Statistics and Its Interface Volume 9 (2016) 79–91

Identification of significant B cell associations with undetected observations using a Tobit model*

Tian Chen, Shujie Ma, James Kobie, Alexander Rosenberg, Ignacio Sanz, and Hua Liang†

To study the relationship of serum antibody neutralization activity (determined by IC50) and the B cell immune response, we face two challenges: (i) IC50 values can not be observed when they are below the detected limitation, and (ii) the number of factors is larger than the number of observations. To address these two challenges, we propose a Tobit model for the analysis of the study, and an adaptive LASSO penalized variable selection procedure to identify important factors. Furthermore, we suggest extended Bayesian information criterion for selecting the tuning parameter. Our analysis indicates that three measured B cells, specifically the frequency of CD19+CD20+, CD19-CD20+, and IgD-B220-CD27- peripheral blood B cell subsets have significant effects on IC50. We have also run simulation studies to evaluate the numerical performance of the proposed method.

Keywords and phrases: Extended Bayesian information criterion, LASSO, Penalized likelihood, High-dimensional Tobit model.

1. INTRODUCTION

Due to the extensive mutation of HIV-1, a major objective for a preventative vaccine against HIV is the induction of antibodies that are capable of recognizing diverse HIV isolates. Approximately 20% of HIV infected patients develop potent antibodies that are capable of neutralizing a broad range of HIV isolates, yet due to the viruses' rapid mutation typically fail to recognize the individual's contemporary isolate [1]. However, the presence of these HIV broadly neutralizing antibodies demonstrates the potential of the human immune response, and it is suggested that by understanding their development in these HIV infected patients effective vaccination strategies can be developed. To this end as antibodies develop from B cells, we conducted a study that focused on defining characteristics of the B cell compartment that are associated with the presence of serum antibodies that neutralize HIV infectivity, we collected 42 observations from HIV-infected patients including frequency of peripheral blood B cell subsets (51 covariates) [2, 3], abundance of serum auto-reactive antibodies (4 covariates) which we [4] and others [5] have previously shown to be associated with the presence of HIV neutralizing antibodies, and clinical characteristics including age, time since diagnosis, CD4 cell counts, and HIV viral load (4 covariates). Additionally, the HIV neutralizing activity of their serum was determined by the half maximal inhibitory concentration (IC50). We are interested in the relationship between the IC50 and these covariates to identify significant B cell characteristics. Because of the limitation of the technology, the exact IC50 values cannot be exactly observed when the the values are below 20, instead; we only know that such observations are less than 20, that is, such observations are left censored. In our dataset about 30% IC50 values are less than 20.

In the absence of undetected observations, linear regression models may be used to study the relationship between the covariates and the response variable. To handle the cases with undetected observations, we may simply (a) exclude the observations with undetected values, or (b) dichotomize the response variable (IC50 < 20 vs IC50 ≥ 20), or (c) replace the unobserved values with the threshold. Although these options serve as a convenient way to analyze data with undetected observations, all of them have serious limitations. Option (a) will reduce the sample size (in our data application, only 28 of the original 42 observations remain after deleting 30% of them); option (b) will cause loss of information as we equally treat the samples with large IC50 values and small values equally as long as they are larger than the threshold; and option (c) may lead to biased estimators of the parameters and is not justified theoretically. To remedy these limitations, we use the Tobit model to investigate the relationship between response IC50 and the 59 covariates as this model incorporates the likelihood function representing the censoring information.

*This research was partially supported by National Institutes of Health (NIH) grants R01AI084808, R21AI078459, and R37AI049660 to IS, and the University of Rochester Developmental Center for AIDS Research grant P30 AI078498 (NIH/NIAID). Liang's research was also partially supported by NSF grants DMS-1207444 and DMS 1418042, by Award Number 11228103, made by National Natural Science Foundation of China.

†Corresponding author.

To identify significant B cell characteristics, which is essentially a concern of variable selection, we propose a penalization approach, which can gain variable selection to detect significant B cells or other important covariates. Variable selection for linear regression models by penalized estimates has attracted a lot of attention in the past decades. A variety of penalties have been proposed for such a purpose. Examples include the bridge penalty [6], the nonnegative garrote penalty [7], the least absolute shrinkage and selection operator (LASSO) penalty [8, 9, 10], and the smoothly clipped absolute deviation (SCAD) penalty [11]. These penalization-based variable selection approaches can somewhat avoid drawbacks such as heavy computational burden and instability [7, 11] occurred in the classical variable selection methods like best subset selection, stepwise selection and criterion based methods (AIC and BIC) when the dimension is high. Meanwhile, these penalization-based methods have nice statistical properties under certain assumptions [11].

It is worth pointing out that the LASSO method [9] can shrink some coefficients to 0, and thus gains the goal of variable selection. However, LASSO still has its own limitations like lack of the oracle property [11]. As a remedy, Zou [10] proposed a variant, adaptive LASSO by using a weighted L_1 penalty that allows larger penalty for zero coefficients and smaller penalty for the nonzero coefficients. Zou [10] and and Huang, Ma and Zhang [12] further established the oracle property of the adaptive LASSO under certain assumptions for linear regression models. Meanwhile, LASSO and its variants have been adopted for variable selection in survival settings. For example, Tibshirani [13] applied the LASSO method in the Cox model and Ishwaran [14] extended it to the high-dimensional survival data. Liu and Zeng [15] proposed adaptive LASSO in general transformation models for right-censored data and Zou [16] proposed least absolute deviations (LAD) variable selection for linear models with randomly censored data. Recently, Liu, Wang and Wu [17] applied grouped LASSO in the Tobit censored response model.

However, there has been few attempts at using penalized regression on the Tobit models. Furthermore, we face additional challenges that the dimension of covariates is larger than the sample size. In this paper we develop an adaptive LASSO-based variable selection procedure for the Tobit model with high-dimensional covariates. The application of L_1 penalty on the Tobit model achieves the goal of variable selection and overcomes the potential problems when handling undetected observations using deletion, dichotomizing or imputation. The variable selection procedure has the oracle property [11] in the sense that it estimates as well as when zero components and nonzero components are known *a prior*.

The paper is organized as follows. Section 2 introduces the model and the estimation procedure and presents the theoretical results of the estimators. Section 2.3 discusses the computational algorithm and tuning parameter selection. Section 3 illustrates the method through the analysis of an HIV study and Section 4 describes simulation studies. A discussion is given in Section 5. Technical proofs are given in the Appendix.

2. MODEL AND METHODS

Consider the Tobit model [18]

$$Y_i^1 = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i,\ i = 1, 2, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_n})^{\mathrm{T}}$ is the vector of unknown parameters, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip_n})^{\mathrm{T}}$, $\varepsilon_i \sim N(0, \sigma^2)$, and Y_i^1 is a latent variable. The observable variable Y_i is defined as $Y_i = Y_i^1$ if $Y_i^1 > \tau$ and τ otherwise, where τ is a pre-specified value. For notational simplicity, without loss of generality, assume $\tau = 0$. That is, $Y_i = \max(Y_i^1, 0)$. Note that p_n, the dimension of parameters, is allowed to increase with the sample size n and it can be larger than n.

Let d_i be the undetected indicator; i.e., $d_i = 1$ if $Y_i > 0$ and 0 otherwise. Then the likelihood function based on the independent observations Y_i is given by

$$L(\mathbf{Y}; \mathbb{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \phi\left(\frac{Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma} \right) / \sigma \right\}^{d_i} \left\{ 1 - \Phi\left(\frac{\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma} \right) \right\}^{(1-d_i)},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\mathbb{X} = (\mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_n^{\mathrm{T}})^{\mathrm{T}}$, and ϕ and Φ are the density and distribution functions of the standard normal distribution, respectively. The log-likelihood function (up to a constant) is

$$l(\mathbf{Y}; \mathbb{X}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[d_i \left\{ -\log\sigma - \frac{(Y_i - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})^2}{2\sigma^2} \right\} + (1 - d_i) \log\left\{ 1 - \Phi\left(\frac{\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma} \right) \right\} \right]. \tag{2}$$

The first part of (2) is the log-likelihood for uncensored observations, and the second part is the log-likelihood for censored observations, for which we only know that they are smaller than 0.

2.1 Penalized-likelihood estimation

To achieve simultaneous variable selection and estimation for the Tobit model, we apply the adaptive Lasso method proposed by Zou [10] to the log-likelihood function given in (2). The adaptive LASSO estimators are defined as

$$\widehat{\boldsymbol{\beta}}_n(\lambda_n) = argmin \left\{ -l(\mathbf{Y}; \mathbb{X}, \boldsymbol{\beta}) + \lambda_n \sum_{j=2}^{p_n} w_j |\beta_j| \right\}, \tag{3}$$

where λ_n is the tuning parameter controlling the degree of the shrinkage and chosen by some data-driven method, w_j's are adaptive weights $1/|\widetilde{\beta}_{nj}|$, which we will discuss how to choose later. Note that we usually don't put any penalty

on the intercept. The LASSO penalty [9] is a special case of the adaptive LASSO penalty with all w_j being 1. The solution to (3) can be obtained through a locally quadratic approximation. We describe selection of tuning parameter and the estimation procedure below.

To find an initial estimator for providing appropriate weights, we conduct marginal regression (response vs each covariate, respectively) for uncensored observations. The reciprocal of corresponding estimators are served as the weights. This strategy is motivated by the conclusion drawn by Huang, Ma and Zhang [12] that the marginal regression estimators are good candidates for the weights without censored observations.

Without loss of generality, let $\mathbf{Y}_{n_1} = (Y_1, \ldots, Y_{n_1})$ be the collection of the response values for the uncensored observations. Let $\mathbb{X}_{n_1} = (\mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_{n_1}^{\mathrm{T}})^{\mathrm{T}}$ be the matrix containing the first n_1 rows of $\mathbb{X}$ and $\mathbf{X}_{n_1 \cdot j}$ be the j^{th} column of $\mathbb{X}_{n_1}$ for $1 \le j \le p_n$. The initial estimator of β_j is given as

$$\widetilde{\beta}_{nj} = (\mathbf{X}_{n_1 \cdot j}^{\mathrm{T}} \mathbf{X}_{n_1 \cdot j})^{-1} \mathbf{X}_{n_1 \cdot j}^{\mathrm{T}} \mathbf{Y}_{n_1}. \tag{4}$$

The simulation study shows that this initial estimator works very well. And it has been shown that using such initial estimators could theoretically guarantee that our adaptive LASSO estimators possesses an oracle property under certain regularity conditions, that is, with proper choice of the weights, the nonzero parameters can be correctly identified with probability approaching 1 and their estimators have the same asymptotic distributions as the estimators when the true model is known.

2.2 Tuning parameter selection

We know that the tuning parameter λ_n controls the degree of shrinkage. So, with a different λ_n, the estimates of β_j's would be different, especially for the components where exact zeros are present. Tuning parameter selection is equivalent to model selection or model comparison among multiple competing models. One of the most popular model selection criterions is Bayesian information criterion ([BIC 19]).

Denote $\boldsymbol{\beta}$ by $\boldsymbol{\beta}(s)$, whose components not in the candidate model s are set to 0 or some prespecified value. Define

$$\mathrm{BIC}(s) = -2 \log L(\widehat{\boldsymbol{\beta}}(s)) + k \log n \tag{5}$$

where $\widehat{\boldsymbol{\beta}}(s)$ is the maximum likelihood estimator of $\boldsymbol{\beta}(s)$ and k is the size of s, that is, the number of components in s. The model that minimizes $\mathrm{BIC}(s)$ is favored. However, the classical definition of BIC assumes constant prior and thus it would assign probabilities to S_j (the class of models with j covariates) proportional to their sizes. This would be strongly against the purpose of variable selection in the large space model scenario, for example, when sample size is smaller than the number of covariates under consideration, a model with a larger number of covariates would receive much higher prior probabilities than models with fewer covariates because the former would have much larger sizes and therefore the classic BIC defined in (5) would tend to select a model with many spurious covariates. This problem was first noticed by Broman and Speed [20] when they used BIC for quantitative trait loci mapping, and it was later observed [21, 22] as well. To improve the performance of BIC in variable selection in large model spaces, Chen and Chen [23] proposed a class of extended Bayesian information criteria, defined as

$$\mathrm{BIC}_\gamma(s) = -2 \log L(\widehat{\boldsymbol{\beta}}(s)) + k \log n + 2\gamma \log \binom{p}{k}, \tag{6}$$

for $0 \le \gamma \le 1$, where k is the number of covariates that are estimated to be nonzero, L is the corresponding likelihood value and p is the total number of covariates under consideration. They suggested $\gamma = 1 - 1/(2 \log_n p)$, and proved that this extended BIC is consistent in model selection. We, therefore, applied the extended BIC to select the tuning parameter in our numerical experiments.

2.3 Implementation algorithm

We now discuss the computational algorithm. The censored observations and the L_1 penalty complicate the search of $\boldsymbol{\beta}$ that minimizes the objective function (3). We apply the locally quadratic approximation for implementation [11] and use the coordinate descent algorithm [24] to find the minimizer.

Let $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}$ be the current estimates of $\boldsymbol{\beta}$ and σ, respectively. Using Taylor expansion on $l(\mathbf{Y}; \mathbf{X}, \boldsymbol{\beta})$, we have a locally quadratic approximation of $l(\mathbf{Y}; \mathbf{X}, \boldsymbol{\beta})$ on $\boldsymbol{\beta}$:

$$\begin{aligned} l_Q(\boldsymbol{\beta}) &= l(\mathbf{Y}; \mathbf{X}, \widetilde{\boldsymbol{\beta}}) + \sum_{i=1}^{n} \Big\{ w_{i1} \mathbf{X}_i^{\mathrm{T}} (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) \\ &\quad - \frac{1}{2} w_{i2} \left[\mathbf{X}_i^{\mathrm{T}} (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}) \right]^2 \Big\} \\ &= -\frac{1}{2} \sum_{i=1}^{n} w_{i2} \left(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta} - \mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}} - \frac{w_{i1}}{w_{i2}} \right)^2 + C(\widetilde{\boldsymbol{\beta}}), \end{aligned}$$

where $w_{i1} = d_i (Y_i - \mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}}) / \widetilde{\sigma}^2 - (1 - d_i) \widetilde{\nu}_i / \widetilde{\sigma}$, $w_{i2} = d_i / \widetilde{\sigma}^2 + (1 - d_i) \widetilde{\nu}_i (\widetilde{\nu}_i - \widetilde{z}_i) / \widetilde{\sigma}^2$, $\widetilde{z}_i = \mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}} / \widetilde{\sigma}$, $\widetilde{\nu}_i = \phi(\widetilde{z}_i) / \{1 - \Phi(\widetilde{z}_i)\}$, and $C(\widetilde{\boldsymbol{\beta}})$ is a constant. As a result, the penalized likelihood function can be rewritten as

$$\begin{aligned} Q(\boldsymbol{\beta}) &= -l_Q(\boldsymbol{\beta}) + \lambda \sum_{j=2}^{p_n} w_j |\beta_j| \\ &= \frac{1}{2} \left\{ \sum_{i=1}^{n} w_{i2} \left(\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta} - \mathbf{X}_i^{\mathrm{T}} \widetilde{\boldsymbol{\beta}} - \frac{w_{i1}}{w_{i2}} \right)^2 \right\} - C(\widetilde{\boldsymbol{\beta}}) \\ &\quad + \lambda \sum_{j=2}^{p_n} w_j |\beta_j|. \end{aligned} \tag{7}$$

Suppose we want to partially optimize (7) with respect to β_j given current estimates $\widetilde{\beta}_k^{[r]}$ for $k \neq j$ at the r^{th} step

of the iteration. Thus, we calculate the gradient at $\beta_j = \widetilde{\beta}_j^{[r+1]}$ for $\widetilde{\beta}_j^{[r]} \neq 0$, and obtain the expression of $\partial Q/\partial\beta_j$ at $\beta_k = \widetilde{\beta}_k^{[r]}, k \neq j, \beta_j = \widetilde{\beta}_j^{[r+1]}$ as follows:

$$\begin{aligned}
&\sum_{i=1}^{n} w_{i2}[\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}^{[r]} - \frac{w_{i1}}{w_{i2}}]X_{ij} \\
&\qquad + \lambda w_j \mathrm{sign}(\beta_j)\Big|_{\beta_k=\widetilde{\beta}_k^{[r]}, k\neq j, \beta_j=\widetilde{\beta}_j^{[r+1]}} \\
&= \sum_{i=1}^{n} w_{i2}[X_{ij}\widetilde{\beta}_j^{[r+1]} - X_{ij}\widetilde{\beta}_j^{[r]} - \frac{w_{i1}}{w_{i2}}]X_{ij} \\
&\qquad + \lambda w_j \mathrm{sign}(\beta_j).
\end{aligned}$$

Recall $w_1 = 0$ (no penalty imposed to the intercept). A simple calculation shows that the coordinate-wise update has the form

$$\widetilde{\beta}_j^{[r+1]} = S(\widetilde{\beta}_j^{[r]} + \frac{\sum_{i=1}^n w_{i1}^{[r]} X_{ij}}{\sum_{i=1}^n w_{i2}^{[r]} X_{ij}^2}, w_j\lambda), \qquad j = 1, \ldots, p_n,$$

where $S(z, \gamma)$ is the soft-thresholding operator with value

$$\mathrm{sign}(z)(|z| - \gamma)_+ = \begin{cases} z - \gamma & \text{if } z > 0 \text{ and } \gamma < |z|, \\ z + \gamma & \text{if } z < 0 \text{ and } \gamma < |z|, \\ 0 & \text{if } \gamma \geq |z|. \end{cases}$$

$\widetilde{\sigma}$ will be updated to maximize the log-likelihood (2) given the updated estimates $\widetilde{\beta}_j$'s, which can be easily implemented with existing optimization algorithm. The iteration can be repeated until all parameters converge according to some stopping criterion.

3. ANALYSIS OF THE HIV STUDY

A primary goal of vaccine strategies to prevent HIV infection is the induction of a protective humoral response. Some HIV infected patients develop potent serum antibodies that are able to neutralize a broad range of HIV isolates. By studying the characteristics of the B cells in such HIV-infected patients, mechanisms for the induction of potent neutralizing antibodies may be revealed. In this section, we apply the proposed method to a cross-sectional study focused on measuring B cell-related parameters in HIV infected patients with varying degrees of HIV neutralizing serum antibody. There are 59 variables including AGE, TIME, CD4, VLOAD, BCELL1-51, ANTIBODY1-4, while there are only 42 observations and among them one third of IC50 values are left censored at 20. We apply the proposed adaptive LASSO method with the Tobit model to analyze this dataset, and compare with three approaches aforementioned.

All covariates were logarithm transformed and then standardized such that linearity relationship is appropriate. For the response variable IC50, the neutralizing activity of the patients' sera against the Tier 2 HIV clade B virus 6535.3, centering was not appropriate due to the presence of censoring, and log-transformation was applied because there was some extreme values as large as 471 and as small as 20 (censored). One of the observations was deleted prior to analysis for its corresponding covariates value was an outlier and therefore there were 41 observations in our analysis. Our target is to find the covariates which have significant effects on IC50. To achieve this, we apply our proposed variable selection and estimation procedure to the Tobit model with all of the 59 variables as predictors and IC50 as the response variable. The extended BIC given in Section 2.3 was used and the selected tuning parameter was 0.105.

The selected variables and their estimated coefficients are listed in the last column of Table 1, where the standard errors were obtained by using the expression given in Theorem A.2. Our proposed method adaptive LASSO with Tobit model identified 3 variables which are subsets of B cells identified by flow cytometry and defined by their expression of surface proteins. These included BCELL6, (CD19+CD20+, percentage of total B cells. See [25] for more details), BCELL8 (CD19-CD20+, percentage of total B cells), and BCELL49 (IgD-B220-CD27-, percentage of IgD- B cells), and LASSO with Tobit model (the LASSO penalty with weight $w_j = 1$ is applied to the Tobit model) detected an additional covariate ANTIBODY2 (anti-dsDNA). To compare the results, we also used other selection procedures including (i) adaptive LASSO with deletion of the censored observations; (ii) adaptive LASSO with threshold imputation; (iii) LASSO with deletion of the censored observation and (iv) LASSO with threshold imputation. R function `glmnet` was applied for cases (i)–(iv) and the tuning parameter was selected by both 5-fold cross-validation and extended BIC. When excluding the censored observation, no covariate was selected except the intercept so we didn't list the results for the deletion method.

Clearly, with such a small number of observations and a large number of parameters, simply excluding the undetected observations dramatically shrinks the sample size and therefore no significant covariates can be identified. Imputing the undetected values with threshold identified two variables: BCELL8 and BCELL49. We can observe that BCELL49 is detected by all the methods as shown in Table 1. The covariate BCELL6 which is selected by the (adaptive) LASSO is not selected by the imputation method. BCELL6 and BCELL8 are directly inversely linked to each other as they are defined with the same markers CD19 and CD20, since BCELL8 is the minor population perhaps it is a more sensitive indicator.

The model has resolved additional features associated with HIV neutralizing activity, although the data set was limited to examining activity against only a single viral isolate, intriguing biological insight was obtained. The B cell subset described by BCELL49, the IgD-B220-CD27- population is the dominate subset within the IgD-CD27- pop-

Table 1. Estimated values and the associated standard errors (se) obtained by using LASSO and adaptive LASSO for the HIV study

	LASSO			Adaptive LASSO		
	Imputation		Tobit	Imputation		Tobit
	eBIC	CV		eBIC	CV	
BCELL6			0.058(0.389)			0.233(0.397)
BCELL8		−0.042(0.137)	−0.255(0.285)	−0.185(0.137)	−0.179(0.137)	−0.318(0.291)
BCELL49	−0.230(0.132)	−0.279(0.127)	−0.506(0.219)	−0.406(0.185)	−0.398(0.127)	−0.555(0.224)
ANTIBODY2			0.113(0.185)			

ulation, which is infrequently observed in healthy subjects, however increases in instances of B cell dysregulation such as in patients with Systemic Lupus Erythematosus [26]. This population overlaps with CD21-CD27- B cells which have been described as "tissue-like memory" and "exhausted" and expanded in viremic HIV patients [4], although its association with HIV neutralization is unknown. The negative correlation of the IgD-B220-CD27- subset with IC50 HIV neutralizing activity may suggest B cell exhaustion negatively impacts the development and maintenance of HIV neutralizing antibodies. Similarly, a negative correlation of BCELL8 (CD19-CD20+ B cells) with IC50 was observed. The CD19-CD20+ subset is a rare and poorly studied subset, and may also be a hallmark of a dysregulated B cell compartment. The positive correlation of ANTIBODY2 (anti-dsDNA) with IC50, HIV neutralizing activity is consistent with previous observations of increased auto-reactive antibodies in patients with increased HIV neutralizing activity [4, 27, 28], which may in part be a direct consequence of a population of HIV-specific antibodies also having reactivity to these self antigens, a phenomenon that has been observed previously [29, 30]. A critical goal of HIV vaccine strategies is to induce antibodies with neutralizing activity against multiple HIV strains, and as such it will be important to extend our model to incorporating multiple IC50 parameters.

4. SIMULATION STUDIES

In this section the proposed variable selection and estimation procedure is evaluated by Monte Carlo simulation studies through assessing accuracy of variable selection and prediction performance measured mean square error. That is, we evaluate the frequency of correctly identifying zero and nonzero coefficients, and the discrepancy between the predicted and the true values of responses, the latter is evaluated in an independent test sample. We compare the results obtained by applying adaptive LASSO to the linear regression model after deleting the censored observations (deletion), or replacing the unobserved values with the threshold (imputation) or the Tobit model, respectively. Marginal regression estimators are used as the initial estimators in the simulation study. The tuning parameter is selected by using the extended BIC described in Section 2.3.

The data are generated from the linear model $Y_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i$, where ε_i are generated independently from $N(0, \sigma^2)$ with $\sigma = 1.5$. Eight examples with $p_n > n$ are considered, representing eight different and commonly encountered scenarios. In each case, the covariate vector is normally distributed with mean zero and the covariance matrix is specified below. To examine the performance of adaptive LASSO with Tobit models under different censoring rate, we let Y be censored at two different values in each case, corresponding to approximately 15% and 38% censoring rates, respectively. Summary statistics are computed based on 100 replications.

The eight simulation examples considered are given as follows:

Ex 1. $p = 81$ and $n = 40$. For the i^{th} row of $\mathbb{X}$, $X_{i1} = 1$ is for the intercept, the first 9 covariates $(X_{i,2}, \ldots, X_{i,10})$ and the remaining 71 covariates $(X_{i,11}, \ldots, X_{i,81})$ are independent; The pairwise correlation between the k^{th} and the j^{th} components of $(X_{i,2}, \ldots, X_{i,10})$ for $k, j = 2, \ldots, 10$; and of $(X_{i,11}, \ldots, X_{i,81})$ for $k, j = 11, \ldots, 81$ is $r^{|k-j|}$ with $r = 0.5$. $\beta_1 = 5, \beta_2 = \beta_3 = \beta_4 = 2.5$, $\beta_5 = \beta_6 = \beta_7 = 1.5$, $\beta_8 = \beta_9 = \beta_{10} = 0.5$, and $\beta_j = 0$ for $11 \le j \le 81$.

Ex 2. The same as Example 1, except that $r = .95$.

Ex 3. The same as Example 1, except that $p = 201$ and $n = 100$.

Ex 4. The same as Example 3, except that $r = .95$.

Ex 5. $p = 81$ and $n = 40$; the pairwise correlation between the j^{th} and the k^{th} components of $(X_{i,2}, \ldots, X_{i,81})$ is $r^{|j-k|}$ with $r = 0.5$, $j, k = 2, \ldots, 81$; and $\beta_1 = 5, \beta_2 = \beta_3 = \beta_4 = 2.5$, $\beta_5 = \beta_6 = \beta_7 = 1.5$, $\beta_8 = \beta_9 = \beta_{10} = 0.5$, and $\beta_j = 0$ for $11 \le j \le 81$.

Ex 6. The same as Example 5, except that $r = .95$.

Ex 7. The same as Example 5, except that $p = 201$ and $n = 100$.

Ex 8. The same as Example 7, except that $r = .95$.

In all examples, the sample size is smaller than the number of unknown coefficients. The values 2.5, 1.5 and 0.5 correspond to strong, moderate and weak coefficients. It is worth pointing that partial orthogonal condition is satisfied in Example 1–4, yet in Example 5–8 since the covariates with nonzero coefficients are correlated to the rest. Furthermore,

Table 2. Simulation results. C: median of number of correctly identifying zero coefficients. I: median of number of incorrectly missing the nonzero coefficients. Low: relatively lower censor rate, inside "()" are the approximated mean censor rate. High: relatively higher censor rate. ALasso: adaptive LASSO for Tobit model

n	p	r	Method	Low (15%)			High (38%)		
				C	I	PMSE	C	I	PMSE
Examples 1–4									
40	81	0.5	Deletion	70	2	4.14	71	3	5.80
			Imputation	71	3	9.80	71	4	25.97
			ALasso	71	2	3.84	60	2	4.82
40	81	0.95	Deletion	71	2	2.91	71	3	4.11
			Imputation	71	3	5.48	71	6	59.49
			ALasso	71	1	2.77	70	2	3.62
100	201	0.5	Deletion	191	1	3.09	190	2	4.64
			Imputation	190	2	4.83	191	3	21.95
			ALasso	191	1	2.91	191	2	3.53
100	201	0.95	Deletion	191	2	2.60	191	2	3.08
			Imputation	191	2	6.57	191	4	37.98
			ALasso	191	1	2.55	191	1	2.66
Examples 5–8									
40	81	0.5	Deletion	71	2	4.18	71	2	5.75
			Imputation	71	2	7.60	71	4	25.88
			ALasso	71	2	3.78	70	2	4.64
40	81	0.95	Deletion	71	3	3.37	71	3	3.63
			Imputation	71	4	7.83	71	5	46.05
			ALasso	70	2	3.01	70	3	3.45
100	201	0.5	Deletion	191	1	2.92	191	1	4.71
			Imputation	191	1	4.74	191	3	24.25
			ALasso	191	1	2.76	191	1	3.23
100	201	0.95	Deletion	191	2	2.73	191	2	3.12
			Imputation	191	3	11.11	191	4	55.59
			ALasso	190	1	2.60	190	1	2.67

Examples 1, 3, 5, 7 have moderately to weakly correlated covariates and Examples 2, 4, 6 and 8 have strongly correlated covariates. For each example, using the three different methods: deletion, imputation and adaptive LASSO with Tobit model, we report 3 values: C, the number of correctly identifying zero coefficients; I, the number of incorrectly missing the nonzero coefficients, and the estimated prediction mean square errors (PMSE) defined below.

In each example, we generate a training set and a test set. The tuning parameter is selected using the extended BIC described in Section 2.3 with the training set only. After selecting the tuning parameter, the adaptive LASSO estimates are computed using the training set. We calculate the PMSE using the test set, which is defined as $\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2/n$ with $\widehat{Y}_i (= X_i\widehat{\beta})$ being obtained by using the training set estimate and Y_i being the independent test set. The reported values are the median from 100 replications.

The results are presented in Table 2, where we can observe that the imputation method has comparable capability to detect zero coefficients but is more likely to miss the important nonzero coefficients. In terms of PMSE, imputation is the worst. This is because when simply replacing the unobserved values with the threshold, the resulting initial estimator may be far away from the true value and thus the weight applied to the L_1 penalty is inappropriately chosen, which causes estimators of the parameter severely tortured, and leads to larger biases. Deletion method has worse performance than the proposed method, especially when sample size is small and censor rate is large. This is not surprising because our method uses the censored information and therefore we have more accurate variable selection results. When partial orthogonality holds, the performance of adaptive LASSO for Tobit model is better than that when this assumption doesn't hold. We also notice that with stronger correlation ($r = 0.95$), the PMSE values based on our proposed method are smaller. This trend has been observed in the literature for similar simulation settings [12, 31].

5. DISCUSSION

We have proposed an adaptive LASSO procedure for simultaneous variable selection and estimation in sparse high-dimensional Tobit models. We have shown that the adaptive LASSO estimator for Tobit models has the oracle property under certain regularity conditions such that the model can be correctly selected with probability approaching 1 and the

estimators for the nonzero coefficients have the same asymptotic distributions as the estimators if the true model were known. Moreover, we develop a Newton-Ralphson computational algorithm by combining locally quadratic approximation and the coordinate descent algorithm. In both simulation studies and the real data applications, we illustrate that the proposed method has superior performance over the commonly used approaches such as the deletion and imputation methods. Our method provides an intuitively appealing, theoretically reliable, and computationally efficient tool for the analysis of data with censored observations and high-dimensional covariates. Recalling the discussions on the deficiency of the deletion and imputation methods, and the results based on these two methods for the real dataset, we expect that the Tobit model and the associated method should increase the power. However, theoretically justifying such a superiority is very difficult, if is not impossible. We hope this can be addressed in the future. The method can be extended to semiparametric Tobit models to relax the linearity assumption of the relationship between response and covariates, which will be the focus of our future work.

We are concerned with the case that the threshold is a constant. We have not studied the case with different censoring values. It should be possible to extend our method in a similar way. The procedure would change only in the sense that the likelihood function would be modified to a more complex expression. However, the obvious analogue of main results presented in the Appendix holds in such a case.

It would appear possible in principle to extend the method to a nonlinear relationship. We need only replace the likelihood function accordingly. However, in general nonlinear models, it is not guaranteed that the minimizer is well-defined, and this would appear to be the difficulty in extending our approach. We expect that the theoretical justification and numerical implementation can be achieved without substantial difficulties. The detailed investigation of these issues is interesting, but beyond the scope of this paper.

We have compared the situations of total deletion and replacement by the threshold. There are alternatives proposed in the literature. For example, replacement by half threshold, or threshold divided by root 2, or even more complex methods such as probabilistic imputation [32] of the values below the detection limit by the left tail distribution of the response variable. This is currently under our investigation. We have used PMSE for a comparison of several methods. Alternative measures proposed in the literature [33] may be used for comparison of prediction errors as well.

It should bear in mind that the oracle property of the proposed estimators only holds with the assumption of partial orthogonality; i.e, the covariates with zero and nonzero coefficients only have a weak correlation, which may not be satisfied sometimes. Interpretation should be careful when this assumption doesn't hold, though our simulation study indicates that the numerical performance of the proposed procedure still works well when even covariates with zero and nonzero coefficients have a moderate or even strong correlation.

APPENDICES

A.5.1 Initial statements

Let $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^{\mathrm{T}}$ denote the true parameters. Assume that model (1) is sparse, that is, some components in $\boldsymbol{\beta}_0$ are exactly zero corresponding to predictors that are irrelevant to the response. Without loss of generality, we assume that the true model has parameters $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^{\mathrm{T}}, \boldsymbol{\beta}_{02}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{01}$ is the $k_n \times 1$ vector with nonzero components and $\boldsymbol{\beta}_{02}$ is the $(p_n - k_n) \times 1$ vector with zero components, and k_n is much smaller than $(p_n - k_n)$. Correspondingly, we write $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n1}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}_{n2}^{\mathrm{T}})^{\mathrm{T}}$, where $\widehat{\boldsymbol{\beta}}_{n1}$ and $\widehat{\boldsymbol{\beta}}_{n2}$ are the estimators of $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\beta}_{02}$, respectively.

Since we have censored observations in the response, it is not appropriate to center $\mathbf{Y}$. We, however, can still center and standardize the covariates as

$$\sum_{i=1}^{n} X_{ij} = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_{ij}^2 = 1, \ j = 2, \ldots, p_n.$$

Let $\mathbf{X}_{\cdot j} = (X_{1j}, \ldots, X_{nj})^{\mathrm{T}}$ for $j = 1, \ldots, p_n$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip_n})^{\mathrm{T}}$ for $i = 1, \ldots, n$, $\mathbb{X} = (\mathbf{X}_{\cdot j}, 1 \le j \le p_n)_{n \times p_n}$ and $\mathbf{D}$ be the diagonal matrix with diagonal elements $I(Y_i > 0) \triangleq d_i$. Let $J_{n1} = \{j : \beta_{0j} \neq 0\}$ and $\mathbb{X}^1 = (\mathbf{X}_{\cdot j}, j \in J_{n1})_{n \times k_n}$, which is the design matrix corresponding to the nonzero coefficients. Denote $\mathbf{X}_i^1 = (X_{ij}, j \in J_{n1})^{\mathrm{T}}$ and $\Sigma_{n1} = n^{-1}(\mathbb{X}^1)^{\mathrm{T}}\mathbb{X}^1$. Write $Z_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0/\sigma$, $\phi_i = \phi(Z_i)$, $\Phi_i = \Phi(Z_i)$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^{\mathrm{T}}$ where $\nu_i = \phi_i/(1 - \Phi_i)$. For any vector $\mathbf{c} = (c_1, c_2, \ldots, c_s)^{\mathrm{T}}$, its sign vector is denoted by $\mathrm{sgn}(\mathbf{c}) = \{\mathrm{sgn}(c_1), \mathrm{sgn}(c_2), \ldots, \mathrm{sgn}(c_s)\}^{\mathrm{T}}$, with the convention that $\mathrm{sgn}(0) = 0$. Following Zhao and Yu [34], we say that $\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}$ if and only if $\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_n) = \mathrm{sgn}(\boldsymbol{\beta})$. Define

$$b_{n1} = \min\{|\beta_{0j}| : j \in J_{n1}\} \tag{A.8}$$

A.5.2 Assumptions

(A1) $n^{-1/2}\max_{1 \le i \le n}\{(\mathbf{X}_i^1)^{\mathrm{T}}\mathbf{X}_i^1\}^{1/2} \to 0$, as $n \to \infty$ and there exists a constant $0 < \tau_2$, such that $\tau_{n2} \ge \tau_2$ for all n, where τ_{n2} is the smallest eigenvalue of $\mathbf{A} = n^{-1}\{\sum_{i=1}^{n} \mathbf{X}_i^1(\mathbf{X}_i^1)^{\mathrm{T}}\sigma_{1i}^2\}$ with $\sigma_{1i} = \sigma^2\{\Phi_i - Z_i\phi_i + \phi_i^2/(1 - \Phi_i)\}$.

(A2) The initial estimators $\widetilde{\beta}_{nj}$ are r_n-consistent for the estimation of certain η_{nj}:

$$r_n \max_{1 \le j \le p_n} |\widetilde{\beta}_{nj} - \eta_{nj}| = O_P(1), \ r_n \to \infty,$$

where η_{nj} are unknown constants depending on $\boldsymbol{\beta}$ and satisfy

$$\max_{j \notin J_{n1}} |\eta_{nj}| \le M_{n2},$$

$$\left\{\sum_{j \in J_{n1}} \left(\frac{1}{|\eta_{nj}|} + \frac{M_{n2}}{|\eta_{nj}|^2}\right)^2\right\}^{1/2} \le M_{n1} = o(r_n).$$

(A3) (Adaptive irrepresentable condition) Define $\boldsymbol{s}_{n1} = \{|\eta_{nj}|^{-1}\text{sgn}(\beta_{0j}), j \in J_{n1}\}^{\mathrm{T}}$. There exists a constant $0 < \kappa < 1$ such that

$$n^{-1}|\mathbf{X}_{\cdot j}^{\mathrm{T}}\mathbb{X}^1\Sigma_{n1}^{-1}\boldsymbol{s}_{n1}| \le \frac{\kappa}{|\eta_{nj}|}, \ \forall j \notin J_{n1}.$$

(A4) The tuning parameter λ_n and the number of nonzero (k_n) and zero ($p_n - k_n$) coefficients satisfy the following order requirements:

$$\frac{(\log k_n)^{1/2}}{\sqrt{n}b_{n1}} + \{\log(p_n - k_n)\}^{1/2}\frac{\sqrt{n}}{\lambda_n}(M_{n2} + \frac{1}{r_n}) + \frac{M_{n1}\lambda_n}{b_{n1}n} \to 0.$$

(A5) There exists a constant $0 < \tau_1$, such that $\tau_{n1} \ge \tau_1$ for all n, where τ_{n1} is the smallest eigenvalue of Σ_{n1}.

Remark: Condition (A1) is needed for the proof of asymptotic normality of the estimators for the nonzero coefficients. It makes restriction on k_n implicitly, for example, if the covariates in $\mathbf{X}_i^1$ are bounded below by a constant $0 < c_0 < 1$, then $(\mathbf{X}_i^1)^{\mathrm{T}}\mathbf{X}_i^1 \ge k_n c_0^2$, and thus we must have $k_n = o(n)$ to ensure Condition (A1) holds. Because we assume sparsity in the true model, this condition is reasonable. Condition (A2) assumes that the initial estimator $\widetilde{\beta}_{nj}$ can, at least, estimate some proxy η_{nj} of β_{0j}, so that as the sample size grows, the weights $w_j = |\widetilde{\beta}_{nj}|^{-1} \approx |\eta_{nj}|^{-1}$ for predictors with zero coefficients is not too small, and the weights for predictors with nonzero coefficients is not too large. In Condition (A3), constraints on η_{nj} are imposed so that it performs as a surrogate of β_{0j}. Condition (A4) is the order requirement of the tuning parameter and the number of zero and nonzero coefficients. It restricts the number of covariates allowed in the model. For example, we usually have r_n increasing somewhat slower than $\sqrt{n}$ so $n^{\delta-1/2}r_n \to \infty$ for some small $\delta > 0$ and $\lambda_n = n^a$ for some $0 < a < 1$. If assuming $1/b_{n1} = O(1)$ and $M_{n1} = O(\sqrt{k_n})$, by condition (A4), the number $(p_n - k_n)$ of zero coefficients can be as large as $\exp(n^{2(a-\delta)})$. While the number of nonzero coefficient k_n can only be allowed of the order $\min\{n^{2(1-a)}, n^{1-2\delta}\}$. Condition (A5) requires that the smallest eigenvalue of Σ_{n1} is bounded away from zero but does not put any restriction on its largest eigenvalue and is reasonable under sparsity assumption. Condition (A4) is a special case when $d = 2$ as in Huang, Ma and Zhang [12] and Conditions (A2), (A3) and (A5) have been imposed by these authors.

A.5.3 Preliminary lemmas

The following lemmas are used in the proofs of the theorems presented in Sections A.5.4 and A.5.5. Lemma A.1 is a variation of Lemma 1 of Huang, Ma and Zhang [12] in their online supplement of that article. Lemma A.2 is the same as theirs [12]. Let $\psi_d(x) = \exp(x^d) - 1$ for $d \ge 1$. The ψ_d-Orlicz norm of random variable X is defined as $\|X\|_{\psi_d} = \inf\{C > 0 : E\{\psi_d(|X|/C)\} \le 1\}$. More details about Orlicz norm can be found in [35].

Lemma A.1. *Suppose $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables with $E\varepsilon_i = 0$ and $Var(\varepsilon_i) \le \sigma^2$. Furthermore, suppose that their tail probabilities satisfy $P(|\varepsilon_i| > x) \le K\exp(-Cx^2)$, $i = 1, \ldots, n$, for constants C and K. Then, for all constants a_i satisfying $\sum_{i=1}^n a_i^2 = 1$,*

$$\|\sum_{i=1}^n a_i\varepsilon_i\|_{\psi_2} \le K_1\{\sigma + (1+K)^{1/2}C^{-1/2}\}$$

where K_1 is a constant. Consequently

$$q_n^*(t) = \sup_{a_1^2+\cdots+a_n^2=1} P\{\sum_{i=1}^n a_i\varepsilon_i > t\} \le \exp(-\frac{t^2}{M})$$

for a certain constant M depending on $\{K, C\}$ only.

Proof. Since ε_i satisfies $P(|\varepsilon_i| > x) \le K\exp(-Cx^2)$ for all $x > 0$ and some positive constants K and C, then it follows for its Orlicz norm $||\varepsilon_i||_{\psi_2} \le \{(1+K)/C\}^{1/2}$ by using Lemma 2.2.1 of van der Vaart and Wellner [35]). According to Proposition A.1.6 [35], there exists a constant K_1 such that

$$\begin{aligned}
\|\sum_{i=1}^n a_i\varepsilon_i\|_{\psi_2} &\le K_1\{E|\sum_{i=1}^n a_i\varepsilon_i| + [\sum_{i=1}^n ||a_i\varepsilon_i||_{\psi_2}^2]^{1/2}\} \\
&\le K_1\{[E(\sum_{i=1}^n a_i\varepsilon_i)^2]^{1/2} \\
&\quad + (1+K)^{1/2}C^{-1/2}[\sum_{i=1}^n |a_i|^2]^{1/2}\} \\
&\le K_1\{[Var(\sum_{i=1}^n a_i\varepsilon_i)]^{1/2} \\
&\quad + (1+K)^{1/2}C^{-1/2}[\sum_{i=1}^n |a_i|^2]^{1/2}\} \\
&\le K_1\{\sigma + (1+K)^{1/2}C^{-1/2}\}.
\end{aligned}$$

Based on the definition of $||X||_{\psi_2}$ we have $E\{\exp(|X|/||X||_{\psi_2})^2 - 1\} = E\{\psi_2(\frac{|X|}{||X||_{\psi_2}})\} \le 1$ and therefore

$$\begin{aligned}
P(X > t||X||_{\psi_2}) &= P(\frac{X}{||X||_{\psi_2}} > t) \\
&\le \exp(-t^2)\left\{1 + E\{\psi_2\left(\frac{X}{||X||_{\psi_2}}\right)\}\right\} \\
&\le 2\exp(-t^2), \ \forall t > 0
\end{aligned}$$

The last inequality in the lemma is an immediate consequence of this inequality.

Let $\tilde{\mathbf{s}}_{n1} = (|\tilde{\beta}_{nj}|^{-1}\text{sgn}(\beta_{0j}), j \in J_{n1})'$ and $\mathbf{s}_{n1} = (|\eta_{nj}|^{-1}\text{sgn}(\beta_{0j}), j \in J_{n1})'$. □

Lemma A.2 (*Huang, Ma and Zhang [12]*). *Suppose (A2) holds. Then,*

(A.9)
$$||\tilde{\mathbf{s}}_{n1}|| = (1+o_P(1))M_{n1},\ \max_{j\notin J_{n1}} \left|\left||\tilde{\beta}_{nj}|\tilde{\mathbf{s}}_{n1} - |\eta_{nj}|\mathbf{s}_{n1}\right|\right| = o_P(1).$$

Lemma A.3. *Recall $Y_i^1 = \mathbf{X}_i^T\boldsymbol{\beta}_0 + \varepsilon_i$, for $i = 1,\ldots,n$, where ε_i's are independent distributed with $\varepsilon_i \sim N(0,\sigma^2)$. Write $Y_i^* = d_iY_i^1 + (1-d_i)(\mathbf{X}_i^T\boldsymbol{\beta}_0 - \sigma\nu_i)$ and $\varepsilon_i^* = Y_i^* - \mathbf{X}_i^T\boldsymbol{\beta}_0, i = 1,\ldots,n$. Then,*

1. *$\varepsilon_1^*,\ldots,\varepsilon_n^*$ are independent distributed with $E(\varepsilon_i^*) = 0$ and $Var(\varepsilon_i^*) \le \sigma^2$*
2. *There exists constants K and C such that $P(|\varepsilon_i^*| > t) \le K\exp(-Ct^2)$*

Consequently, Lemma A.1 can be applied to such ε_i^*'s.

Proof. According to Amemiya [36], when $Y_i > 0$, we have

$$Y_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + U_i^*, \tag{A.10}$$

where U_i^* is the random variable with the density $h(u)$ given by $h(u) = 1/\Phi_i 1/\sqrt{2\pi\sigma^2}\exp\{-(u/\sigma)^2/2\}$, $-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 < u < \infty$ with $E(U_i^*) = \sigma\phi_i/\Phi_i$ and $E(U_i^{*2}) = (\sigma^2 - \sigma^2 Z_i\phi_i/\Phi_i)$ and is independent of d_i. Therefore, Y_i^* can be re-expressed as

$$\begin{aligned} Y_i^* &= d_i(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + U_i^*) + (1-d_i)(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 - \sigma\nu_i) \\ &= \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 + d_iU_i^* - \sigma(1-d_i)\nu_i, \end{aligned}$$

where $\nu_i = \phi_i/(1-\Phi_i)$. Therefore, $\varepsilon_i^* = Y_i^* - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 = d_iU_i^* - \sigma(1-d_i)\nu_i$. It is ready to verify that $E(d_i) = \Phi_i$ so $E(\varepsilon_i^*) = 0$ and $Var(\varepsilon_i^*) = \sigma^2 - Var(Y_i|Y_i < -\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0) \le \sigma^2$ (See [37]). The first part is then proved.

For the tail probability, when $t > \max_i\{\sigma|Z_i|, \sigma\nu_i\}$,

$$\begin{aligned} P(|\varepsilon_i^*| > t) &= P(\varepsilon_i^* > t) + P(\varepsilon_i^* < -t) \\ &= P(\varepsilon_i^* > t, d_i = 1) + P(\varepsilon_i^* > t, d_i = 0) \\ &\quad + P(\varepsilon_i^* < -t, d_i = 1) \\ &\quad + P(\varepsilon_i^* < -t, d_i = 0) \\ &= P(d_i = 1)P(U_i^* > t|d_i = 1) \\ &= P(d_i = 1)P(U_i^* > t) \\ &= 1 - \Phi(t/\sigma), \end{aligned}$$

and there exists constants K_1 and C_1 such that $1-\Phi(t/\sigma) \le K_1\exp(-C_1t^2)$. When $0 \le t \le \max_i\{\sigma|Z_i|, \sigma\nu_i\}$, we can find constants K_2 and C_2 satisfying $K_2\exp(-C_2t^2) \ge 1$ and therefore let $K = \max\{K_1, K_2\}, C = \min\{C_1, C_2\}$. We finish the proof of the second part. □

A.5.4 Statistical properties of the proposed estimators

Theorem A.1 (*Consistency in variable selection*). *Suppose that conditions (A2)–(A5) hold. Let $\widehat{J}_{n1} = \{j : \widehat{\beta}_j \neq 0\}$. Then $lim_{n\to\infty}P(\widehat{J}_{n1} = J_{n1}) = 1$, or equivalently, $lim_{n\to\infty}P(\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0) = 1$.*

This theorem shows that zero coefficients can be correctly identified with probability tending to 1, so the adaptive LASSO method for the Tobit model has the model selection consistency property. The following theorem presents the asymptotic normality of the adaptive LASSO estimators for the nonzero coefficients in the Tobit model. Write $s_n^2 = n^{-1}\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}\{\sum_{i=1}^n(\mathbf{X}_i^1)(\mathbf{X}_i^1)^{\mathrm{T}}\sigma_{1i}^2\}\Sigma_{n1}^{-1}\boldsymbol{\alpha}_n$ with $\sigma_{1i}^2 = \sigma^2\{\Phi_i - Z_i\phi_i + \phi_i^2/(1-\Phi_i)\}$.

Theorem A.2 (*Asymptotic normality*). *Suppose that conditions (A1)–(A5) hold. For any $k_n \times 1$ vector $\boldsymbol{\alpha}_n$ satisfying $\boldsymbol{\alpha}_n^T\boldsymbol{\alpha}_n \le 1$. If $M_{n1}\lambda_n/\sqrt{n} = o(1)$, then $n^{1/2}s_n^{-1}\boldsymbol{\alpha}_n^T(\widehat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{01}) \to_D N(0,1)$.*

Proof of Theorem A.1. From Karush-Kunh-Tucker conditions we know that $\widehat{\boldsymbol{\beta}}_n = (\widehat{\beta}_{n1},\ldots,\widehat{\beta}_{np_n})'$ is the unique solution of the adaptive LASSO if

$$\begin{cases} \mathbf{X}_{\cdot j}^{\mathrm{T}}(\mathbf{Y}^* - \mathbb{X}\widehat{\boldsymbol{\beta}}_{\mathbf{n}}) = \sigma^2\lambda_n w_{nj}\mathrm{sgn}(\widehat{\beta}_{nj}), & \widehat{\beta}_{nj} \neq 0, \\ |\mathbf{X}_{\cdot j}^{\mathrm{T}}(\mathbf{Y}^* - \mathbb{X}\widehat{\boldsymbol{\beta}}_{\mathbf{n}})| < \sigma^2\lambda_n w_{nj}, & \widehat{\beta}_{nj} = 0, \end{cases} \tag{A.11}$$

where $\mathbf{Y}^* = (Y_1^*,\ldots,Y_n^*)^{\mathrm{T}}$ is the $n \times 1$ vector with $Y_i^* = d_iY_i + (1-d_i)(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_n - \sigma\frac{\phi_i}{1-\Phi_i})$. The vectors $\{\mathbf{X}_{\cdot j} : \widehat{\beta}_{nj} \neq 0\}$ are linearly independent. Let $\widetilde{\boldsymbol{s}}_{n1} = (w_{nj}\mathrm{sgn}(\beta_{0j}), j \in J_{n1})^{\mathrm{T}}$ and

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{n1} &= \{(\mathbb{X}^1)^{\mathrm{T}}\mathbb{X}^1\}^{-1}\{(\mathbb{X}^1)^{\mathrm{T}}\mathbf{Y}^* - \sigma^2\lambda_n\widetilde{\boldsymbol{s}}_{n1}\} \\ &= \boldsymbol{\beta}_{01} + \frac{1}{n}\Sigma_{n1}^{-1}\{(\mathbb{X}^1)^{\mathrm{T}}\boldsymbol{\varepsilon}^* - \sigma^2\lambda_n\widetilde{\boldsymbol{s}}_{n1}\}, \end{aligned} \tag{A.12}$$

where $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*,\ldots,\varepsilon_n^*)$ with $\varepsilon_i^* = Y_i^* - (\mathbf{X}_i^1)^{\mathrm{T}}\boldsymbol{\beta}_{01}$. Notice that $(\mathbf{X}_i^1)^{\mathrm{T}}\boldsymbol{\beta}_{01} = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0$ so $\varepsilon_i^* = Y_i^* - \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0 = d_iU_i - \sigma(1-d_i)\nu_i$ as described in Lemma A.3. If $\widehat{\boldsymbol{\beta}}_{n1} =_s \boldsymbol{\beta}_{01}$, then the equation in (A.11) holds for $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n1}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$. Thus for this particular $\widehat{\boldsymbol{\beta}}_n$ we have $\mathbb{X}\widehat{\boldsymbol{\beta}}_n = \mathbb{X}^1\widehat{\boldsymbol{\beta}}_{n1}$ and $\{\mathbf{X}_{\cdot j}^{\mathrm{T}}, j \in J_{n1}\}^{\mathrm{T}}$ are linearly independent. Therefore, $\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0$ if

$$\begin{cases} \widehat{\boldsymbol{\beta}}_{n1} =_s \boldsymbol{\beta}_{01}, & \forall j \in J_{n1} \\ |\mathbf{X}_{\cdot j}^{\mathrm{T}}(\mathbf{Y}^* - \mathbb{X}^1\widehat{\boldsymbol{\beta}}_{n1})| < \sigma^2\lambda_n w_{nj}, & \forall j \notin J_{n1}. \end{cases} \tag{A.13}$$

Write $\mathbf{H}_n = \mathbf{I}_n - \mathbb{X}^1\Sigma_{n1}^{-1}(\mathbb{X}^1)^{\mathrm{T}}/n$, which is the projection to the null of $(\mathbb{X}^1)^{\mathrm{T}}$. From (A.12) we have $\mathbf{Y}^* - \mathbb{X}^1\widehat{\boldsymbol{\beta}}_{n1} = (\mathbf{Y}^* - \mathbb{X}^1\boldsymbol{\beta}_{01}) - \mathbb{X}^1(\widehat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{01}) = \mathbf{H}_n\boldsymbol{\varepsilon}^* + \sigma^2\mathbb{X}^1\Sigma_{n1}^{-1}\widetilde{\mathbf{s}}_{n1}\lambda_n/n$, and thus we have, $\widehat{\boldsymbol{\beta}}_n =_s \boldsymbol{\beta}_0$ if

$$\begin{cases} \mathrm{sgn}(\beta_{0j})(\beta_{0j} - \widehat{\beta}_{nj}) < |\beta_{0j}|, & \forall j \in J_{n1} \\ |\mathbf{X}_{\cdot j}^{\mathrm{T}}(\mathbf{H}_n\boldsymbol{\varepsilon}^* + \sigma^2\mathbb{X}^1\Sigma_{n1}^{-1}\widetilde{\mathbf{s}}_{n1}\frac{\lambda_n}{n})| < \sigma^2\lambda_n w_{nj}, & o.t. \end{cases} \tag{A.14}$$

Therefore, by (A.12) and (A.14), for any $0 < \kappa < \kappa + \epsilon < 1$

$$P\left\{\widehat{\boldsymbol{\beta}}_n \neq_s \boldsymbol{\beta}_0\right\} \le P\left\{\frac{1}{n}|\mathbf{e}_j^{\mathrm{T}}\Sigma_{n1}^{-1}\left(\mathbb{X}^1\right)^{\mathrm{T}}\boldsymbol{\varepsilon}^*| \ge \frac{|\beta_{0j}|}{2} \right.$$
$$\left. \text{for some } j \in J_{n1}\right\}$$

$$
\begin{aligned}
&+P\left\{\sigma^2|\mathbf{e}_j^{\mathrm{T}}\Sigma_{n1}^{-1}\widetilde{\mathbf{s}}_{n1}|\frac{\lambda_n}{n}\geq\frac{|\beta_{0j}|}{2}\ \text{for some } j\in J_{n1}\right\}\\
&+P\left\{|\mathbf{X}_{\cdot j}^{\mathrm{T}}\mathbf{H}_n\boldsymbol{\varepsilon}^*|\geq(1-\kappa-\epsilon)\sigma^2\lambda_n w_{nj}\right.\\
&\qquad\left.\text{for some } j\notin J_{n1}\right\}\\
&+P\left\{\frac{1}{n}|\mathbf{X}_{\cdot j}^{\mathrm{T}}\mathbb{X}^1\Sigma_{n1}^{-1}\widetilde{\mathbf{s}}_{n1}|\geq(\kappa+\epsilon)w_{nj}\right.\\
&\qquad\left.\text{for some } j\notin J_{n1}\right\}\\
&=P\{B_{n1}\}+P\{B_{n2}\}+P\{B_{n3}\}+P\{B_{n4}\},
\end{aligned}
$$

where $\mathbf{e}_j$ is the unit vector in the direction of the j^{th} coordinate.

Since $||(\mathbf{e}_j^{\mathrm{T}}\Sigma_{n1}^{-1}(\mathbb{X}^1)^{\mathrm{T}})^{\mathrm{T}}||_2/n\leq(n\tau_{n1})^{-1/2}$ and $|\beta_{0j}|\geq b_{n1}$ for $j\in J_{n1}$,

$$
P\{B_{n1}\}=P\left\{\frac{1}{n}|\mathbf{e}_j^{\mathrm{T}}\Sigma_{n1}^{-1}\left(\mathbb{X}^1\right)^{\mathrm{T}}\boldsymbol{\varepsilon}^*|\geq\frac{|\beta_{0j}|}{2},\right.
\left.\text{for some } j\in J_{n1}\right\}\leq k_n q_n^*(\frac{\sqrt{\tau_{n1}n}b_{n1}}{2})
$$

with the tail probability $q_n^*(t)$ defined in Lemma A.1. Thus, $P\{B_{n1}\}\to 0$ as $n\to\infty$, by Lemmas A.1 and A.3, Conditions (A4) and (A5).

For $P\{B_{n2}\}$, by Lemma A.2 and Conditions (A4) and (A5)

$$
\begin{aligned}
\sigma^2|\mathbf{e}_j^{\mathrm{T}}\Sigma_{n1}^{-1}\widetilde{\mathbf{s}}_{n1}|\frac{\lambda_n}{n}&\leq\frac{\sigma^2||\widetilde{\mathbf{s}}_{n1}||\lambda_n}{\tau_{n1}n}\\
&=O_P(\sigma^2\frac{M_{n1}\lambda_n}{\tau_{n1}n})=o_P(b_{n1}),
\end{aligned}
$$

where $b_{n1}=\min\{|\beta_{0j}|,j\in J_{n1}\}$. Therefore, we have $P\{B_{n2}\}\to 0$ as $n\to\infty$.

Since $w_{nj}^{-1}=|\widetilde{\beta}_{nj}|\leq M_{n2}+O_P(1/r_n)$ and $||(\mathbf{X}_{\cdot j}^{\mathrm{T}}\mathbf{H}_n)^{\mathrm{T}}||_2\leq\sqrt{n}$, for large C,

$$
\begin{aligned}
P\{B_{n3}\}&\leq P\left\{|\mathbf{X}_{\cdot j}^{\mathrm{T}}\mathbf{H}_n\boldsymbol{\varepsilon}^*|\geq\frac{(1-\kappa-\epsilon)\lambda_n}{C(M_{n2}+\frac{1}{r_n})}\right.\\
&\qquad\left.\text{for some } j\notin J_{n1}\right\}+o(1)\\
&\leq m_n q_n^*\left\{\frac{(1-\kappa-\epsilon)\lambda_n}{C(M_{n2}+\frac{1}{r_n})\sqrt{n}}\right\}.
\end{aligned}
$$

Thus by Lemmas A.1 and A.3, and Condition (A4), $P\{B_{n3}\}\to 0$ as $n\to\infty$.

Finally for $P\{B_{n4}\}$, it comes from Lemma A.2 and Condition (A5) that

$$
\begin{aligned}
&\max_{j\notin J_{n1}}\left(\frac{|\mathbf{X}_{\cdot j}^{\mathrm{T}}\mathbb{X}^1\Sigma_{n1}^{-1}\widetilde{\mathbf{s}}_{n1}|}{nw_{nj}}-\frac{\left|\eta_{nj}\mathbf{X}_{\cdot j}^{\mathrm{T}}\mathbb{X}^1\Sigma_{n1}^{-1}\boldsymbol{s}_{n1}\right|}{n}\right)\\
&\leq\max_{j\notin J_{n1}}\left(\frac{\left\|\left(\mathbf{X}_{\cdot j}^{\mathrm{T}}\mathbb{X}^1\Sigma_{n1}^{-1}\right)^{\mathrm{T}}\right\|}{n}\right)\left\||\tilde{\beta}_{nj}|\widetilde{\boldsymbol{s}}_{n1}-|\eta_{nj}|\boldsymbol{s}_{n1}\right\|\\
&\leq\tau_{n1}^{-1/2}o_p(1)=o_p(1).
\end{aligned}
$$

By Condition (A3), we have $|\eta_{nj}\mathbf{X}_{\cdot j}^{\mathrm{T}}\mathbb{X}^1\Sigma_{n1}^{-1}\boldsymbol{s}_{n1}|/n\leq\kappa$. So $P\{B_{n4}\}\to 0$ as $n\to\infty$. □

Proof of Theorem A.2. By (A.12), we have

$$
\begin{aligned}
n^{1/2}\boldsymbol{\alpha}_n^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{01})=&\,n^{-1/2}\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}\left(\mathbb{X}^1\right)^{\mathrm{T}}\boldsymbol{\varepsilon}^*\\
&-n^{-1/2}\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}\sigma^2\lambda_n\widetilde{\boldsymbol{s}}_{n1},
\end{aligned}
$$

where $\boldsymbol{\varepsilon}^*=(\varepsilon_1^*,\ldots,\varepsilon_n^*)$ with $\varepsilon_i^*=Y_i^*-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0$. When $||\boldsymbol{\alpha}_n||_2\leq 1$,

$$
|n^{-1/2}\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}\sigma^2\lambda_n\widetilde{\boldsymbol{s}}_{n1}|\leq 2n^{-1/2}\tau_{n1}^{-1}M_{n1}\sigma^2\lambda_n.
$$

Therefore, by $M_{n1}\lambda_n/\sqrt{n}\to 0$, we have $n^{1/2}s_n^{-1}\boldsymbol{\alpha}_n^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{01})=n^{-1/2}s_n^{-1}\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}(\mathbb{X}^1)^{\mathrm{T}}\boldsymbol{\varepsilon}^*+o_p(1)=n^{-1/2}s_n^{-1}\sum_{i=1}^n\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}\mathbf{X}_i^1\varepsilon_i^*+o_p(1)$, where $(\mathbf{X}_i^1)^{\mathrm{T}}$ is the i^{th} row of $\mathbb{X}^1$ and ε^* is the i^{th} component of $\boldsymbol{\varepsilon}^*$, which is as discussed in Lemma A.3. It suffices to prove Theorem A.2 by verifying the conditions of the Lindeberg-Feller central limit theorem.

Let $v_i=n^{-1/2}s_n^{-1}\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}\mathbf{X}_i^1$ and $w_i=v_i\varepsilon_i^*$. Note that $\mathrm{E}(\varepsilon_i^*)=0$ and $\mathrm{Var}(\varepsilon_i^*)=\sigma^2\{\Phi_i-Z_i\phi_i+\phi_i^2/(1-\Phi_i)\}=\sigma_{1i}^2$. Then we have

$$
\begin{aligned}
\mathrm{var}(\sum_{i=1}^n w_i)=n^{-1}s_n^{-1}\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}\left\{\sum_{i=1}^n\mathbf{X}_i^1\left(\mathbf{X}_i^1\right)^{\mathrm{T}}\sigma_{1i}^2\right\}\\
\Sigma_{n1}^{-1}\boldsymbol{\alpha}_n s_n^{-1}=1.
\end{aligned}
$$

For any $\epsilon>0$, $\sum_{i=1}^n E[w_i^2I\{|w_i|>\epsilon\}]=\sum_{i=1}^n v_i^2E[(\varepsilon_i^*)^2I\{|\varepsilon_i^*v_i|>\epsilon\}]$. Since $\sum_{i=1}^n v_i^2\sigma_{1i}^2=1$, it suffices to show that, $\max_{1\leq i\leq n}E(\varepsilon_i^*)^2I\{|\varepsilon_i^*v_i|>\epsilon\}\to 0$ or equivalently,

$$
\max_{1\leq i\leq n}|v_i|=n^{-1/2}s_n^{-1}\max_{1\leq i\leq n}|\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}\mathbf{X}_i^1|\to 0. \tag{A.15}
$$

Write $\mathbf{A}=n^{-1}\{\sum_{i=1}^n\mathbf{X}_i^1(\mathbf{X}_i^1)^{\mathrm{T}}\sigma_{1i}^2\}$. Then

$$
\begin{aligned}
|\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}\mathbf{X}_i^1|&=|\boldsymbol{\alpha}_n^{\mathrm{T}}\left(\Sigma_{n1}^{-1}\mathbf{A}\Sigma_{n1}^{-1}\right)\left(\Sigma_{n1}\mathbf{A}^{-1}\mathbf{X}_i^1\right)|\\
&\leq\left(\boldsymbol{\alpha}_n^{\mathrm{T}}\Sigma_{n1}^{-1}\mathbf{A}\Sigma_{n1}^{-1}\boldsymbol{\alpha}_n\right)^{1/2}\left(\left(\mathbf{X}_i^1\right)^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{X}_i^1\right)^{1/2}\\
&=s_n\left(\left(\mathbf{X}_i^1\right)^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{X}_i^1\right)^{1/2}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\max_{1\leq i\leq n}|v_i|&\leq n^{-1/2}\max_{1\leq i\leq n}\left(\left(\mathbf{X}_i^1\right)^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{X}_i^1\right)^{1/2}\\
&\leq\tau_2^{-1/2}n^{-1/2}\max_{1\leq i\leq n}\left(\left(\mathbf{X}_i^1\right)^{\mathrm{T}}\mathbf{X}_i^1\right)^{1/2}\to 0
\end{aligned}
$$

due to Conditions (A1) and (A5). This completes the proof of Theorem A.2. □

A.5.5 Consistency of the marginal regression estimators

Without loss of generality, let $\mathbf{Y}_{n_1} = (Y_1, \ldots, Y_{n_1})$ be the collection of the response values for the un-censored observations. Correspondingly, let $\mathbb{X}_{n_1} = (\mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_{n_1}^{\mathrm{T}})^{\mathrm{T}}$ be the matrix containing the first n_1 rows of $\mathbb{X}$ and $\mathbf{X}_{n_1 \cdot j}$ be the j^{th} column of $\mathbb{X}_{n_1}$ for $1 \le j \le p_n$. The initial estimator of β_j is given as

$$\widetilde{\beta}_{nj} = (\mathbf{X}_{n_1 \cdot j}^{\mathrm{T}} \mathbf{X}_{n_1 \cdot j})^{-1} \mathbf{X}_{n_1 \cdot j}^{\mathrm{T}} \mathbf{Y}_{n_1}. \tag{A.16}$$

According to Amemiya [36], let U_i^* be the random variable with the density $h(u)$ given by

$$h(u) = \frac{1}{\Phi_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(u/\sigma)^2/2\}, \ -\mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0 < u < \infty.$$

Then we have

$$Y_i = \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + U_i^*, \ \forall Y_i > 0, \tag{A.17}$$

with $E(U_i^*) = \sigma\phi_i/\Phi_i$. By (A.16) and (A.17) we obtain

$$\widetilde{\beta}_{nj} = (\mathbf{X}_{n_1 \cdot j}^{\mathrm{T}} \mathbf{X}_{n_1 \cdot j})^{-1} \mathbf{X}_{n_1 \cdot j}^{\mathrm{T}} \mathbf{Y}_{n_1} = \frac{\mathbf{X}_{n_1, j}^{\mathrm{T}} (\mathbb{X}_{n_1} \boldsymbol{\beta}_0 + \mathbf{U}^*)}{\sum_{i=1}^{n_1} X_{ij}^2},$$

where $\mathbf{U}^* = (U_1^*, \ldots, U_{n_1}^*)^{\mathrm{T}}$. Define $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{n_1})^{\mathrm{T}}$ with $\gamma_i = \phi_i/\Phi_i$. Take η_{nj} given in Condition (A2) as $E(\widetilde{\beta}_{nj}) = \frac{\mathbf{X}_{n_1, j}^{\mathrm{T}} (\mathbb{X}_{n_1} \boldsymbol{\beta}_0 + \sigma\boldsymbol{\gamma})}{\sum_{i=1}^{n_1} X_{ij}^2}$, and consider the following conditions:

(B1) (Partial Orthogonality) The covariates with zero coefficients and those with nonzero coefficients are weakly correlated such that

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_{ij} X_{ik} \right| = \left| \frac{\mathbf{X}_{\cdot j}^{\mathrm{T}} \mathbf{X}_{\cdot k}}{n} \right| \le \rho_n, \ j \notin J_{n1}, \ k \in J_{n1},$$

where ρ_n satisfies that there is a constant $0 < \kappa < 1$, such that

$$c_n = (\max_{j \notin J_{n1}} |\eta_{nj}|)(\sum_{j \in J_{n1}} \frac{|\eta_{nj}|^{-2}}{k_n})^{1/2} \le \frac{\kappa\tau_{n1}}{k_n \rho_n}, \tag{A.18}$$

with κ given in Condition (A3).

(B2) The minimum $\widetilde{b}_{n1} = \min\{|\eta_{nj}|, j \in J_{n1}\}$ satisfies

$$\frac{k_n^{1/2}(1 + c_n)}{\widetilde{b}_{n1} r_n} \to 0, \ r_n = \frac{\sqrt{n}}{\{\log(p_n - k_n)\}^{1/2}}.$$

(B3) There exists a constant $0 < a < 1$, such that $a \le \frac{1}{n} \sum_{i=1}^{n_1} X_{ij}^2 \le 1$.

Condition (B1) indicates that the covariates with zero and nonzero coefficients are weakly correlated. Condition (B2) implies that the nonzero coefficients are bounded away from zero at certain rates which depend on the growth of k_n and $(p_n - k_n)$ and is the special case of Condition (B3) in Huang, Ma and Zhang [12] with $d = 2$. Condition (B3) requires that square sum of the j^{th} covariate for uncensored part is bounded away from zero. Notice that $\sum_{i=1}^{n} X_{ij}^2 = n$. This assumption generally holds as long as the censoring rate is not very large.

Theorem A.3. *Suppose that conditions (B1)–(B3) hold. Then the initial estimator $\widetilde{\beta}_{nj}$ in (A.16) satisfies r_n-consistency for the estimation of η_{nj} such that $r_n \max_{1 \le j \le p_n} |\widetilde{\beta}_{nj} - \eta_{nj}| = O_P(1)$, as $r_n \to \infty$, and η_{nj} satisfies Condition (A2) and the adaptive irrpresentable condition in (A3).*

Proof of Theorem A.3. For all $\epsilon > 0$,

$$P\{r_n \max_{1 \le j \le p_n} |\widetilde{\beta}_{nj} - \eta_{nj}| > \epsilon\} = P\left\{ r_n \max_{1 \le j \le p_n} \frac{|\mathbf{X}_{n_1, j}^{\mathrm{T}} (\mathbf{U}^* - \sigma\boldsymbol{\gamma})|}{\sum_{i=1}^{n_1} X_{ij}^2} > \epsilon \right\}.$$

It is easy to verify that $E(U_i^* - \sigma\gamma_i) = 0$ and $Var(U_i^* - \sigma\gamma_i) = \sigma^2(1 - Z_i\phi_i/\Phi_i - \phi_i^2/\Phi_i^2) \le \sigma^2$. Also it is ready to show that there exists constants C and K such that $P(|U_i^* - \sigma\gamma_i| > t) \le K\exp(-Ct^2), \forall t > 0$. So by Lemma A.1 and Condition (B3), we have

$$P\{r_n \max_{1 \le j \le p_n} |\widetilde{\beta}_{nj} - \eta_{nj}| > \epsilon\} \le p_n q_n^* (\frac{\sqrt{an}\epsilon}{r_n}) = o(1).$$

As to the second part of Condition (A2) with $M_{n2} = \max_{j \notin J_{n1}} |\eta_{nj}|$, it comes from Condition (B2) that

$$\sum_{j \in J_{n1}} \left(\frac{1}{\eta_{nj}^2} + \frac{M_{n2}^2}{\eta_{nj}^4} \right) \le \frac{k_n}{\widetilde{b}_{n1}^2} (1 + c_n^2) = o(r_n^2).$$

For Condition (A3), with the facts that

$$\left\| (\mathbb{X}_1)^{\mathrm{T}} \mathbf{X}_{\cdot j} \right\|^2 \le k_n n^2 \rho_n^2$$

and $|\eta_{nj}| \times ||\boldsymbol{s}_{n1}|| \le k_n^{1/2} c_n$ for all $j \notin J_{n1}$, it comes from Condition (B1) that

$$|\eta_{nj}| n^{-1} |\mathbf{X}_{\cdot j}^{\mathrm{T}} \mathbb{X}^1 \Sigma_{n1}^{-1} \boldsymbol{s}_{n1}| \le \frac{c_n k_n \rho_n}{\tau_{n1}} \le \kappa, \ \forall j \notin J_{n1}.$$

The proof is completed. □

Received 29 August 2014

REFERENCES

[1] Mascola, J. R. and Haynes, B. F. Hiv-1 neutralizing antibodies: understanding nature's pathways. *Immunological Reviews* 2013; **254**(1):225–244.

[2] Wei, C., Jung, J., and Sanz, I. OMIP-003: phenotypic analysis of human memory B cells. *Cytometry. Part A* 2011; **79**(11):894–896.

[3] Sanz, I., Wei, C., Lee, F., and Anolik, J. Phenotypic and functional heterogeneity of human memory B cells. *Seminars in Immunology* 2008; **20**(1):67–82.
[4] Kobie, J. J., Alcena, D. C., Zheng, B., Bryk, P., Mattiaccio, J. L., Brewer, M., LaBranche, C., Young, F. M., Dewhurst, S., and Montefiori, D. C., et al. 9G4 autoreactivity is increased in HIV-infected patients and correlates with HIV broadly neutralizing serum activity. *PLoS One* 2012; **7**(4):e35356.
[5] Schiller, J. and Chackerian, B. Why HIV virions have low numbers of envelope spikes: implications for vaccine development. *PLoS Pathog* 2014; **10**(8):e1004254.
[6] Frank, I. and Friedman, J. A statistical view of some chemometrics regression tools (with discussion). *Technometrics* 1993; **35**:109–148.
[7] Breiman, L. Heuristics of instability and stabilization in model selection. *The Annals of Statistics* 1996; **24**:2350–2383. MR1425957
[8] Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 2005; **67**:301–320. MR2137327
[9] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 1996; **58**:267–288. MR1379242
[10] Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 2006; **101**:1418–1429. MR2279469
[11] Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 2001; **96**:1348–1360. MR1946581
[12] Huang, J., Ma, S., and Zhang, C. H. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* 2008; **18**:1603–1618. MR2469326
[13] Tibshirani, R., et al. The lasso method for variable selection in the Cox model. *Statistics in Medicine* 1997; **16**(4):385–395.
[14] Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. High-dimensional variable selection for survival data. *Journal of the American Statistical Association* 2010; **105**(489):205–217. MR2757200
[15] Liu, X. and Zeng, D. Variable selection in semiparametric transformation models for right-censored data. *Biometrika* 2013; **100**(4):859–876. MR3142337
[16] Zhou, Z., Jiang, R., and Qian, W. LAD variable selection for linear models with randomly censored data. *Metrika* 2013; **76**(2):287–300. MR3018834
[17] Liu, X., Wang, Z., and Wu, Y. Group variable selection and estimation in the tobit censored response model. *Computational Statistics & Data Analysis* 2013; **60**:80–89. MR3007020
[18] Tobin, J. Estimation of relationships for limited dependent variables. *Econometrica* 1958; **26**:24–36. MR0090462
[19] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics* 1978; **6**:461–464. MR0468014
[20] Broman, K. and Speed, T. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; **64**(4):641–656. MR1979381
[21] Siegmund, D. Model selection in irregular problems: applications to mapping quantitative trait loci. *Biometrika* 2004; **91**(4):785–800. MR2126033
[22] Bogdan, M., Ghosh, J., and Doerge, R. Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* 2004; **167**(2):989–999.
[23] Chen, J. and Chen, Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 2008; **95**(3):759–771. MR2443189
[24] Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. Pathwise coordinate optimization. *The Annals of Applied Statistics* 2007; **1**:302–332. MR2415737
[25] Delves, P., Martin, S., Burton, D., and Roitt, I. *Roitt's Essential Immunology (Essentials)*. 11th edn. Wiley-Blackwell, 2006.
[26] Wei, C., Anolik, J., Cappione, A., Zheng, B., Pugh-Bernard, A., Brooks, J., Lee, E. H., Milner, E. C., and Sanz, I. A new population of cells lacking expression of CD27 represents a notable component of the B cell memory compartment in systemic lupus erythematosus. *The Journal of Immunology* 2007; **178**(10):6624–6633.
[27] Gray, E. S., Taylor, N., Wycuff, D., Moore, P. L., Tomaras, G. D., Wibmer, C. K., Puren, A., DeCamp, A., Gilbert, P. B., and Wood, B., et al. Antibody specificities associated with neutralization breadth in plasma from human immunodeficiency virus type 1 subtype C-infected blood donors. *Journal of Virology* 2009; **83**(17):8925–8937.
[28] Petrovas, C., Vlachoyiannopoulos, P., Kordossis, T., and Moutsopoulos, H. Anti-phospholipid antibodies in HIV infection and SLE with or without anti-phospholipid syndrome: comparisons of phospholipid specificity, avidity and reactivity with beta2-GPI. *Journal of Autoimmunity* 1999; **13**(3):347–355.
[29] Haynes, B. F., Fleming, J., Clair, E. W. S., Katinger, H., Stiegler, G., Kunert, R., Robinson, J., Scearce, R. M., Plonk, K., and Staats, H. F., et al. Cardiolipin polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies. *Science* 2005; **308**(5730):1906–1908.
[30] Scheid, J. F., Mouquet, H., Ueberheide, B., Diskin, R., Klein, F., Oliveira, T. Y., Pietzsch, J., Fenyo, D., Abadir, A., and Velinzon, K., et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* 2011; **333**(6049):1633–1637.
[31] Huang, J., Horowitz, J. L., and Ma, S. G. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics* 2008; **36**:587–613. MR2396808
[32] Batini, C. and Scannapieco, M. *Data Quality: Concepts, Methodologies and Techniques*. Springer Science & Business Media, 2006.
[33] Veall, M. R. and Zimmermann, K. F. Goodness of fit measures in the tobit model. *Oxford Bulletin of Economics and Statistics* 1994; **56**(4):485–99.
[34] Zhao, P. and Yu, B. On model selection consistency of lasso. *Journal of Machine Learning Research* 2007; **7**(2):2541. MR2274449
[35] van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, Springer-Verlag, New York, 1996. MR1385671
[36] Amemiya, T. Regression analysis when the dependent variable is truncated normal. *Econometrica* 1973; **41**:997–1016. MR0440773
[37] James, I. and Smith, P. Consistency results for linear regression with censored data. *The Annals of Statistics* 1984; **12**(2):590–600. MR0740913

Tian Chen
Department of Biostatistics and Computational Biology
University of Rochester Medical Center
Rochester, NY 14642
USA
E-mail address: Tian_Chen@urmc.rochester.edu

Shujie Ma
Department of Statistics
University of California at Riverside
Riverside, CA 92521
USA
E-mail address: shujie.ma@ucr.edu

James Kobie
Division of Infectious Diseases
Department of Medicine
University of Rochester Medical Center
Rochester, NY 14642
USA
E-mail address: James_Kobie@urmc.rochester.edu

Alexander Rosenberg
Division of Allergy, Immunology, and Rheumatology
Department of Medicine
University of Rochester Medical Center
Rochester, NY 14642
USA
E-mail address: Alex_Rosenberg@URMC.Rochester.edu

Ignacio Sanz
Division of Rheumatology and
Lowance Center for Human Immunology
Emory University
Atlanta, GA 30322
USA
E-mail address: gnacio.sanz@emory.edu

Hua Liang
Department of Statistics
George Washington University
Washington, DC 20052
USA
E-mail address: hliang@gwu.edu