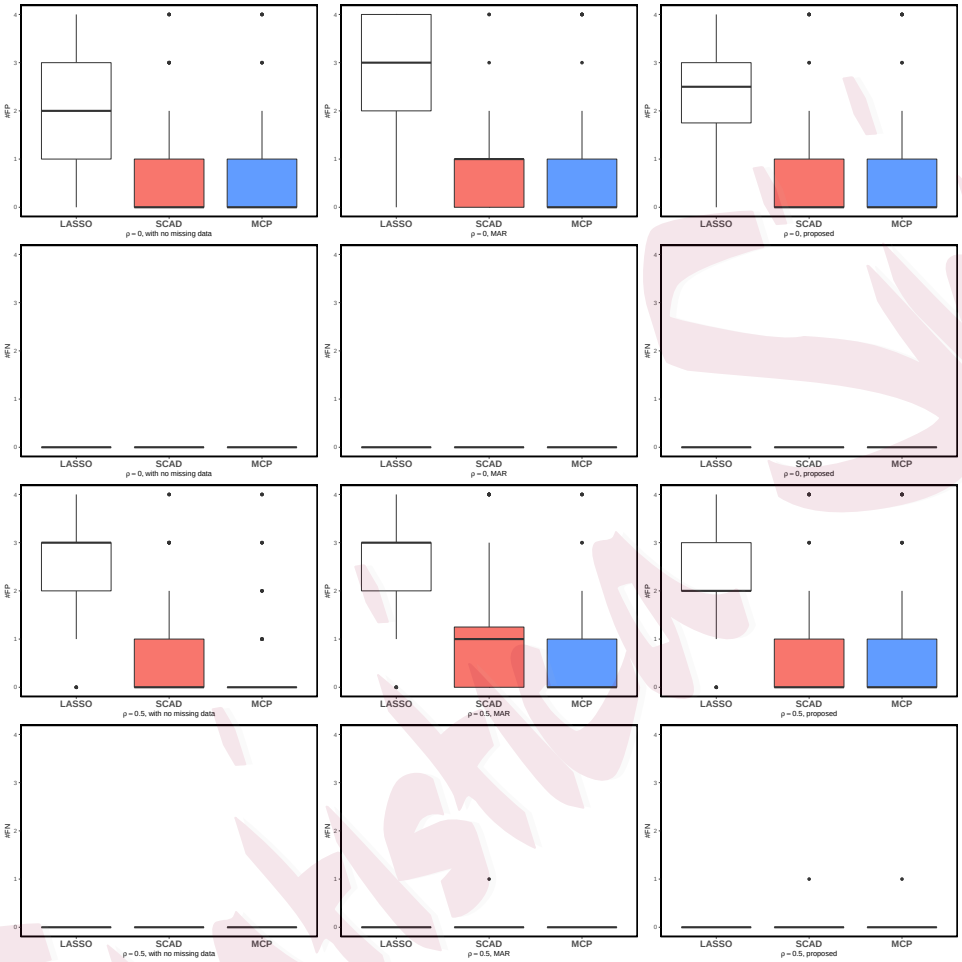

Figure 3: Boxplots of #FP and #FN in simulation setting (S3). The three columns represent the methods with no missing data, MAR and proposed, respectively. The first and third rows show #FP while the second and fourth rows show #FN. The first two rows are for the case with $\rho = 0$ and the last two rows are for the case with $\rho = 0.5$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

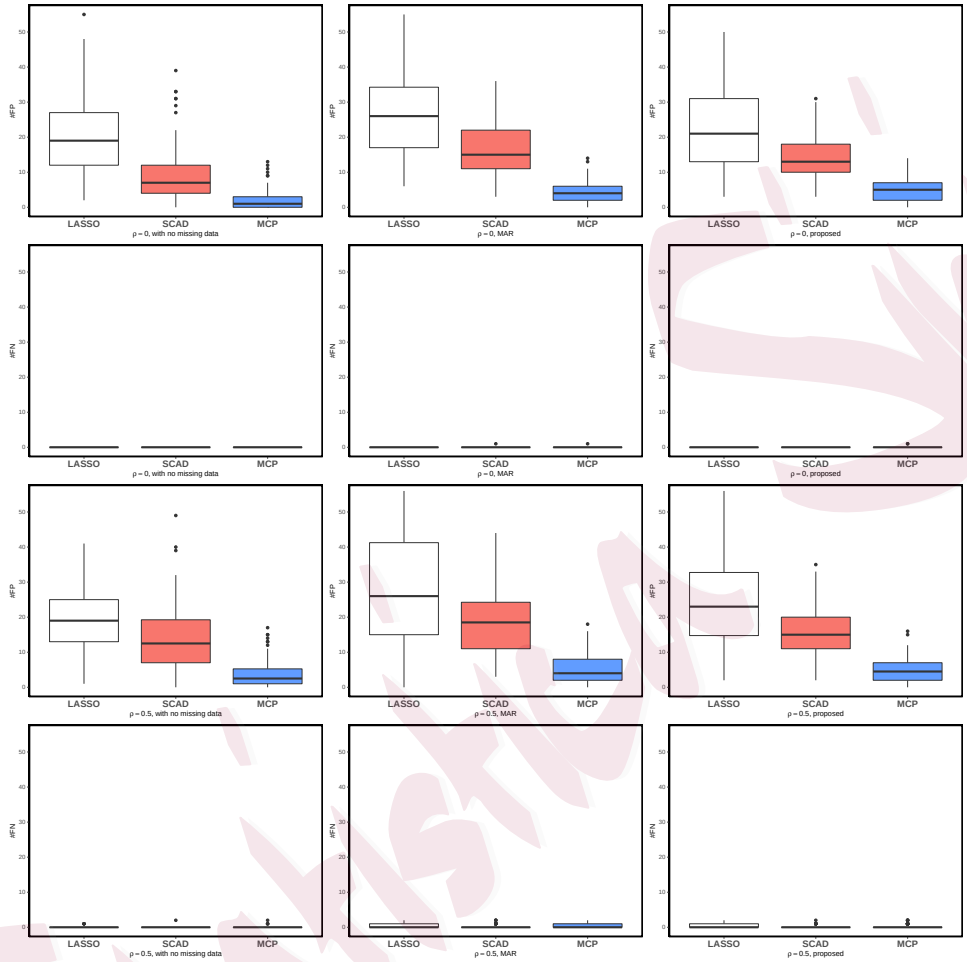

Figure 4: Boxplots of #FP and #FN in simulation setting (S4). The three columns represent the methods with no missing data, MAR and proposed, respectively. The first and third rows show #FP while the second and fourth rows show #FN. The first two rows are for the case with $\rho = 0$ and the last two rows are for the case with $\rho = 0.5$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

Table 1: Mean and standard deviation (SD; in parentheses) of #FP and #FN in simulation settings (S1)–(S2). The proposed method is compared to two other methods: the method with no missing data, which uses all simulated data; and the method assuming MAR, which uses completely observed samples only.

	Method	Penalty	$\rho = 0$		$\rho = 0.5$	
			#FP	#FN	#FP	#FN
p=8	with no missing data	LASSO	1.72 (1.57)	0 (0)	1.28 (1.39)	0 (0)
		SCAD	0.92 (1.36)	0 (0)	0.62 (1.06)	0 (0)
		MCP	0.73 (1.48)	0 (0)	0.45 (0.99)	0.01 (0.10)
	MAR	LASSO	2.50 (1.53)	0 (0)	1.76 (1.56)	0 (0)
		SCAD	1.30 (1.40)	0 (0)	0.93 (1.29)	0.01 (0.10)
		MCP	1.04 (1.59)	0 (0)	0.68 (1.29)	0 (0)
	proposed	LASSO	2.34 (1.39)	0 (0)	2.28 (1.33)	0 (0)
		SCAD	0.98 (1.25)	0 (0)	0.98 (1.22)	0.02 (0.14)
		MCP	0.78 (1.31)	0 (0)	0.63 (1.12)	0.04 (0.20)
p=200	with no missing data	LASSO	12.45 (9.90)	0 (0)	8.88 (9.54)	0 (0)
		SCAD	3.72 (3.22)	0 (0)	1.41 (1.28)	0.02 (0.14)
		MCP	1.38 (2.10)	0 (0)	1.09 (1.14)	0.01 (0.10)
	MAR	LASSO	14.06 (10.99)	0.01 (0.10)	9.89 (8.57)	0 (0)
		SCAD	6.57 (6.92)	0.01 (0.10)	7.13 (6.60)	0.14 (0.35)
		MCP	2.23 (3.59)	0.05 (0.22)	2.42 (2.81)	0.19 (0.39)
	proposed	LASSO	12.54 (6.83)	0.01 (0.10)	9.81 (5.84)	0 (0)
		SCAD	3.91 (2.86)	0.01 (0.10)	4.35 (2.58)	0.04 (0.20)
		MCP	1.96 (1.79)	0.03 (0.17)	2.71 (1.99)	0.10 (0.30)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0312

Table 2: Mean and standard deviation (SD; in parentheses) of #FP and #FN in simulation settings (S3)–(S4). The proposed method is compared to two other methods: the method with no missing data, which uses all simulated data; and the method assuming MAR, which uses completely observed samples only.

	Method	Penalty	$\rho = 0$		$\rho = 0.5$	
			#FP	#FN	#FP	#FN
p=8	with no missing data	LASSO	2.09 (1.13)	0 (0)	2.42 (1.12)	0 (0)
		SCAD	0.76 (1.09)	0 (0)	0.64 (1.10)	0 (0)
		MCP	0.52 (1.00)	0 (0)	0.40 (0.96)	0 (0)
	MAR	LASSO	2.72 (1.07)	0 (0)	2.56 (1.13)	0 (0)
		SCAD	0.81 (0.92)	0 (0)	1.03 (1.34)	0.01 (0.10)
		MCP	0.56 (1.07)	0 (0)	0.54 (1.01)	0 (0)
	proposed	LASSO	2.32 (1.16)	0 (0)	2.47 (1.10)	0 (0)
		SCAD	0.78 (1.08)	0 (0)	0.66 (1.09)	0.01 (0.10)
		MCP	0.65 (1.12)	0 (0)	0.58 (1.12)	0.01 (0.10)
p=500	with no missing data	LASSO	20.16 (10.93)	0 (0)	19.94 (8.28)	0.04 (0.20)
		SCAD	9.77 (8.15)	0 (0)	14.08 (9.40)	0.02 (0.20)
		MCP	2.28 (2.83)	0 (0)	3.78 (3.95)	0.05 (0.26)
	MAR	LASSO	27.91 (14.38)	0 (0)	27.65 (14.95)	0.46 (0.56)
		SCAD	16.87 (8.02)	0.01 (0.10)	18.77 (8.85)	0.22 (0.48)
		MCP	4.33 (3.31)	0.01 (0.10)	5.28 (4.28)	0.47 (0.70)
	proposed	LASSO	23.09 (13.16)	0 (0)	24.62 (13.96)	0.39 (0.53)
		SCAD	13.73 (5.88)	0 (0)	15.59 (6.76)	0.14 (0.38)
		MCP	4.84 (3.08)	0.02 (0.14)	4.85 (3.49)	0.22 (0.50)