Stability enhanced variable selection for a semiparametric model with flexible missingness mechanism and its application to the ChAMP study

Yang Yang^a, Jiwei Zhao , Gregory Wilding^b, Melissa Kluczynski and Leslie Bisson^c

^aAbbVie Inc., North Chicago, IL, USA; ^bDepartment of Biostatistics, School of Public Health and Health Professions, State University of New York at Buffalo, Buffalo, NY, USA; ^cDepartment of Orthopaedics, Jacobs School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, NY, USA

ABSTRACT

This paper is motivated by the analytical challenges we encounter when analyzing the ChAMP (Chondral Lesions And Meniscus Procedures) study, a randomized controlled trial to compare debridement to observation of chondral lesions in arthroscopic knee surgery. The main outcome, WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) pain score, is derived from the patient's responses to the questionnaire collected in the study. The major goal is to identify potentially important variables that contribute to this outcome. In this paper, the model of interest is a semiparametric model for the pain score. To address the missing data issue, we adopt a flexible missingness mechanism that is much more versatile in practice than a single parametric model. Then we propose a pairwise conditional likelihood approach to estimate the unknown parameter in the semiparametric model without the need of modeling its nonparametric counterpart nor the missingness mechanism. For variable selection, we apply a regularization approach with a variety of stability enhanced tuning parameter selection methods. We conduct comprehensive simulation studies to evaluate the performance of the proposed method. We also apply the proposed method to the ChAMP study to demonstrate its usefulness.

ARTICLE HISTORY

Received 23 October 2018
Accepted 10 August 2019

KEYWORDS

Missing data mechanism; pairwise conditional likelihood; semiparametric model; stability; variable selection; ChAMP study

1. Introduction

Data driven scientific discovery leads to an emerging demand for statistical methodologies to analyze data with complex structure in biomedical studies. Variable selection based on the regularization approach is one of the most evolving areas in the past two decades in these settings. The rationale behind the regularization approach is that it is usually believed that only a small number of variables are truly informative to the response variable while others are not, and the regularization provides a way of identifying these variables.

Although researchers have developed theoretical properties and computational methods for regularization procedure for commonly seen models, there are still many unsolved

problems in their application. For example, in the ChAMP (Chondral Lesions And Meniscus Procedures) study [2], which is a randomized controlled trial to compare debridement to observation of chondral lesions seen during arthroscopic knee surgery, we encounter several analytical challenges.

The main aim in the ChAMP study is to investigate the potential risk factors for the WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) pain score, a patient reported outcome (PRO). In many situations, the PRO has become one of the most commonly used clinical tools to assess patient's health status [7,33]. However, statistically, the PRO measures are usually skewed in distribution, bounded in a finite range, and may have multiple modes [6]. Hence, traditional models are often not applicable when one attempts to examine the relation between a given PRO and a set of independent variables.

The second problem pertains to missing data. The issue of missing data is a rule rather than an exception in the analysis of PRO, and it has consistently drawn researchers' attention in the literature, see, e.g. [9,15,24,26]. The traditional missing data methodologies [20], such as multiple imputation [25] and sensitivity analysis [27], have been applied to PRO analysis and they have been shown to be useful in many situations. In the ChAMP study, all the patients were followed up for 1 year, with the primary outcome being the pain score 1 year after surgery. The loss to follow up is most likely due to complete resolution of pain, but it could also be increased pain. Using the standard missing data terminology, it is very plausible that such missingness is nonignorable, in that missing data may be strongly related to unobserved variables. For nonignorable missingness mechanism, a strong parametric model is restrictive and is easily misspecified. In this paper, we propose a mechanism model which only needs an assumption at a lowest level so that the model misspecification could be minimized.

The third challenge lies in the variable selection using the regularization approach. This could be regarded as one of the most evolving areas in modern statistical research with a vast literature. The basic idea is to assume that there only exists a small number of variables that are truly informative, i.e. the sparsity assumption, and then impose a penalty function onto the unpenalized objective to achieve this sparsity, with a to-be-determined tuning parameter to balance the weights between the penalty function and the unpenalized objective. Popular penalty functions include the LASSO (Least Absolute Shrinkage and Selection Operator [31]) and its various extensions such as the adaptive LASSO [41], and nonconvex counterparts such as the SCAD (Smoothly Clipped Absolute Deviation [11]) and the MCP (Minimax Concave Penalty [36]). Most of the time, one adopts the standard cross validation procedure or some information criterion to determine the tuning parameter. In the ChAMP study, since we have to establish a statistical model for the PRO and also need to tackle the nonignorable missing data issue, it is not clear whether and how we could use the standard penalized likelihood approach. Furthermore, it is also not clear whether we could directly use currently existing tuning parameter selection methods, and how we could improve those methods in our case.

In this paper, we address these problems in the context of the ChAMP study. Our contributions can be summarized in the following three aspects. First, we propose a semi-parametric model, the proportional likelihood ratio model [22], for the outcome variable which is skewed, lies in a finite range and may have multiple modes. This model makes less stringent assumptions than a parametric model and we find that it is an effective

model for the PRO. Second, we assume a very flexible missing data mechanism. This mechanism contains a variety of ignorable and nonignorable missingness scenarios. Third, motivated by the stability criterion well advocated for massive data computation [23,35], we propose novel stability enhanced tuning parameter selection methods in our penalized likelihood framework. We show that these methods perform better than those established alternatives.

The structure of the paper is as follows. In Section 2, we introduce our semiparametric model, the missing data mechanism and how the regularization approach can be adopted in our framework. Section 3 briefly summarizes the popular tuning parameter selection methods and proposes some stability enhanced tuning parameter selection methods. We conduct comprehensive simulation studies in Section 4 to demonstrate the effectiveness of our proposed method. We analyze the ChAMP trial data in Section 5. The paper is concluded with a discussion in Section 6.

2. Methodology

2.1. A semiparametric model for the WOMAC pain score

We first introduce a semiparametric model for the WOMAC pain score. As we discussed in Section 1, the WOMAC pain score data are usually skewed in distribution, bounded in a finite range, and may have multiple modes. Various simplistic parametric models were fitted and compared for this type of PRO in [1], however, those stringent parametric assumptions are less likely to be correct for real application.

In this paper, we consider the following semiparametric model for the WOMAC pain score. This model was first studied in [22] and is referred to as the proportional likelihood ratio (PLR) model. This model features a p -dimensional unknown parameter β and a unknown function $g(\cdot)$, the density of some unknown baseline distribution function $G(\cdot)$ with respect to a dominating measure. Given both, the PLR model assumes that the conditional pdf of the outcome Y given the p -dimensional covariate X is

$$p(y | x; \beta) = \frac{\exp(\beta^T xy)g(y)}{\int \exp(\beta^T xy) dG(y)}. \quad (1)$$

It can be seen that the unknown function $g(y)$, to be estimated by the data, will capture the intrinsic characteristics of the PRO variable Y , such as the skewness, boundedness and mode multiplicity; therefore, this model is much less stringent than a single parametric model, hence much more appropriate for the analysis of PRO. Also, same as the simple parametric linear regression model, this semiparametric model still shares the same convenient interpretation for the relation between Y and X . In (1), the only component with both x and y in it is $\exp(\beta^T xy)$, parametrically specified, while the y -only term and the x -only term are both left as unspecified. Therefore, the parameter β solely determines the association between Y and X . If $\beta = 0$, Y and X are independent.

This semiparametric model (1) is general and it contains many commonly used parametric models as special cases. For example, any generalized linear model (GLM) can be written as a special case of the PLR model with appropriate parametrization of β and $g(\cdot)$. Let the parameter β in PLR model be $\gamma/a(\tau)$ and $g(y)$ be $\exp\{\alpha\gamma/a(\tau) + c(y; \tau)\}$, we can

write the model in (1) as

$$p(y | \mathbf{x}; \boldsymbol{\beta}) = \exp \left\{ \frac{\boldsymbol{\theta}y - b(\boldsymbol{\theta})}{a(\tau)} + c(y; \tau) \right\},$$

where $\boldsymbol{\theta} = \boldsymbol{\alpha} + \boldsymbol{\gamma}^T \mathbf{x}$. This is actually the standard GLM with canonical link, and $a(\tau)$ is the positive dispersion parameter. More detailed discussion on the connection of the PLR model with other models can be found in [22].

2.2. Missing data mechanism

To model the missingness mechanism, we introduce the random variable R , which only takes two values: $R = 1$ means that there is no missing data and it is a fully observed subject; $R = 0$ implies a partially observed subject. We consider the data $\{r_i, y_i, \mathbf{x}_i\}$, $i = 1, \dots, N$, as independent and identically distributed copies of random variables $(R, Y, \mathbf{X})$. Without loss of generality, we let the first n subjects be fully observed.

To address missing data, one of the most difficult and foremost steps is to propose a correct model for $\text{pr}(R = 1 | y, \mathbf{x})$, the missing data mechanism. The simplistic way is to impose a parametric model; however, it has two major limitations. First, a given parametric assumption may not be robust to model misspecification. Second, in the most practical cases, the missing mechanism is nonignorable, and the underlying mechanism is unknown and is very hard to be verified. Therefore it is necessary to propose a missing data mechanism that is as robust and flexible as possible. In this paper, we impose a general assumption

$$\text{pr}(R = 1 | y, \mathbf{x}) = s(y)t(\mathbf{x}), \quad (2)$$

where $s(\cdot)$ and $t(\cdot)$ are merely some unspecified functions. The essence of this assumption is that it only assumes $\text{pr}(R = 1 | y, \mathbf{x})$ to be a product of an $\mathbf{x}$ -only function and a y -only function. An exciting feature, which we will show in Sections 2.3 and 2.4, is that, there is no need to know or even estimate the exact forms of $s(\cdot)$ or $t(\cdot)$, hence our method is robust to misspecification of the $s(\cdot)$ or $t(\cdot)$ function.

The assumption (2) is very general and it includes many special cases in the literature. These situations include missing response, missing covariate, or missing both. For example, when only the response Y has missing values and $t(\cdot)$ is constant, it corresponds to the nonignorable nonresponse assumption considered in [28,30,38]; when only the covariate $\mathbf{X}$ has missing values and $s(\cdot)$ is constant, then it is the nonignorable missing covariate assumption considered by Fang *et al.* [13].

Assumption (2) was explored in the literature in different aspects. Chan [8] considered the invariance properties under this assumption; Zhao [37] considered an improved estimation procedure with reduced bias using a resampling technique; and Zhao and Shao [39] studied the identifiability conditions in a GLM with non-canonical link under this assumption.

2.3. Pairwise conditional likelihood and identifiability

In the model introduced in Section 2.1 and the assumption in Section 2.2, four quantities $\boldsymbol{\beta}$, $g(\cdot)$, $s(\cdot)$ and $t(\cdot)$ are all unknown. We are not aiming to estimate all of them. Our foremost

aim is to identify informative variables for the WOMAC pain score, hence we introduce a pairwise conditional likelihood strategy to only estimate β so that the variable selection can still be performed without the need of knowing the nuisance nonparametric components $g(\cdot)$, $s(\cdot)$ or $t(\cdot)$. In other words, our β estimation method is robust to the misspecification of all other nuisance $g(\cdot)$, $s(\cdot)$ and $t(\cdot)$.

The difficulty to incorporate the assumption (2) into the likelihood framework is that $s(\cdot)$ and $t(\cdot)$ are both unknown functions. Note that

$$p(Y | X, R = 1) = \frac{\text{pr}(R = 1 | Y, X)}{w(X)} p(Y | X) = \frac{s(Y)t(X)}{w(X)} p(Y | X), \quad (3)$$

where $w(X) = \int \text{pr}(R = 1 | Y, X) p(Y | X) dY = \text{pr}(R = 1 | X)$. This means, the conditional pdf based on the completely observed subjects $p(Y | X, R = 1)$ equals the original pdf of Y given X , $p(Y | X)$, multiplies a ratio, and similar to the mechanism $\text{pr}(R = 1 | Y, X) = s(Y)t(X)$, this ratio is still a multiplier of an X -only function and a Y -only function.

Hence, if we only concentrate on the first n completely observed subjects, decompose the data $\{y_1, \dots, y_n\}$ into its rank statistic and its order statistic $\{y_{(1)}, \dots, y_{(n)}\}$, consider the conditional pdf of its rank statistic given its order statistic, and follow the idea of conditional likelihood [17], we can obtain the following likelihood function for β :

$$p(y_1, \dots, y_n | r_1 = \dots = r_n = 1, \mathbf{x}_1, \dots, \mathbf{x}_n, y_{(1)}, \dots, y_{(n)}) = \frac{\prod_{i=1}^n p(y_i | \mathbf{x}_i; \beta)}{\sum_c \prod_{i=1}^n p(y_{(i)} | \mathbf{x}_i; \beta)}, \quad (4)$$

where the summation in the denominator corresponds to all possible permutations of $\{1, \dots, n\}$. We can see that all the nuisance functions $s(\cdot)$, $t(\cdot)$ and $w(\cdot)$ are all cancelled out.

Although (4) only uses data from the completely observed subjects, we reiterate here that it is totally different from the so-called complete case analysis. In general, the complete case analysis studies $\prod_{i=1}^n p(y_i | \mathbf{x}_i; \beta)$ which does not incorporate the missingness mechanism (2) and is in general regarded as biased in our context.

To proceed with (4), there is a tremendous computational burden with an order of $n!$. To reduce the computational burden to n^2 [40], we consider a pairwise fashion of the conditional likelihood:

$$\prod_{1 \leq i < j \leq n} \frac{p(y_i | \mathbf{x}_i; \beta) p(y_j | \mathbf{x}_j; \beta)}{p(y_i | \mathbf{x}_i; \beta) p(y_j | \mathbf{x}_j; \beta) + p(y_i | \mathbf{x}_j; \beta) p(y_j | \mathbf{x}_i; \beta)}. \quad (5)$$

Plugging the pdf of the PLR model (1) into (5), both the $g(y)$ in its numerator and the denominator $\int \exp(\beta^T \mathbf{x} y) dG(y)$ are cancelled out. Thus the negative part of the log version of (5), after adding a normalizing constant, can be written as

$$\mathcal{L}(\beta) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \log\{1 + \exp(-\beta^T (\mathbf{x}_i - \mathbf{x}_j)(y_i - y_j))\}. \quad (6)$$

A very nice feature of this approach is that the nonparametric component $g(y)$ in the PLR model cancels out and there is no need to estimate it. The unknown parameter β can be estimated from minimizing the objective function (6).

The identifiability issue for nonignorable missing data is notorious. In general for parametric models, the identifiability conditions have to be derived on a case-by-case basis. In this paper, due to the consideration of modeling the WOMAC pain score, we propose a semiparametric model containing unknown β and $g(\cdot)$ in Section 2.1 and introduce a partially specified missingness mechanism containing unknown $s(\cdot)$ and $t(\cdot)$ in Section 2.2. Here, we adopt an approach of only identifying and estimating β without the knowledge of other nuisance components $g(\cdot)$, $s(\cdot)$ and $t(\cdot)$, so that variable selection relied on β can still be carried out without the identification or specification of all other components. Since our goal here is the variable selection, the proposed method is robust to any misspecification of all those $g(\cdot)$, $s(\cdot)$ and $t(\cdot)$ nuisance components.

2.4. Regularization for variable selection and its algorithm

For variable selection, we propose to minimize

$$\mathcal{L}_\lambda(\beta) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \log\{1 + \exp(-\beta^T(\mathbf{x}_i - \mathbf{x}_j)(y_i - y_j))\} + \sum_{j=1}^p p_\lambda(|\beta_j|), \quad (7)$$

where $p_\lambda(t)$ denotes a penalty function and $\lambda > 0$ is the tuning parameter. Conceptually any penalty function studied in the literature can be applied here. In this paper, we focus on three single-stage approaches with three commonly used penalty functions: LASSO [31], SCAD [11] and MCP [36]. For LASSO, $p_\lambda(t) = \lambda|t|$ and $p'_\lambda(t) = \lambda$ for $t > 0$. In SCAD, $p'_\lambda(t) = \lambda \mathbb{1}(t \leq \lambda) + (a\lambda - t)_+ / (a - 1) \mathbb{1}(t > \lambda)$ for some $a > 2$ and $t > 0$, where $\mathbb{1}(\cdot)$ is the indicator function. In MCP, $p'_\lambda(t) = (a\lambda - t)_+ / a$ for some $a > 1$ and $t > 0$. Note that the SCAD and the MCP belong to nonconvex penalizations. Also, some other two-stage approaches such as the adaptive LASSO [16,41] with the first stage pursuing an initial estimator used for the adaptive weights can also be applied here, but we skip their discussion for simplicity. We use $\hat{\beta}_\lambda$ to indicate minimizer of (7) for a given λ .

To implement the minimization of (7), note that the unpenalized objective function $\mathcal{L}(\beta)$ can be written as

$$\begin{aligned} \mathcal{L}(\beta) &= \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \log\{1 + \exp\{-\text{sgn}(y_i - y_j)\beta^T(\mathbf{x}_i - \mathbf{x}_j)|y_i - y_j|\}\} \\ &= \frac{1}{M} \sum_{k=1}^M \log\{1 + \exp(z_k \beta^T \mathbf{v}_k)\}, \end{aligned} \quad (8)$$

where $z_k = -\text{sgn}(y_i - y_j)$, $\mathbf{v}_k = (\mathbf{x}_i - \mathbf{x}_j)|y_i - y_j|$ and $M = n(n-1)/2$. After defining

$$u_k = \begin{cases} 1 & \text{if } y_i - y_j > 0, \\ 0 & \text{if } y_i - y_j < 0, \end{cases}$$

we can see that $\mathcal{L}(\beta)$ can be treated as the negative log-likelihood function of a regular logistic regression with response u_k , covariate $\mathbf{v}_k$, without the intercept term. Therefore, after the aforementioned data manipulation, the minimization of (7) can be carried out directly as a regular penalized logistic regression forcing the intercept to zero. In R, this procedure can be directly implemented using the package, e.g. `glmnet` [14].

3. Tuning parameter selection

The tuning parameter λ determines the model complexity and the prediction accuracy. When λ increases, the final model tends to have less selected variables and the prediction accuracy is decreased, and vice versa. There are two classical tuning parameter selection methods in the literature. One is the K -fold cross validation (CV). In this method, one randomly partitions the entire dataset D into K equally sized subsamples. Each time, for a given λ , one retains the k th subsample as the testing dataset $D^{(k)}$, where $k = 1, \dots, K$, and keeps the remaining $K-1$ subsamples $D \setminus D^{(k)}$ as the training dataset. Then, we choose the value of λ , λ_{CV} , which minimizes $CV(\lambda) = (1/K) \sum_{k=1}^K \mathcal{L}^{(k)}(\hat{\beta}_{\lambda}^{(-k)})$, where $\mathcal{L}^{(k)}(\cdot)$ is the $\mathcal{L}(\cdot)$ in (6) evaluated with dataset $D^{(k)}$. The Bayesian information criterion (BIC) is the other classical method to choose the tuning parameter λ . With this approach, one chooses λ_{BIC} that is the minimizer of $BIC(\lambda) = 2\mathcal{L}(\hat{\beta}_{\lambda}) + |\mathcal{S}|(\log(n)/n)$, where $|\mathcal{S}|$ is the size of the selected model, i.e. the number of nonzero estimators in the selected model. In high-dimensional settings, the definition of the BIC needs to be modified in order to guarantee the selection consistency [12,34].

Stability is advocated in the recent literature [23,35] as one of the criteria for data analysis due to the imperative role of reproducibility for any scientific discovery. The stability criterion can be brought to bear on estimation, variable selection, or the information criterion. In the following, we propose three different stability enhanced tuning parameter selection methods under our framework.

3.1. Estimation stability (sEST)

The key idea of estimation stability is that, if multiple samples are available from the same distribution, the parameter estimates from these different samples should be close in value to each other. Lim and Yu [19] considered the LASSO problem in the simple least square framework and studied an estimation stability method that is relevant to CV, to determining the tuning parameter.

The difficulty of directly employing this method under our framework is that, we consider a semiparametric model and use a conditional likelihood approach, thus we can only estimate the parameter β but not the nonparametric component $g(\cdot)$. Therefore we cannot compute residuals or make predictions based on the original data.

Instead, we propose to use the following technique to develop the estimation stability tuning parameter selection method for our framework. Note that our unpenalized objective function $\mathcal{L}(\beta)$ in (6) is a U-statistic of order 2. If the parameter is estimated by $\hat{\beta}_{\lambda}$, we compute $\hat{\beta}_{\lambda}^T(\mathbf{x}_i - \mathbf{x}_j)$ in the objective function, instead of $\hat{\beta}_{\lambda}^T \mathbf{x}_i$. Therefore, we create a vector with dimensionality $n(n-1)/2$ with its each element equal to $\hat{\beta}_{\lambda}^T(\mathbf{x}_i - \mathbf{x}_j)$ for $1 \leq i < j \leq n$. Using a similar idea as CV, we compute $\bar{\mathbf{z}}_{\lambda} = (1/K) \sum_{k=1}^K \hat{\mathbf{z}}_{\lambda}^{(-k)}$, where $\hat{\mathbf{z}}_{\lambda}^{(-k)} = (\hat{\beta}_{\lambda}^{(-k)})^T(\mathbf{x}_i - \mathbf{x}_j)$, and find a value of λ , denoted λ_{sest} , which minimizes

$$\text{sEST}(\lambda) = \frac{\frac{1}{K} \sum_{k=1}^K \|\hat{\mathbf{z}}_{\lambda}^{(-k)} - \bar{\mathbf{z}}_{\lambda}\|_2^2}{\|\bar{\mathbf{z}}_{\lambda}\|_2^2}. \quad (9)$$

To guarantee a more parsimonious model than the CV, similar to Lim and Yu [19], our final estimation stability enhanced tuning parameter is defined as $\lambda_{\text{sEST}} = \lambda_{\text{sest}} \vee \lambda_{CV}$.

3.2. Variable selection stability (sVS)

Using the same idea of stability, the key idea behind variable selection stability focuses on the concept that the sets of selected informative variables across different samples should not vary much. Sun *et al.* [29] proposed to divide the whole data set into two equally sized subsets and use Cohen's kappa [10] to assess the similarity of the sets of selected informative variables between the two subsets.

It is of interest to generalize the results in [29] to q equally sized subsets. When $q > 2$, the Cohen's kappa cannot be used immediately. Instead, we propose to use the Fleiss' kappa to assess the similarity of q binary data sets. Similar to Sun *et al.* [29], we repeatedly divide our completely observed data set into q equally sized subsets B times. For $b = 1, \dots, B$, we compute

$$\kappa_{\lambda}^b = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \quad (10)$$

where $\bar{P} = (1/pq(q-1))(\sum_{i=1}^p \sum_{j=0,1} s_{ij}^2 - pq)$, $\bar{P}_e = \sum_{j=0,1} ((1/pq) \sum_{i=1}^p s_{ij})^2$, $s_{i1} = \sum_{k=1}^q I_{\{\hat{\beta}_{ik} \neq 0\}}$, $s_{i0} = \sum_{k=1}^q I_{\{\hat{\beta}_{ik} = 0\}}$, and $\hat{\beta}_{ik}$ is the estimation of β_i from k th subset. Then we define the variable selection stability as $\Psi(\lambda) = (1/B) \sum_{b=1}^B \kappa_{\lambda}^b$.

We need to maximize $\Psi(\lambda)$ to find the optimal λ . But as pointed in [29], some informative variables may have a relatively weak effect compared with others. A large value of λ may produce an active set that consistently overlooks the weakly informative variables, which leads to an underfitted model with large variable selection stability. To remedy this issue, our final variable selection stability enhanced tuning parameter λ_{sVS} is defined as the minimizer of the set $\{\lambda : \Psi(\lambda)/\max_{\lambda'} \Psi(\lambda') \geq 1 - \alpha\}$. The theoretical derivation in [29] showed that the value of α needs to be small and tends to zero as the sample size n goes to infinity.

Under our framework, to divide the whole data set into q subsets with $q > 2$ may potentially reduce the computation time. Note that our objective function (6) has the computation burden of order $O(n^2)$. When we divide it to q subsets, the computation burden equals $O(n^2/q^2 \times q) = O(n^2/q)$. In our numerical studies, we evaluate the performance when $q = 2$ and $q = 5$. Note that this advantage does not exist in the traditional least square framework in [29] without missing data.

3.3. BIC stability (sBIC)

The idea of BIC stability is similar as above, but focuses on the closeness of the BIC across different subsets. The other motivation of the BIC stability is as follows. From the current literature, it is the general consensus that the CV method performs better in prediction accuracy while the BIC performs better in selection consistency. The BIC stability tuning parameter selection method we propose here is a mix of the CV and the BIC: when evaluating the CV, we change the loss function to the BIC. To be more specific, the optimal λ under this scheme is defined as

$$\lambda_{sBIC} = \operatorname{argmin}_{\lambda} \frac{1}{K} \sum_{k=1}^K \left\{ 2\mathcal{L}^{(k)}(\hat{\beta}_{\lambda}^{(-k)}) + |\mathcal{S}^{(-k)}| \frac{\log(n_k)}{n_k} \right\}, \quad (11)$$

where $S^{(-k)}$ is the set of all nonzero components in $\widehat{\beta}_\lambda^{(-k)}$ and n_k is the sample size in the k th subsample.

4. Simulation studies

In this section, we conduct simulation studies to examine the finite sample performance of our proposed method. We mainly compare the performance in terms of estimation and variable selection under different tuning parameter selection methods: CV, BIC, sEST, sVS2 for sVS method with $q = 2$, sVS5 for $q = 5$, and sBIC. We consider the convex penalty LASSO and two nonconvex penalties SCAD and MCP. We examine two types of PRO type outcomes: setting 1 with positive real outcomes and setting 2 with nonnegative integer outcomes.

In the first setting, we consider a classic scenario studied in [11,31]. We let the true parameter $\beta = (3, 1.5, 0, 0, 2, 0, \dots, 0)$ and we consider two scenarios, $p = 8$ and $p = 30$, with $N = 200$. We generate $\mathbf{X}$ from $N(\mathbf{0}, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$. We let $g(y) = \Phi(\sigma^2)^{-1/2} (2\pi\sigma^2)^{-1/2} e^{-(y-\sigma^2)^2/2\sigma^2}$ with $\sigma = 0.5$ and $y > 0$. The missing data mechanism is defined as $\text{pr}(R = 1 \mid y, \mathbf{x}) = \mathbb{1}(y > \gamma_1) \mathbb{1}(x_1 > \gamma_2)$ with $\gamma_1 = 0.15$ and $\gamma_2 = -1$ such that around 30% samples have missing values.

The second setting is similar to the first one except that, we let $g(y) = \lambda^y e^{-\lambda}/y!$ with $\lambda = 0.5$ and $y = 0, 1, \dots$, and $\text{pr}(R = 1 \mid y, \mathbf{x}) = \mathbb{1}(y < \gamma_1) \mathbb{1}(x_1 < \gamma_2)$ with $\gamma_1 = 5$, $\gamma_2 = 1$ and $N = 500$. Similar to setting 1, this mechanism also produces around 30% samples with missing data.

To evaluate the performance of estimation, we report the L_1 , L_2 , and L_∞ norms of the estimation bias

$$\|\widehat{\beta} - \beta\|_1, \quad \|\widehat{\beta} - \beta\|_2, \quad \|\widehat{\beta} - \beta\|_\infty$$

in all situations. To evaluate the performance of variable selection, we summarize the following measures: #FP, the number of false positives (the ones with true zero value but falsely estimated as nonzero); #FN, the number of false negatives (the ones with true nonzero value but falsely estimated as zero); and the F -measure, the harmonic mean of precision and sensitivity, which is defined as

$$F = \frac{2\#\text{TP}}{2\#\text{TP} + \#\text{FP} + \#\text{FN}},$$

where TP stands for true positive (the one with true nonzero value and also correctly estimated as nonzero). We also report the proportion of under-fit, correct-fit, and over-fit, where under-fit represents the situation of excluding any nonzero coefficients, correct-fit means the situation of selecting the exact subset model, and over-fit stands for the situation of including all three significant variables and some noise variables. We also report the time consumed (in seconds) per simulation replication for each tuning parameter selection method. Based on 100 replicates, we report the estimation and selection results under setting 1 in Tables 1 and 2, and those under setting 2 in Tables 3 and 4.

The conclusions from the simulation studies are apparent. The stability enhanced methods, especially sVS and sBIC, perform better than other methods in either estimation or variable selection. The behavior of BIC is not bad, and it is as good as stability enhanced

Table 1. Under setting 1, the mean and standard deviation (in parentheses) of the L_1 , L_2 , L_∞ norms of the estimation bias, and the average time (seconds) consumed for each method.

			L_1 -norm	L_2 -norm	L_∞ -norm	time (s)
$p = 8$	LASSO	CV	2.25 (0.105)	1.05 (0.046)	0.73 (0.035)	0.65
		BIC	1.43 (0.076)	0.83 (0.040)	0.64 (0.032)	0.06
		sEST	1.33 (0.080)	0.80 (0.047)	0.64 (0.038)	2.72
		sVS2	1.26 (0.085)	0.81 (0.052)	0.65 (0.044)	4.43
		sVS5	1.37 (0.102)	0.88 (0.066)	0.72 (0.055)	1.43
		sBIC	1.24 (0.077)	0.77 (0.047)	0.62 (0.040)	1.47
	SCAD	CV	1.60 (0.089)	0.90 (0.044)	0.68 (0.033)	2.34
		BIC	1.21 (0.069)	0.76 (0.041)	0.61 (0.033)	0.08
		sEST	1.39 (0.098)	0.86 (0.060)	0.69 (0.048)	6.16
		sVS2	1.26 (0.092)	0.81 (0.057)	0.67 (0.048)	10.87
		sVS5	1.34 (0.105)	0.87 (0.066)	0.71 (0.054)	3.13
		sBIC	1.14 (0.067)	0.73 (0.041)	0.59 (0.033)	3.11
	MCP	CV	1.46 (0.091)	0.85 (0.045)	0.66 (0.034)	2.40
		BIC	1.14 (0.068)	0.74 (0.041)	0.60 (0.033)	0.08
		sEST	1.57 (0.121)	1.00 (0.074)	0.80 (0.058)	6.26
		sVS2	1.20 (0.082)	0.80 (0.052)	0.66 (0.043)	10.50
		sVS5	1.25 (0.086)	0.81 (0.051)	0.67 (0.041)	3.24
		sBIC	1.12 (0.070)	0.74 (0.045)	0.61 (0.037)	3.21
$p = 30$	LASSO	CV	5.82 (0.341)	1.90 (0.094)	1.05 (0.057)	1.86
		BIC	1.56 (0.099)	0.89 (0.048)	0.66 (0.036)	0.13
		sEST	1.79 (0.155)	0.99 (0.072)	0.74 (0.055)	5.64
		sVS2	1.48 (0.103)	0.95 (0.068)	0.78 (0.059)	10.39
		sVS5	1.37 (0.087)	0.85 (0.053)	0.68 (0.044)	3.04
		sBIC	1.67 (0.150)	1.06 (0.093)	0.85 (0.074)	3.07
	SCAD	CV	3.23 (0.203)	1.38 (0.072)	0.86 (0.047)	8.88
		BIC	1.70 (0.112)	0.95 (0.052)	0.70 (0.037)	0.15
		sEST	1.78 (0.133)	1.00 (0.068)	0.74 (0.052)	19.76
		sVS2	1.49 (0.108)	0.96 (0.070)	0.78 (0.059)	36.11
		sVS5	1.42 (0.095)	0.89 (0.056)	0.71 (0.045)	10.09
		sBIC	1.54 (0.130)	0.97 (0.079)	0.77 (0.063)	10.04
	MCP	CV	2.31 (0.164)	1.14 (0.066)	0.78 (0.043)	9.27
		BIC	1.39 (0.098)	0.85 (0.052)	0.67 (0.039)	0.15
		sEST	1.92 (0.150)	1.19 (0.087)	0.94 (0.069)	20.65
		sVS2	1.46 (0.109)	0.97 (0.072)	0.80 (0.060)	35.23
		sVS5	1.42 (0.105)	0.89 (0.058)	0.71 (0.045)	10.58
		sBIC	1.45 (0.110)	0.95 (0.068)	0.78 (0.055)	10.58

methods under a few situations. Although the method sVS5 has slightly worse performance compared to sVS2 under some scenarios, it takes much less time than sVS2. In general, there is no obvious winner among the stability enhanced methods, but we will focus on BIC, sVS5, and sBIC to analyze our motivating example, the ChAMP study.

5. Application to the ChAMP study

The ChAMP study is a double-blinded randomized controlled trial designed to investigate the effect of debridement versus observation of unstable chondral lesions on pain in patients undergoing arthroscopic partial meniscectomy [2]. In our data analysis, the primary outcome variable is the WOMAC pain score 1 year after the surgery. This pain score is a PRO ranging from 0 to 100, and a larger value indicates more severe pain. Our purpose is to identify important factors that are associated with this WOMAC pain score.

Table 2. Under setting 1, the mean and standard deviation (in parentheses) of #FP, #FN, and F-measure, and the proportion of under-fit, correct-fit, and over-fit, for each method.

			#FP	#FN	F-measure	Proportion of		
						u-fit	c-fit	o-fit
$p = 8$	LASSO	CV	3.06 (0.139)	0.00 (0.000)	0.68 (0.011)	0.00	0.03	0.97
		BIC	0.97 (0.099)	0.00 (0.000)	0.88 (0.011)	0.00	0.37	0.63
		sEST	0.65 (0.095)	0.05 (0.030)	0.91 (0.012)	0.03	0.54	0.43
		sVS2	0.24 (0.049)	0.08 (0.031)	0.95 (0.009)	0.06	0.73	0.21
		sVS5	0.24 (0.049)	0.12 (0.043)	0.94 (0.012)	0.08	0.71	0.21
		sBIC	0.42 (0.061)	0.02 (0.020)	0.94 (0.009)	0.01	0.63	0.36
	SCAD	CV	1.05 (0.137)	0.00 (0.000)	0.88 (0.014)	0.00	0.48	0.52
		BIC	0.23 (0.053)	0.00 (0.000)	0.97 (0.007)	0.00	0.81	0.19
		sEST	0.33 (0.075)	0.12 (0.046)	0.93 (0.013)	0.07	0.70	0.23
		sVS2	0.11 (0.035)	0.07 (0.033)	0.97 (0.009)	0.04	0.86	0.10
		sVS5	0.11 (0.035)	0.11 (0.042)	0.96 (0.011)	0.07	0.83	0.10
		sBIC	0.14 (0.038)	0.00 (0.000)	0.98 (0.005)	0.00	0.87	0.13
	MCP	CV	0.70 (0.131)	0.00 (0.000)	0.92 (0.013)	0.00	0.66	0.34
		BIC	0.09 (0.032)	0.00 (0.000)	0.99 (0.004)	0.00	0.92	0.08
		sEST	0.21 (0.071)	0.27 (0.063)	0.91 (0.016)	0.17	0.70	0.13
		sVS2	0.03 (0.017)	0.11 (0.037)	0.97 (0.009)	0.09	0.88	0.03
		sVS5	0.12 (0.036)	0.07 (0.033)	0.97 (0.009)	0.05	0.84	0.11
		sBIC	0.03 (0.017)	0.04 (0.024)	0.99 (0.006)	0.03	0.94	0.03
$p = 30$	LASSO	CV	8.98 (0.381)	0.00 (0.000)	0.43 (0.012)	0.00	0.00	1.00
		BIC	0.86 (0.100)	0.01 (0.010)	0.89 (0.011)	0.01	0.44	0.55
		sEST	0.98 (0.169)	0.08 (0.037)	0.87 (0.016)	0.05	0.49	0.46
		sVS2	0.25 (0.061)	0.19 (0.046)	0.93 (0.012)	0.15	0.68	0.17
		sVS5	0.31 (0.063)	0.10 (0.033)	0.94 (0.010)	0.09	0.68	0.23
		sBIC	0.16 (0.044)	0.31 (0.076)	0.89 (0.023)	0.17	0.70	0.13
	SCAD	CV	3.65 (0.231)	0.00 (0.000)	0.66 (0.016)	0.00	0.08	0.92
		BIC	0.83 (0.109)	0.01 (0.010)	0.90 (0.012)	0.01	0.52	0.47
		sEST	0.98 (0.144)	0.08 (0.037)	0.87 (0.016)	0.05	0.53	0.42
		sVS2	0.19 (0.049)	0.22 (0.052)	0.93 (0.013)	0.16	0.69	0.15
		sVS5	0.28 (0.064)	0.12 (0.036)	0.94 (0.010)	0.11	0.69	0.20
		sBIC	0.22 (0.050)	0.18 (0.059)	0.92 (0.018)	0.10	0.72	0.18
	MCP	CV	1.72 (0.158)	0.01 (0.010)	0.81 (0.015)	0.00	0.28	0.72
		BIC	0.33 (0.067)	0.04 (0.020)	0.95 (0.009)	0.03	0.73	0.24
		sEST	0.33 (0.083)	0.38 (0.071)	0.87 (0.018)	0.24	0.57	0.19
		sVS2	0.07 (0.026)	0.24 (0.053)	0.94 (0.013)	0.19	0.74	0.07
		sVS5	0.24 (0.061)	0.10 (0.033)	0.95 (0.010)	0.08	0.75	0.17
		sBIC	0.11 (0.035)	0.21 (0.050)	0.94 (0.013)	0.14	0.76	0.10

The major treatment variable Trt has three levels in this study: Trt = 1 indicates the patients who had unstable chondral lesions that were exercised during the surgery; Trt = 2 indicates the patients who also had unstable chondral lesions but that were left unaltered during the surgery; Trt = 3 indicates the patients who did not have unstable chondral lesions. Based on previous analyses of this study [4,5,18], some other potential covariate variables of clinical interest are: age, sex, BMI, leg, MLTear, TFPT, and LaMe, where leg is a binary variable indicating which leg the surgery is on (left/right), MLTear is a two level categorical variable meaning the locations of meniscus tear, either in the medial or lateral compartment or both, TFPT is a five-level categorical variable showing the number of locations (Tibia, Femur, Patella, and Trochlea) where chondral lesions exist, and LaMe is a three-level categorical variable representing bone bruising as none, present in medial or lateral compartment, or present in both medial and lateral compartments.

Table 3. Under setting 2, the mean and standard deviation (in parentheses) of the L_1 , L_2 , L_∞ norms of the estimation bias, and the average time (seconds) consumed for each method.

			L_1 -norm	L_2 -norm	L_∞ -norm	Time (s)
$p = 8$	LASSO	CV	1.30 (0.068)	0.61 (0.035)	0.43 (0.028)	2.68
		BIC	0.93 (0.059)	0.53 (0.033)	0.41 (0.027)	0.39
		sEST	0.86 (0.060)	0.51 (0.034)	0.40 (0.027)	16.00
		sVS2	0.83 (0.060)	0.51 (0.036)	0.40 (0.028)	29.33
		sVS5	0.97 (0.096)	0.59 (0.057)	0.45 (0.040)	10.21
		sBIC	0.86 (0.057)	0.51 (0.033)	0.40 (0.027)	10.24
	SCAD	CV	0.84 (0.064)	0.51 (0.035)	0.40 (0.027)	7.50
		BIC	0.78 (0.054)	0.49 (0.033)	0.39 (0.026)	0.45
		sEST	0.78 (0.054)	0.49 (0.033)	0.39 (0.026)	25.31
		sVS2	0.78 (0.054)	0.49 (0.032)	0.39 (0.026)	48.21
		sVS5	0.83 (0.057)	0.51 (0.033)	0.40 (0.026)	14.75
		sBIC	0.78 (0.054)	0.49 (0.033)	0.39 (0.026)	14.71
	MCP	CV	0.83 (0.060)	0.51 (0.034)	0.40 (0.027)	6.58
		BIC	0.77 (0.053)	0.49 (0.032)	0.39 (0.026)	0.44
		sEST	0.77 (0.053)	0.49 (0.032)	0.39 (0.026)	23.79
		sVS2	0.77 (0.053)	0.49 (0.032)	0.39 (0.026)	43.94
		sVS5	0.82 (0.057)	0.50 (0.033)	0.40 (0.026)	14.06
		sBIC	0.77 (0.053)	0.49 (0.032)	0.39 (0.026)	14.18
$p = 30$	LASSO	CV	3.33 (0.157)	1.04 (0.048)	0.65 (0.038)	6.03
		BIC	0.89 (0.045)	0.50 (0.024)	0.38 (0.020)	0.72
		sEST	0.87 (0.068)	0.52 (0.041)	0.40 (0.033)	22.87
		sVS2	0.71 (0.041)	0.44 (0.024)	0.35 (0.019)	42.84
		sVS5	0.76 (0.044)	0.46 (0.025)	0.36 (0.020)	13.38
		sBIC	0.76 (0.043)	0.46 (0.024)	0.36 (0.019)	13.54
	SCAD	CV	0.95 (0.060)	0.53 (0.027)	0.40 (0.021)	18.66
		BIC	0.70 (0.041)	0.44 (0.024)	0.35 (0.019)	0.78
		sEST	0.80 (0.067)	0.49 (0.041)	0.39 (0.033)	48.23
		sVS2	0.69 (0.041)	0.44 (0.024)	0.35 (0.020)	93.29
		sVS5	0.72 (0.042)	0.45 (0.025)	0.36 (0.020)	26.11
		sBIC	0.69 (0.042)	0.44 (0.025)	0.35 (0.020)	26.23
	MCP	CV	0.85 (0.058)	0.49 (0.028)	0.38 (0.021)	18.81
		BIC	0.68 (0.041)	0.44 (0.025)	0.35 (0.020)	0.77
		sEST	0.77 (0.070)	0.49 (0.043)	0.39 (0.035)	48.35
		sVS2	0.68 (0.041)	0.43 (0.025)	0.35 (0.020)	88.87
		sVS5	0.72 (0.044)	0.45 (0.025)	0.35 (0.020)	26.31
		sBIC	0.68 (0.041)	0.43 (0.025)	0.35 (0.020)	26.36

In this study, we have in total $N = 266$ patients. Many variables have missing values, for instance, the WOMAC pain score has 48 missing values, LaMe has 25 missing values, BMI has 2 missing values. It is highly suspect that the missingness in pain score is nonignorable because it may be due to complete resolution of pain or increased pain. Given the complex unknown reasons for the missingness, we posit the partially specified assumption (2) for this study. The number of patients without any missing values is $n = 196$.

Without selecting variables, we fit the PLR model for the WOMAC pain score and obtain the estimate of the unknown parameter by minimizing (6). We minimize (7), advocating to select informative variables, to obtain the estimate of the unknown parameter. In this step, for simplicity, we only present the results with the MCP penalty and tuning parameter selection methods BIC, sVS5 and sBIC.

We first treat Trt as a three-level variable and the results are shown in the upper panel in Table 5. Using the variable selection technique, the coefficient of the variable Trt2vs1 is

Table 4. Under setting 2, the mean and standard deviation (in parentheses) of #FP, #FN, and F-measure, and the proportion of under-fit, correct-fit, and over-fit, for each method.

						Proportion of		
						u-fit	c-fit	o-fit
			#FP	#FN	F-measure			
$p = 8$	LASSO	CV	4.10 (0.097)	0.00 (0.000)	0.60 (0.007)	0.00	0.00	1.00
		BIC	1.07 (0.100)	0.00 (0.000)	0.86 (0.011)	0.00	0.34	0.66
		sEST	0.55 (0.078)	0.00 (0.000)	0.93 (0.010)	0.00	0.59	0.41
		sVS2	0.18 (0.044)	0.01 (0.010)	0.97 (0.006)	0.01	0.83	0.16
		sVS5	0.28 (0.060)	0.05 (0.026)	0.95 (0.009)	0.04	0.75	0.21
		sBIC	0.59 (0.075)	0.00 (0.000)	0.92 (0.010)	0.00	0.56	0.44
	SCAD	CV	0.34 (0.090)	0.00 (0.000)	0.96 (0.010)	0.00	0.82	0.18
		BIC	0.02 (0.014)	0.00 (0.000)	1.00 (0.002)	0.00	0.98	0.02
		sEST	0.03 (0.017)	0.00 (0.000)	1.00 (0.002)	0.00	0.97	0.03
		sVS2	0.03 (0.017)	0.00 (0.000)	1.00 (0.002)	0.00	0.97	0.03
		sVS5	0.21 (0.067)	0.00 (0.000)	0.97 (0.008)	0.00	0.87	0.13
		sBIC	0.02 (0.014)	0.00 (0.000)	1.00 (0.002)	0.00	0.98	0.02
	MCP	CV	0.25 (0.070)	0.00 (0.000)	0.97 (0.008)	0.00	0.84	0.16
		BIC	0.00 (0.000)	0.00 (0.000)	1.00 (0.000)	0.00	1.00	0.00
		sEST	0.00 (0.000)	0.00 (0.000)	1.00 (0.000)	0.00	1.00	0.00
		sVS2	0.00 (0.000)	0.00 (0.000)	1.00 (0.000)	0.00	1.00	0.00
		sVS5	0.17 (0.064)	0.00 (0.000)	0.98 (0.007)	0.00	0.90	0.10
		sBIC	0.00 (0.000)	0.00 (0.000)	1.00 (0.000)	0.00	1.00	0.00
$p = 30$	LASSO	CV	13.71 (0.390)	0.00 (0.000)	0.32 (0.007)	0.00	0.00	1.00
		BIC	1.00 (0.104)	0.00 (0.000)	0.87 (0.012)	0.00	0.38	0.62
		sEST	0.60 (0.092)	0.02 (0.020)	0.92 (0.011)	0.01	0.57	0.42
		sVS2	0.15 (0.039)	0.00 (0.000)	0.98 (0.005)	0.00	0.86	0.14
		sVS5	0.34 (0.064)	0.00 (0.000)	0.95 (0.008)	0.00	0.74	0.26
		sBIC	0.39 (0.063)	0.00 (0.000)	0.95 (0.008)	0.00	0.68	0.32
	SCAD	CV	0.96 (0.157)	0.00 (0.000)	0.89 (0.015)	0.00	0.59	0.41
		BIC	0.07 (0.026)	0.00 (0.000)	0.99 (0.004)	0.00	0.93	0.07
		sEST	0.18 (0.052)	0.02 (0.020)	0.97 (0.008)	0.01	0.86	0.13
		sVS2	0.04 (0.024)	0.00 (0.000)	0.99 (0.003)	0.00	0.97	0.03
		sVS5	0.13 (0.044)	0.00 (0.000)	0.98 (0.006)	0.00	0.90	0.10
		sBIC	0.04 (0.020)	0.00 (0.000)	0.99 (0.003)	0.00	0.96	0.04
	MCP	CV	0.53 (0.119)	0.00 (0.000)	0.94 (0.012)	0.00	0.78	0.22
		BIC	0.01 (0.010)	0.00 (0.000)	1.00 (0.001)	0.00	0.99	0.01
		sEST	0.03 (0.017)	0.03 (0.022)	0.99 (0.006)	0.02	0.95	0.03
		sVS2	0.00 (0.000)	0.00 (0.000)	1.00 (0.000)	0.00	1.00	0.00
		sVS5	0.13 (0.039)	0.00 (0.000)	0.98 (0.005)	0.00	0.89	0.11
		sBIC	0.00 (0.000)	0.00 (0.000)	1.00 (0.000)	0.00	1.00	0.00

estimated as zero, which indicates that there is no difference between $\text{Trt} = 1$ and $\text{Trt} = 2$. Hence, when the unstable chondral lesions exist, there is no evidently significant difference of whether to treat them or not during the surgery, in terms of the patients' WOMAC pain score change 1 year after the surgery. This is consistent with the major clinical findings in the ChAMP study [3]. This result could have a big impact on the future surgery practice on whether to treat the chondral lesions during this surgery.

Furthermore, we combine $\text{Trt} = 1$ and $\text{Trt} = 2$, and conduct a refined analysis and the results are shown in the lower panel in Table 5. The estimate of the coefficient for the variable Trt3vs12 is nonzero, showing that whether the unstable chondral lesions exist does have some effect on the patients' WOMAC pain score change 1 year after the surgery. Additionally, different tuning parameter selection methods provide the same selection results.

Table 5. Data analysis results for the ChAMP study.

	Age	Sex	BMI	Leg	MLTear	TFPT	LaMe	Trt2vs1	Trt3vs1
No selection	−0.0154	−0.1315	−0.0379	0.0023	−0.0180	−0.0793	−0.0317	0.0554	0.2912
MCP BIC	0	0	0	0	0	0	0	0	0.1845
MCP sVS5	0	−0.0104	0	0	0	0	0	0	0.1976
MCP sBIC	0	0	0	0	0	0	0	0	0.0759
	Age	Sex	BMI	Leg	MLTear	TFPT	LaMe	Trt3vs12	
No selection	−0.0198	−0.1312	−0.0323	0.0036	−0.0211	−0.0705	−0.0378	0.2720	
MCP BIC	0	0	0	0	0	0	0	0.1845	
MCP sVS5	0	0	0	0	0	0	0	0.0970	
MCP sBIC	0	0	0	0	0	0	0	0.0291

Figure 1. The solution path in the ChAMP study.

The solution path is also given in Figure 1 to demonstrate how the estimates change across different λ values.

6. Conclusion and discussion

How to handle missing values when we perform variable selection is an interesting but challenging topic. In the literature, methods based on parametric model assumptions, multiple imputation or resampling technique had been proposed in [21,32] and among many others.

In this paper, motivated by analyzing the WOMAC pain score in the ChAMP study, we consider a penalized likelihood framework for variable selection under a flexible missing

data mechanism and propose a variety of stability enhanced methods to determining the tuning parameter.

Flexibility is an overarching theme in our proposed method. For the WOMAC pain score, due to its inherent feature, we propose to fit a semiparametric model, which is more flexible than a purely parametric model. Due to the intrinsic challenges in handling missing data, especially nonignorable missing data, we impose a generally applicable missingness mechanism. In variable selection, we also propose stability enhanced tuning parameter selection methods, which can be brought to bear on estimation, selection, or the information criterion. Also, since our framework only distinguishes whether it is a completely observed subject or not, we do not confine our methodological approach to address either missing response or missing covariates.

Identifiability is always an issue worthy of further discussion. The key idea in our method is that, although we cannot fully identify all the components in the data generation model and the missingness mechanism model, we can still perform variable selection through β which *can* be identified and estimated. The unidentification of all other model components is attributed to the information loss brought by the missing data and is also the price we have to pay in order to successfully estimate β as well as conduct variable selection through β .

Last but not least, due to the space constraint, we do not study the theoretical property of the proposed estimation in this paper. Zhao *et al.* [40] considered a GLM for Y given X and showed the asymptotic theory under their assumptions. Briefly, in their setting, the dimensionality requirement is $\log p = o(n^{1-4\varsigma}/(\log n)^2)$ where $0 < \varsigma < 1/4$. It means, for the high dimensional GLM as in [40], the number of covariates p can grow at most exponentially fast with the complete case sample size n . The similar theory for the semiparametric PLR model studied in this paper should be similar but more complicated, which would be an exciting research topic for future investigation.

Acknowledgments

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors thank the Editor, the Associate Editor and two anonymous referees for their constructive comments and insightful suggestions, which have led to a significantly improved paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR001412.

ORCID

Jiwei Zhao <http://orcid.org/0000-0002-9298-9412>

Melissa Kluczynski <http://orcid.org/0000-0003-3947-4222>

References

- [1] I. Arostegui, V. Núñez-Antón, and J.M. Quintana, *Statistical approaches to analyse patient-reported outcomes as response variables: an application to health-related quality of life*, Stat. Methods Med. Res. 21 (2012), pp. 189–214.
- [2] L.J. Bisson, M.A. Kluczynski, W.M. Wind, M.S. Fineberg, G.A. Bernas, M.A. Rauh, J.M. Marzo, and R.J. Smolinski, *Design of a randomized controlled trial to compare debridement to observation of chondral lesions encountered during partial meniscectomy: The ChAMP (Chondral Lesions And Meniscus Procedures) Trial*, Contemp. Clin. Trials 45 (2015), pp. 281–286.
- [3] L.J. Bisson, M.A. Kluczynski, W.M. Wind, M.S. Fineberg, G.A. Bernas, M.A. Rauh, J.M. Marzo, Z. Zhou, and J. Zhao, *Patient outcomes after observation versus debridement of unstable chondral lesions during partial meniscectomy: The Chondral Lesions And Meniscus Procedures (ChAMP) randomized controlled trial*, J. Bone Joint Surg. 99 (2017), pp. 1078–1085.
- [4] L.J. Bisson, M.A. Kluczynski, W.M. Wind, M.S. Fineberg, G.A. Bernas, M.A. Rauh, J.M. Marzo, Z. Zhou, and J. Zhao, *How does the presence of unstable chondral lesions affect patient outcomes after partial meniscectomy?: The Chondral Lesions And Meniscus Procedures (ChAMP) randomized controlled trial*, Am. J. Sports Med. 46 (2018), pp. 590–597.
- [5] L.J. Bisson, P. Phillips, J. Matthews, Z. Zhou, J. Zhao, W.M. Wind, M.S. Fineberg, G.A. Bernas, M.A. Rauh, J.M. Marzo, et al. *Association between bone marrow lesions, chondral lesions, and pain in patients without radiographic evidence of degenerative joint disease who underwent arthroscopic partial meniscectomy*, Orthop. J. Sports Med. 7 (2019), p. 2325967119830381.
- [6] J.C. Cappelleri and A.G. Bushmakina, *Interpretation of patient-reported outcomes*, Stat. Methods Med. Res. 23 (2014), pp. 460–483.
- [7] J.C. Cappelleri, K.H. Zou, A.G. Bushmakina, J.M.J. Alvir, D. Alemayehu, and T. Symonds, *Patient-reported Outcomes: Measurement, Implementation and Interpretation*, Chapman and Hall/CRC, Boca Raton, FL, 2013.
- [8] K.C.G. Chan, *Nuisance parameter elimination for proportional likelihood ratio models with nonignorable missingness and random truncation*, Biometrika 100 (2013), pp. 269–276.
- [9] E. Chan, C. Jamieson, H. Metin, and S. Hudgens, *Missing data in patient-reported outcomes: Regulatory, statistical, and operational perspectives*, Value Health 21 (2018), p. S227.
- [10] J. Cohen, *A coefficient of agreement for nominal scales*, Educ. Psychol. Meas. 20 (1960), pp. 37–46.
- [11] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Am. Stat. Assoc. 96 (2001), pp. 1348–1360.
- [12] Y. Fan and C.Y. Tang, *Tuning parameter selection in high dimensional penalized likelihood*, J. R. Stat. Soc. Ser. B 75 (2013), pp. 531–552.
- [13] F. Fang, J. Zhao, and J. Shao, *Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values*, Stat. Sin. 28 (2018), pp. 1677–1701.
- [14] J. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, J. Stat. Softw. 33 (2010), p. 1.
- [15] M. Gomes, N. Gutacker, C. Bojke, and A. Street, *Addressing missing data in patient-reported outcome measures (PROMS): Implications for the use of PROMS for comparing provider performance*, Health Econ. 25 (2016), pp. 515–528.
- [16] J. Huang, S. Ma, and C.-H. Zhang, *Adaptive Lasso for sparse high-dimensional regression models*, Stat. Sin. (2008), pp. 1603–1618.
- [17] J.D. Kalbfleisch, *Likelihood methods and nonparametric tests*, J. Am. Stat. Assoc. 73 (1978), pp. 167–170.
- [18] M.A. Kluczynski, J.M. Marzo, W.M. Wind, M.S. Fineberg, G.A. Bernas, M.A. Rauh, Z. Zhou, J. Zhao, and L.J. Bisson, *The effect of body mass index on clinical outcomes in patients without radiographic evidence of degenerative joint disease after arthroscopic partial meniscectomy*, Arthroscopy 33 (2017), pp. 2054–2063.
- [19] C. Lim and B. Yu, *Estimation Stability with Cross-Validation (ESCV)*, J. Comput. Graph. Stat. 25 (2016), pp. 464–492.

- [20] R.J. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed., Wiley, Hoboken, NJ, 2002.
- [21] Q. Long and B.A. Johnson, *Variable selection in the presence of missing data: resampling and imputation*, *Biostatistics* 16 (2015), pp. 596–610.
- [22] X. Luo and W.Y. Tsai, *A proportional likelihood ratio model*, *Biometrika* 99 (2012), pp. 211–222.
- [23] N. Meinshausen and P. Bühlmann, *Stability selection*, *J. R. Stat. Soc. Ser. B* 72 (2010), pp. 417–473.
- [24] R. Mercieca-Bebber, M.J. Palmer, M. Brundage, M. Calvert, M.R. Stockler, and M.T. King, *Design, implementation and reporting strategies to reduce the instance and impact of missing patient-reported outcome (PRO) data: a systematic review*, *BMJ Open* 6 (2016), p. e010938.
- [25] I. Rombach, A.M. Gray, C. Jenkinson, D.W. Murray, and O. Rivero-Arias, *Multiple imputation for patient reported outcome measures in randomised controlled trials: advantages and disadvantages of imputing at the item, subscale or composite score level*, *BMC Med. Res. Methodol.* 18 (2018), p. 87.
- [26] I. Rombach, O. Rivero-Arias, A.M. Gray, C. Jenkinson, and O. Burke, *The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature*, *Qual. Life Res.* 25 (2016), pp. 1613–1623.
- [27] D.O. Scharfstein and A. McDermott, *Global sensitivity analysis of clinical trials with missing patient-reported outcomes*, *Stat. Methods Med. Res.* 28 (2019), pp. 1439–1456.
- [28] J. Shao and J. Zhao, *Estimation in longitudinal studies with nonignorable dropout*, *Stat. Interface* 6 (2013), pp. 303–313.
- [29] W. Sun, J. Wang, and Y. Fang, *Consistent selection of tuning parameters via variable selection stability*, *J. Mach. Learn. Res.* 14 (2013), pp. 3419–3440.
- [30] G. Tang, R.J. Little, and T.E. Raghunathan, *Analysis of multivariate missing data with nonignorable nonresponse*, *Biometrika* 90 (2003), pp. 747–764.
- [31] R. Tibshirani, *Regression shrinkage and selection via the lasso*, *J. R. Stat. Soc. Ser. B* 58 (1996), pp. 267–288.
- [32] C.-h. Tseng and Y.-H. Chen, *Regularized approach for data missing not at random*, *Stat. Methods Med. Res.* (2017), p. 0962280217717760.
- [33] J. Valderas, A. Kotzeva, M. Espallargues, G. Guyatt, C. Ferrans, M.Y. Halyard, D. Revicki, T. Symonds, A. Parada, and J. Alonso, *The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature*, *Qual. Life Res.* 17 (2008), pp. 179–193.
- [34] H. Wang, B. Li, and C. Leng, *Shrinkage tuning parameter selection with a diverging number of parameters*, *J. R. Stat. Soc. Ser. B* 71 (2009), pp. 671–683.
- [35] B. Yu, *Stability*, *Bernoulli* 19 (2013), pp. 1484–1500.
- [36] C.-H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, *Ann. Statist.* 38 (2010), pp. 894–942.
- [37] J. Zhao, *Reducing bias for maximum approximate conditional likelihood estimator with general missing data mechanism*, *J. Nonparametr. Stat.* 29 (2017), pp. 577–593.
- [38] J. Zhao and Y. Ma, *Optimal pseudolikelihood estimation in the analysis of multivariate missing data with nonignorable nonresponse*, *Biometrika* 105 (2018), pp. 479–486.
- [39] J. Zhao and J. Shao, *Approximate conditional likelihood for generalized linear models with general missing data mechanism*, *J. Syst. Sci. Complex.* 30 (2017), pp. 139–153.
- [40] J. Zhao, Y. Yang, and Y. Ning, *Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data*, *Stat. Sin.* 28 (2018), pp. 2125–2148.
- [41] H. Zou, *The adaptive lasso and its oracle properties*, *J. Am. Stat. Assoc.* 101 (2006), pp. 1418–1429.