ABSTRACT

SHOWS, JUSTIN HALL. Sparse Estimation and Inference for Censored Median Regression. (Under the direction of Drs. Wenbin Lu and Hao Helen Zhang).

Censored median regression models have been shown to be useful for analyzing a variety of censored survival data with the robustness property. We study sparse estimation and inference of censored median regression. The new method minimizes an inverse censoring probability weighted least absolute deviation subject to the adaptive LASSO penalty. We show that, with a proper choice of the tuning parameter, the proposed estimator has nice theoretical properties such as root-n consistency and asymptotic normality. The estimator can also identify the underlying sparse model consistently. We propose using a resampling method to estimate the variance of the proposed estimator. Furthermore, the new procedure enjoys great advantages in computation, since its entire solution path can be obtained efficiently. Also, the method can be extended to multivariate survival data, where there is a natural or artificial clustering structure. The performance of our estimator is evaluated by extensive simulations and two real data applications.

Sparse Estimation and Inference for Censored Median Regression

by
Justin Hall Shows

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fullfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2009

Approved By:

Dr. Dennis Boos	Dr. Daowen Zhang
Dr. Wenbin Lu Chair of Advisory Committee	Dr. Hao Helen Zhang Co-Chair of Advisory Committee

DEDICATION

This is dedicated to everyone who has helped me (named in the acknowledgement section).

BIOGRAPHY

I was born January 20, 1980 at Jones County Commumity Hospital in Laurel, Mississippi to Thomas and Sherry Shows. I received a B.S. in Mathematics with Secondary Licensure at the University of Southern Mississppi and an M.S. in Statistics from Mississippi State University before coming to North Carolina State University to pursue a PhD in Statistics.

ACKNOWLEDGMENTS

I first have to thank Dr. Wenbin Lu and Dr. Helen Zhang who have helped me more than words can say. I would also like to thank Dr. Dennis Boos and Dr. Daowen Zhang for their suggestions during my oral exams. I would also like to thank the following people: Dr. Pam Arroway, Dr. Dongfeng Wu, Dr. Jane Harvill, Thomas and Sherry Shows, Anastasia Wong, and Dr. Thomas Spruill.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Review on survival models and methods

Survival or time-to-event data analysis is commonly found in various fields of study, including biomedicine, epidemiology, social science, economics, and engineering. The main goal in this type of analysis is to study the distribution of the time until a certain event of interest, usually called a "failure", occurs. In actual data collection, however, it is common that the event times of interest are censored for various reasons. For a more detailed discussion of censoring, see Kalbfleisch and Prentice (2002) and Klein and Moeschberger (2003). Due to censoring, only partial information of event times on some subjects may be observed, which makes analyzing survival data additonally challenging. In this thesis, we will focus on right censoring, in which something happens prior to the failure time that prevents it from being observed directly. For example, if we are studying the time to death of patients with terminal lung cancer, there may be many events which may preclude the observation of the time of death. An individual may die from another cause or drop out of the study, or commonly, the study period may end before some individuals die. We will assume that for every individual, there is a failure time T and a censoring time C. Usually, it is assumed that the failure time is independent of the censoring time, or at least that they are independent given certain covariates. While we are interested in studying the properties and distribution of T, we only observe $\tilde{T} = \min(T, C)$ and $\delta = I(T \leq C)$. So for a study consisting of n individuals from a homogeneous population, we observe $(\tilde{T}_i, \delta_i)$, $i = 1, \ldots, n$. Based on the observed data, we need to deduce the distribution of T.

The distribution of T can be described by the survival function, which is left-continuous and given by

$$S(t) = P(T \geq t) \quad \text{for} \quad t \geq 0$$

where $S(0) = 1$ and $S(\infty) = 0$. The distribution can also be described by the hazard rate, or instantaneous rate of failure, defined as

$$\lambda(t) = \lim_{h \to 0} h^{-1} P(t \leq T < t + h | T \geq t).$$

The cumulative hazard function is then $\Lambda(t) = \int_0^t \lambda(s) ds$. Any of these may describe the distribution of T, since $S(t) = \exp\{-\Lambda(t)\}$. There are numerous methods of estimating the distribution of T nonparametrically using the observed data. We will make extensive use of the Kaplan-Meier estimator (Kaplan and Meier, 1958) of the survival function, which is given by

$$\hat{S}(t) = \prod_{u \leq t} \{1 - \frac{J(u)}{Y(u)} dN(u)\},$$

where $Y(u) = \sum_{i=1}^n I(\tilde{T}_i \geq u)$ is the total number of subjects at risk at time u, $N(u) = \sum_{i=1}^n I(\tilde{T}_i \leq u, \delta_i = 1)$ is the total number of failures observed by time u, and $J(u) = I(Y(u) > 0)$. In this context, any value of $\frac{0}{0}$ is taken to be 0.

In many practical time to event studies, besides the event time information, a large number of predictors will also be collected. A major objective of these studies is to investigate the relationship between the event time of interest and the predictors. In the last forty years, a variety of semiparametric survival models has been proposed and actively used in literature. One of the most popular models in survival analysis is Cox's proportional hazards (PH) model (Cox, 1972). The hazard function is given by

$$\lambda(t|X_i) = \lambda_0(t) \exp(\beta_0' X_i),$$

where $\lambda_0(t)$ is a baseline hazard function, X_i is a vector of covariates for the ith subject, and β_0 the true value of a vector of unknown regression coefficients. One of the major advantages of this model is that a partial likelihood is available (Cox, 1972), given by

$$L(\beta) = \prod_{i=1}^n \left(\frac{e^{\beta' X_i}}{\sum_{j \in R_i} e^{\beta' X_j}} \right)^{\delta_i},$$

where $R_i = \{j : \tilde{T}_j \geq \tilde{T}_i\}$. The maximum partial likelihood estimator of β can then be easily obtained from $L(\beta)$ (Cox, 1975). Moreover, it has been shown that the resulting estimator is consistent, asymptotically normal, and semi-parametrically efficient (Tsiatis, 1981; Andersen and Gill, 1982).

In many practical applications, however, the assumption of proportional hazards is violated. For example, if X_i is simply a treatment indicator

$$X_i = \begin{cases} 1 & \text{if subject is in the treatment group} \\ 0 & \text{if subject is in the control group} \end{cases}$$

then the proportional hazards assumption requires that the ratio of the hazard function of the treatment group and that of the control group is constant at all times, i.e., the hazard functions are parallel at all times. If the hazard functions converge to the same limit, then the proportional hazards assumption is violated. Under such situations alternative models may be more suitable.

Another popular model is the proportional odds (PO) model (Bennett, 1983a; Bennett, 1983b; Pettitt, 1984). This model assumes that for $S(t|X_i)$, the conditional survival function for the ith individual, we have

$$\frac{1 - S(t|X_i)}{S(t|X_i)} = \frac{1 - S_0(t)}{S_0(t)} \exp(\beta_0' X_i),$$

where $S_0(t)$ is a completely unspecified baseline survival function, β_0 is the true value of a vector of regression parameters, and X_i is a vector of covariates for the ith individual. Under this model, it is not necessary for the hazard function to be monotonic, so that it may be appropriate when the hazard function reaches a peak at some point in time and then declines (Bennett, 1983b). However, for the PO model, the partial likelihood is not available. Therefore, the estimation of parameters, including regression coefficients, becomes more challenging. Various methods have been proposed for the estimation of the PO model. For example, Pettitt (1984), among others, shows that the estimates for the regression parameters can be computed by replacing the survival times with their ranks. In addition, the estimation and inference procedures can be based on a marginal likelihood (Lam and Leung, 2001), which is also a rank-based method. Recently, the semiparametric maximum likelihood estimation method has also been studied for the PO model by various authors (Scharfstein et al, 1998; Zeng and Lin, 2007). Scharfstein et al (1998) construct

an estimator for β that is consistent and asymptotically normal; further, the estimator is semiparametric efficient in that it attains the semiparametric variance bound. Zeng and Lin (2007) develop a unified semi-parametric maximum likelihood estimation method for a general class of linear transformation models, of which the PO model is a special case. While the PO model may be more appropriate than the PH model in some situations, as with the proportional hazards assumption, the assumption of proportional odds may not be met in some applications.

Another attractive alternative is the accelerated failure time (AFT) model (Kalbfleisch and Prentice, 1980; Cox and Oakes, 1984), which relates the logarithm of the failure time linearly to covariates as

$$\log(T) = \beta' Z + \epsilon, \tag{1.1}$$

where T is the failure time of our interest, Z is the p-dimensional vector of covariates and ϵ is the error term with a completely unspecified distribution that is independent of Z. Due to censoring, we only observe $\tilde{T} = \min(T, C)$ and $\delta = I(T \leq C)$, where C is the censoring time that is assumed to be independent of T conditional on Z. Compared to proportional hazards models, the estimates from AFT models are robust to the presence of unmeasured covariates, since they are less affected by the choice of probability distribution. Furthermore, the results of AFT models are easier to interpret (Reid, 1994). Due to these desired properties, the AFT model has been extensively studied in the literature (Prentice, 1978; Buckley and James, 1979; Ritov, 1990; Tsiatis, 1990; Ying, 1993; Jin et al., 2003; among others). Buckley and James (1979) modify the normal equations to accommodate censored observations in order to estimate the regression parameters, while Prentice (1978) uses generalized ranks for inference about β. The large sample properties of rank estimators have been rigously studied by Tsiatis (1990) and Ying (1993). Jin et al (2003) show that the estimators can be obtained via linear programming, and that they are consistent and asymptotically normal. They further use a resampling technique to estimate the limiting covariance matrix that does not involve nonparametric density estimation.

A major assumption of the AFT model is that the error terms are i.i.d., which is often too restrictive in practice (Koenker and Geling, 2001). To handle more complicated problems, say, data with heteroscedastic or heavy-tailed error distributions, censored quantile regression provides a natural remedy. In the context of survival data analysis,

Ying et al. (1995), Bang and Tsiatis (2003), and Zhou (2006) consider median regression with random censoring and derive various inverse censoring probability weighted methods for parameter estimation; Portnoy (2003) considers general censored quantile regression by assuming that all conditional quantiles are linear functions of the covariates and develops recursively re-weighted estimators of regression parameters. In general, quantile regression is more robust against outliers and may be more effective; the regression quantiles may be well estimated even if the mean is not.

1.2 Review on variable selection methods

In studying the distribution of the failure time, it is common to collect auxilary information in the form of measured covariates. One interesting problem is to study the distribution of the failure time conditional on these covariates. With modern technology, it is possible to collect information on a large number of potential predictors. For example, more and more medical studies tend to measure patients' genomic information and combine it with traditional risk factors to improve disease diagnosis and make personalized drugs. Genomic information, often in the form of gene expression patterns, is high dimensional and typically only a small number of genes contains relevant information, so the underlying model is naturally sparse. As in standard regression problems, we are often interested in investigating the relationships between the response (survival time in our case) and a large number of covariates. We would then select those covariates that are relevant to predicting the survival distribution and estimate their effects on the survival time. For some of the covariates, the true value of the corresponding coefficients may be 0, i.e., they do not affect survival times or prediction of a survival outcome. Simultaneous variable selection and parameter estimation then becomes important since it can produce a parsimonious model with better risk assessment and model interpretation. However, how to identify the subset of relevant predictors with censored data is an important yet challenging question.

1.2.1 Variable selection for uncensored data

Before we discuss variable selection for censored data, we will review some variable selection methods in the context of uncensored data. Suppose that we have iid observations

data $(X_i, Y_i)_{i=1}^n$ so that

$$Y_i = \theta' X_i + \epsilon_i \quad \text{for } i = 1, 2, \ldots, n,$$

where $X_i = (1, X_{i1}, \ldots, X_{ip})'$; $\theta = (\alpha, \beta_1, \ldots, \beta_p)'$, where $\theta_0 = (\alpha_0, \beta_{10}, \ldots, \beta_{p0})'$ is the true value of θ; ϵ_i iid with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. In standard least squares regression, some of the most popular methods of variable selection are forward selection, backward elimination, and stepwise regression (Montgomery and Peck, 1991; Hocking, 1976). In these methods, variables are added or deleted from the model one at a time. In forward selection, the model begins with only an intercept. The first variable added to the model is the one with the highest simple correlation with the response variable. At each subsequent step, the variable is added that causes the largest decrease in the sum of squares error (SSE). At each step, the partial sums of squares are used to test if the entering variable meets the prespecified entry significance level α_{ENTER}. If not, then the process is stopped. In backward elimination, the beginning model includes all possible variables. At each step, the variable whose deletion causes the smallest decrease in SSE is eliminated. At each deletion, the partial sums of squares are used to test if the variables meets the prespecified significance level α_{STAY} to stay in the model. If so, then the process stops. Stepwise regression starts with one variable and adds variables one by one to the model as in forward selection. At each step, all the variables in the model are tested to see if they meet the α_{STAY} criteria to stay in the model. The process is stopped when all the variables in the model meet the α_{STAY} criterion to stay in the model, and the variables not in the model meet the α_{ENTER} criterion. While these methods can be easily carried out in SAS, many authors note that they are not guaranteed to yield the best model in any sense (Montgomery and Peck, 1991; Hocking, 1976). There are many criteria available that allow for best subset selection. Some of them include the adjusted correlation coefficient, Mallow's C_p statistic (Mallows, 1973), Akaike information criteria (AIC) (Akaike, 1973), Schwarz-Bayesian information criteria (BIC) (Schwarz, 1978), and various cross-validation criteria.

Luo et al (2006) note that these methods do not adapt the tuning parameters of the criteria based on the data. They propose to tune the parameters by adding controlled amounts of random noise to the response variable and performing a variable selection method with the noise-inflated data. The method discussed is forward selection, but can be extended to other methods as well. The noise addition model selection (NAMS) method is based on the fact that underfitting and overfitting the model results in biased estimates

of the error variance (Luo et al, 2006). In the case of forward selection, if α_{ENTER} is too large, then the selected model will be overfit and the error variance will be underestimated; conversely, if α_{ENTER} is too small, the selected model will be underfit and the error variance will tend to be overestimated. The noise-inflated responses are given by

$$Y_i^* = Y_i + \tau\sqrt{\lambda}W_i, \quad i = 1, \ldots, n,$$

where τ and λ are known and $(W_i)_{i=1}^n$ are randomly generated i.i.d. standard normal variables. (Luo et al (2006) recommend using $\tau = \sqrt{\hat{\sigma}_F^2}$, where $\hat{\sigma}_F^2$ is the estimate of σ^2 from a fit of the full model.) Then Y_i^* has the same conditional expectation of Y_i, but the error variance is $Var(Y_i^*|X_i) = \sigma^2 + \tau^2\lambda$. Forward selection is then applied to the noise-inflated data $(Y_i^*, X_i)_{i=1}^n$. If α_{ENTER} is too large, then expexted mean square error (MSE) of the selected model is less than $\sigma^2 + \tau^2\lambda$, and vice versa. The ideal value for α_{ENTER} will yield a model with expected MSE $\sigma^2 + \tau^2\lambda$, which is linear in λ with slope τ^2. The algorithm for selecting the ideal value of α_{ENTER} is given by Luo et al (2006) as follows. First choose a grid of noise levels $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_m \leq 4$, and a grid of α_{ENTER} levels $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_s < 1$ and generate a large number of i.i.d. $N(0,1)$ random variables $(W_{i,k,b}, i = 1, \ldots, n; k = 1, \ldots, m; b = 1, \ldots, B)$. Then for each combination of $\lambda_k, k = 1, \ldots$ and $\alpha_j, j = 1, \ldots, s$, there are B i.i.d. remeasured data sets $(Y_{i,k,b}^*, X_i)_{i=1}^n$ for $b = 1, \ldots, B$, where $Y_{i,k,b}^* = Y_i + \sqrt{\lambda_k}\tau W_{i,k,b}$. For each of these remeasured data sets, perform forward selection with $\alpha_{ENTER} = \alpha_j$ and compute $MSE_{\alpha_j,b}(\lambda_k)$ from the selected model. The average of the MSE's across the B data sets is given by $\bar{MSE}_{\alpha_j}(\lambda_k) = \frac{1}{B}\sum_{b=1}^B MSE_{\alpha_j,b}(\lambda_k)$. This gives s simple linear regression data sets $(\lambda_k, \bar{MSE}_{\alpha_j}(\lambda_k))_{k=1}^m$ for $j = 1, 2, \ldots, s$. Compute the estimated slope of each of these data sets and choose the one that is closest to τ^2. The corresponding α_j is chosen as α_{ENTER}.

The variable addition model selection (VAMS) method of Wu et al (2007) chooses the tuning parameter by adding nonimportant phony variables to the model and performing a variable selection method on the new model. By keeping track of how many of these variables are included and excluded by the selection method, Wu et al (2007) can reveal the underfitting or overfitting tendencies of the method. As with the NAMS method, forward selection is discussed, but other methods can be used also. The parameter α_{ENTER} is tuned to control the false selection rate (FSR), which is the proportion of nonimportant variables

included in the selected model. The FSR of the real variables cannot be directly observed because it is not known which of them are important. When nonimportant pseudovariables are added to the data, and forward selection is applied to the the data $(Y_i, X_i, P_i)_{i=1}^n$, where P_i is a vector of nonimportant phony variables, then the proportion of falsely selected phony variables can be observed since it is known that these variables are nonimportant. Replicating this process allows the FSR of the real variables to be estimated by using the proportions of phony variables selected. This can be done for a grid of values for the tuning parameter, and the value of α_{ENTER} is chosen so that the estimated FSR of the real variables is approximately equal to a desired level. Boos et al (2008) modify this method so that no simulation is required and the computation is more simple. This method uses the fast FSR tuning method and can be applied even when the number of covariates is larger than the sample size and is applied to clinical trial data for linear, logistic, and Cox proportional hazards regression.

Recently, several shrinkage methods that perform simultaneous variable selection and estimation have been studied. Tibshirani (1996) notes that the ordinary least squares (OLS) estimators usually have small bias but may have large variances and therefore not be optimal in terms of prediction accuracy and seeks to develop a method that will perform variable selection and also yield estimates with adequate prediction accuracy. He proposes that shrinking some coefficients and setting others to 0 could reduce the variance of predicted values by allowing more bias. Further, setting coefficients to 0 allows the elimination of the corresponding variables. He notes that any method of best subset selection is a discrete process and is not always stable in terms of prediction accuracy. Ridge regression, a continuous process, is more stable, but it does not perform variable selection. While it shrinks some coefficients, it cannot set them to 0 and eliminate the corresponding variables. Tibshirani (1996) proposes a technique called "least absolute shrinkage and selection operator (Lasso)", which seeks to combine the desirable properties of ridge regression and best subset selection. It has the continuous shrinkage of ridge regression while still performing variable selection. The lasso procedure, having an L_1 penalty, is able to set some coefficient estimates to 0, hence eliminating the least important variables.

The Lasso estimates are obtained by minimizing

$$\sum_{i=i}^{n} \left(Y_i - \theta' X_i\right)^2 + n\lambda \sum_{j=1}^{p} |\beta_j|,$$

where λ is a tuning parameter. The tuning parameter is the penalty on the number of variables selected, and therefore determines how much shrinkage is applied to the resulting estimates (Tibshirani 1996).

Zou (2006) notes that in some situations, the Lasso can be inconsistent in variable selection. He proposes using different weights for different coefficients, thereby placing larger penalties on the non-important coefficients and smaller penalties on the important ones. If the weights are data-dependent and chosen properly, Zou (2006) shows that the weighted Lasso

$$\arg\min_{\theta} \sum_{i=i}^{n} \left(Y_i - \theta' X_i\right)^2 + n\lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where w is a known vector of weights, can asymptotically have the oracle properties as defined by Fan and Li (2001). These properties are that for an estimator $\hat{\theta}$, $a = \{j : \beta_{j0} \neq 0\} = \{j : \hat{\beta}_j \neq 0\}$ (the true model is identified) with probability tending to one and that $\sqrt{n}(\hat{\beta}_a - \beta_{a0})$ converges in distribution to Normal with mean vector 0 and covariance matrix Σ^*, where Σ^* is the covariance matrix of the true subset model.

Zou (2006) defines the adaptive Lasso estimator as

$$\arg\min_{\theta} \sum_{i=i}^{n} \left(Y_i - \theta' X_i\right)^2 + n\lambda \sum_{j=1}^{p} \frac{1}{|\hat{\beta^*}_j|} |\beta_j|,$$

where $\hat{\beta}^*$ is a root-n consistent estimator for β_0. Since it is a consistent estimator, the values of $\hat{\beta}^*$ should reflect the relative importance of the covariates (Zhang and Lu 2007). With a proper choice of λ, Zou (2006) shows that the adaptive Lasso can have the aforementioned oracle properties. In particular, if $\sqrt{n}\lambda \to 0$ and $n\lambda \to \infty$, then the estimator has these properties.

Fan and Li (2001) propose the smooth clipped absolute deviation (SCAD) estimator, which minimizes

$$\sum_{i=1}^{n} (Y_i - \theta' X_i)^2 + n \sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where $p_\lambda(0) = 0$, and the first derivative is

$$p'_\lambda(\vartheta) = I(\vartheta \leq \lambda) + \frac{(a\lambda - \vartheta)_+}{(a-1)\lambda} I(\vartheta > \lambda)$$

for some $a > 2$ and $\vartheta > 0$ and $(r)_+ = rI(r > 0)$. By minimizing a Bayes risk, Fan and Li (2001, 2002) suggest using $a = 3.7$. They show that with the proper choice of regularization parameters, the SCAD has the aforementioned oracle properties.

Wang et al (2007) note that the L_2 loss function may cause the estimators to be too sensitive to severe outliers and/or heavy-tailed errors. Their method, the LAD-Lasso, uses both an L_1 loss function along with the L_1 penalty of the adaptive Lasso. Since the median is in general more resistant to outliers and heavy-tailed error distributions, the L_1 loss function enjoys many advantages. The LAD-Lasso estimates are obtained by minimizing

$$\sum_{i=1}^{n} |Y_i - \theta' X_i| + n\lambda \sum_{j=1}^{p} w_j |\beta_j|.$$

An added advantage of this method is that the L_1 loss function allows the solution to be obtained by linear programming, as the problem is equivalent to minimizing

$$\sum_{i=1}^{n+p} |Y_i^* - \theta' X_i^*|,$$

where $Y_i^* = Y_i$ for $i = 1, 2, \ldots, n$ and 0 for $i = n+1, \ldots, n+p$. Also, X_i^* is a $(n+p) \times p$ matrix where the first n rows consist of the matrix X. The remaining p rows consist of a column vector of p zeroes and a p-dimemsional diagonal matrix with diagonal elements $n\lambda w_j$ for $j = 1, \ldots, p$. We can see then that this is simply an unpenalized LAD fit with the augmented form of the data.

The major concerns of this method are the tuning procedure and the choice of weights. As with the adaptive Lasso with L_2 loss, with certain conditions on λ and a proper choice of weights, the resulting estimator has many desirable properties. Suppose the data are arranged so that for $\beta_0 = (\beta_{a0}, \beta_{b0})$, where $\beta_{a0} = (\beta_{10}, \ldots, \beta_{p'0})$ represents the significant coefficients, and $\beta_{b0} = (\beta_{(p'+1)0}, \ldots, \beta_{p0})$ are the insignificant ones. Of course then, $\beta_{a0} \neq 0$ and $\beta_{b0} = 0$. Just as in the adaptive Lasso case with L_2 penalty, it can be shown that with the proper choice of weights and tuning parameter, the resulting estimator is sparse and is asymptotically normal. Wang et al (2007) show that if the weights are chosen so that the inverses are root-n consistent, and λ is chosen so that $\sqrt{n}\lambda \to 0$ and $n\lambda \to \infty$ as $n \to \infty$, then the resulting estimator $\hat{\beta}$ is root-n consistent, satisfies $\hat{\beta}_b = 0$ with probability tending to 1 (sparcity), and that $(\beta_{a0} - \hat{\beta}_a)$ converges in distribution to a p'-dimensional normal random vector with mean 0.

Wang et al (2007) derive an appropriate way to select tuning parameters and weights so that the previously mentioned conditions are met. They suggest tuning parameter estimates $\hat{\lambda} = \frac{\log(n)}{n}$ and $\hat{w}_j = \frac{1}{|\tilde{\beta}_j|}$ where $\tilde{\beta}$ is the estimate from an unpenalized LAD fit. These values meet both of the conditions on the weights and λ.

1.2.2 Variable selection for censored data

In the last section, we discussed some variable selection methods for uncensored data. From now on, we will consider failure time data that is possibly right-censored. In a study of n subjects, then, each subject has failure time T_i and censoring time C_i with a p-dimensional vector of covariates Z_i. The observed data consists of $(\tilde{T}_i, \delta_i, Z_i), i = 1, \ldots, n$, where $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. The Cox proportional hazards model (Cox 1972) assumes that the hazard function for a subject is given by

$$\lambda(t|Z_i) = \lambda_0(t) \exp\{\beta' Z_i\},$$

where $\lambda_0(t)$ is a completely unspecified baseline hazard function. This model is one of the most well-studied in the field of survival analysis, including shrinkage methods. There is a rich literature for variable selection in standard linear regression and survival data analysis based on the PH model. Traditional procedures include backward deletion, forward addition and stepwise selection. However, these procedures may suffer from high variability (Breiman, 1996). Recently some shrinkage methods have been proposed based on the penalized likelihood or partial likelihood estimation, including the LASSO (Tibshirani, 1996, 1997), the SCAD (Fan and Li, 2001, 2002) and the adaptive LASSO (Zou 2006, 2008; Zhang and Lu, 2007). Boos et al (2008) applied the fast FSR VAMS method to this model along with foward selection to a data set in order to tune the parameter α_{ENTER}.

Recently, there has been much work done in applying shrinkage methods to censored survival data under the Cox model. Several authors, including Tibshirani (1997), Fan and Li (2002), Zhang and Lu (2007), and Zou (2008), have proposed minimizing the penalized log partial likelihood function based on the Cox model, where the log partial likelihood is given by

$$\ell_n(\beta) = \sum_{i=1}^{n} \delta_i \left[\beta' Z_i - \log \left\{ \sum_{j=1}^{n} I(\tilde{T}_j \geq \tilde{T}_i) \exp(\beta' Z_j) \right\} \right].$$

Tibshirani proposes using the Lasso penalty along with the log partial likelihood. The estimates are obtained by minimizing

$$-\ell_n(\beta) + n\lambda \sum_{j=1}^{p} |\beta_j|.$$

While simulation studies indicated that the Lasso can be more accurate than stepwise selection (Tibshirani 1997), the estimator does not have the oracle properties defined previously

(Zhang and Lu 2007). As is the case in the classical regression setting, having only one tuning parameter may not produce solutions that are sparse enough. Fan and Li (2002) propose the SCAD penalty, where estimators minimize

$$-\ell_n(\beta) + n\sum_{j=1}^{p} p_\lambda(|\beta_j|),$$

where $p_\lambda(0) = 0$, and the first derivative is

$$p'_\lambda(\vartheta) = I(\vartheta \leq \lambda) + \frac{(a\lambda - \vartheta)_+}{(a-1)\lambda} I(\vartheta > \lambda)$$

for some $a > 2$ and $\vartheta > 0$ and $(r)_+ = rI(r > 0)$. By minimizing a Bayes risk, Fan and Li (2001, 2002) suggest using $a = 3.7$. As is the case with uncensored data, they show that with the proper choice of regularization parameters, the SCAD performs as well as an oracle estimator.

However, Zhang and Lu (2007) note that numerical difficulties arise due to the nonconvex form of the penalty. They propose an adaptive Lasso method based on a penalized partial likelihood with adaptively weighted L_1 penalties. Since the penalty has a convex form, the problem can be solved more efficiently. The estimator minimizes

$$-\ell_n(\beta) + n\lambda\sum_{j=1}^{p} w_j|\beta_j|,$$

where the positive weights are data dependent. They suggest using $w_j = \frac{1}{|\tilde{\beta}_j|}$, where $\tilde{\beta}$ is the maximizer of $\ell_n(\beta)$, but note that any consistent estimator of β can be used. Zhang and Lu (2007) show that if $\sqrt{n}\lambda \to 0$ the adaptive Lasso estimator is root-n consistent, and that in addition, if $n\lambda \to \infty$, then the estimator exhibits sparcity and asymptotic normality of the non-zero coefficients. Zou (2008) proposes an efficient and adaptive shrinkage method that uses the adaptive Lasso penalty. Instead of using the partial likelihood $\ell(\beta)$, he uses an efficient quadratic approximation that permits more efficient computations without losing the information about β that is contained in $\ell(\beta)$. Zou (2008) shows that if $\sqrt{n}\lambda \to 0$ and $n\lambda \to \infty$, then the resulting estimators are sparse and asymptotically normal. As mentioned before, however, the assumptions of the proportional hazards model may not be met in many practical applications and other models may be appropriate.

Lu and Zhang (2007) study variable selection for the PO model by maximizing the marginal likelihood subject to both the Lasso and Adaptive Lasso penalties. They

provide an efficient computation algorithm for these methods, and their numerical results indicate that they can produce accurate models, noting that the adaptive Lasso tends to work better than the Lasso. However, since the marginal likelihood has no closed form, it must be computed numerically; therefore, is hard to establish the theoretical properties of the estimator.

For the AFT and other semiparametric linear models with i.i.d. errors, Johnson (2008) examines two procedures for variable selection. The first uses shrinkage estimators with different penalties, including the SCAD and Lasso, using both the rank estimation of Prentice (1978) and the Buckley-James estimator (Buckley and James, 1979). The second method uses the VAMS method of Wu et al (2007) by adding phony variables to the data and controlling the false selection rate (FSR) to adapt the tuning parameter α_{ENTER}. With a proper choice of tuning parameters, he shows that the resulting estimators are root-n consistent, exhibit sparcity, and are asymptotically normal. Johnson et al (2008) develop asymptotic theory for a broad class of penalized estimating functions that may not pertain to the derivatives of any objective functions. Under certain conditions placed on the estimating functions and penalty parameters, the resulting estimates are shown to be consistent, sparse, and asymptotically normal. In addition, Johnson et al (2008) develop both an algorithm for implementing the estimators and a resampling technique for estimating the variances of the estimates. The resampling technique can be used when the asymptotic variances cannot be evaluated directly. Cai et al (2008) propose a regularized rank-based estimation procedure with a Lasso-type penalty for estimation and variable selection under the AFT model. This procedure only requires that the censoring time is conditionally independent of the failure time given the covariates. They further show that their estimator is a solution to a linear optimization procedure and the method is fairly robust against model misspecification. While robust procedures for the AFT model exist, as mentioned before, however, the assumption of i.i.d. errors may be too restrictive. In this case, and in the case of heteroscedatic/heavy-tailed error distributions, censored median regression may be more appropriate.

1.3 Plan of work

To the best of our knowledge, the sparse estimation of censored quantile regression has not been studied in the literature. When there is no censoring, Wang et al. (2007) considered the LAD-LASSO estimation for variable selection in median regression. In this paper, we focus on the sparse estimation of censored median regression. However, the proposed method can be easily extended to arbitrary quantile regression with random censoring. Here we use the inverse censoring probability weighted least absolute deviation estimation method of Zhou (2006) for censored median regression and incorporate adaptive LASSO penalty for consistent variable selection. The entire solution path will be shown to be piecewise linear, and the estimator is evaluated in extensive simulation studies and applied to two data sets. The method will also be extended for use with multivariate or clustered survival data.

Chapter 2

Sparse Censored Median Regression

2.1 The Sparse Censored Median Regression estimator

As discussed before, the assumptions of the proportional hazards, proportional odds, and AFT models are often violated in practice, which leads to the use of other models. Censored quantile regression is attractive due to its robustness against outliers and heteroscedastic or heavy-tailed error distributions. The method developed in this section will use the inverse censoring probability weighted LAD estimation proposed by Zhou (2006) with the adaptive Lasso penalty. Consider a study of n subjects. Let $\{T_i, C_i, Z_i, i = 1, \cdots, n\}$ denote n i.i.d. triplets of failure times, censoring times, and p-dimensional covariates of interest. Conditional on Z_i, the median regression model assumes

$$\log(T_i) = \theta_0' X_i + \epsilon_i, \quad i = 1, \cdots, n, \tag{2.1}$$

where $X_i = (1, Z_i')'$, $\theta_0 = (\alpha_0, \beta_0')'$ is a $(p+1)$-dimensional vector of regression parameters and ϵ_i's are assumed to have a conditional median of 0. Note that the log transformation in (2.1) can be replaced by any specified monotonic transformation. Define $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, then the observed data consist of $(\tilde{T}_i, \delta_i, Z_i, i = 1, \cdots, n)$. As in Ying et al. (1995) and Zhou (2006), we assume that the censoring time C is independent of T and Z. Let $G(\cdot)$ denote the survival function of censoring times. In the absence of censoring, Ying et al (1995) note that the LAD estimates for θ_0 are obtained by minimizing

$\sum_{i=1}^{n} |\log(T_i) - \theta' X_i|$. If we take the derivative of this quantity, we obtain

$$\begin{aligned}
\tfrac{d}{d\theta} \textstyle\sum_{i=1}^{n} |\log(T_i) - \theta' X_i| &= -\textstyle\sum_{i=1}^{n} X_i sign(\log(T_i) - \theta' X_i) \\
&= -\textstyle\sum_{i=1}^{n} X_i \{I(\log(T_i) - \theta' X_i \geq 0) - I(\log(T_i) - \theta' X_i \leq 0)\} \\
&= -\textstyle\sum_{i=1}^{n} X_i \{2I(\log(T_i) - \theta' X_i \geq 0) - 1\} \\
&= -2\textstyle\sum_{i=1}^{n} X_i \left\{I(\log(T_i) - \theta' X_i \geq 0) - \tfrac{1}{2}\right\}.
\end{aligned}$$

Although this derivative is not continuous in θ, Ying et al (1995) propose the estimating equation

$$U_n(\theta) = \sum_{i=1}^{n} X_i \left\{ I(\log(T_i) - \theta' X_i \geq 0) - \frac{1}{2} \right\} = 0$$

motivated by the fact that the expected value (conditional on Z) of $U_n(\theta_0)$ is 0. With censored data, the expected value of $I(\log(\tilde{T}_i) - \theta_0' X_i \geq 0)$ is $\frac{1}{2} G(\theta_0' X_i)$. Therefore, in order to estimate the parameters θ_0, Ying et al. (1995) propose to solve

$$\sum_{i=1}^{n} X_i \left[\frac{I\{\log(\tilde{T}_i) - \theta' X_i \geq 0\}}{\hat{G}(\theta' X_i)} - \frac{1}{2} \right] = 0,$$

where $\hat{G}(\cdot)$ is the Kaplan-Meier estimator of $G(\cdot)$ based on the data $(\tilde{T}_i, 1-\delta_i)$, $i = 1, \cdots, n$. Note that the above equation is neither continuous nor monotone in θ and thus is difficult to solve especially when the dimension of θ is high. Bang and Tsiatis (2002) note that at $\tilde{T}_i$, the probability that the ith individual is not censored is $G(\tilde{T}_i)$. They show that the weighted estimating equation using only the observed failure times, given by

$$\sum_{i=1}^{n} \frac{\delta_i}{G(\tilde{T}_i)} X_i \left[I\{\log(\tilde{T}_i) - \theta' X_i \geq 0\} - \frac{1}{2} \right] = 0,$$

is unbiased. The alternative inverse censoring probability weighted estimating equation as proposed by Bang and Tsiatis (2002) is as follows:

$$\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}(\tilde{T}_i)} X_i \left[I\{\log(\tilde{T}_i) - \theta' X_i \geq 0\} - \frac{1}{2} \right] = 0.$$

As pointed out by Zhou (2006), the solution of the above equation is also a minimizer of

$$\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}(\tilde{T}_i)} |\log(\tilde{T}_i) - \theta' X_i|, \tag{2.2}$$

which is convex in θ and can be easily solved using an efficient linear programming algorithm of Koenker and D'Orey (1987). In addition, Zhou (2006) shows that the above inverse

censoring probability weighted least absolute deviation estimator has better finite sample performance than that of Ying et al. (1995).

Let $\tilde{\theta} \equiv (\tilde{\alpha}, \tilde{\beta}')'$ denote the minimizer of (2.2). To conduct variable selection for censored median regression, we propose to minimize

$$\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}(\tilde{T}_i)} |\log(\tilde{T}_i) - \theta' X_i| + n\lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|}, \tag{2.3}$$

where $\lambda > 0$ is the tuning parameter and $\tilde{\beta} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_p)'$. Here we use the adaptive LASSO penalty (Zou, 2006; Zhang and Lu, 2007; Wang et al., 2007; among others) for variable selection. It has been well studied that if the tuning parameter is properly chosen, the adaptive LASSO procedure can produce consistent variable selection. We have already shown that this is the case for uncensored data and under the Cox proportional hazards model. In the next section, we will show that this is also true for our estimator.

2.2 Theoretical properties

Suppose that $\beta_0 = (\beta'_{a0}, \beta'_{b0})'$ and is arranged so that $\beta_{a0} = (\beta_{10}, \cdots \beta_{q0})' \neq 0$ and $\beta_{b0} = (\beta_{(q+1)0}, \ldots, \beta_{p0})' = 0$. Correspondingly, denote the minimizer of (2.3) by $\hat{\theta} = (\hat{\alpha}, \hat{\beta}'_a, \hat{\beta}'_b)'$. Define $\beta_{a'0} = (\alpha_0, \beta'_{a0})'$ and $\hat{\beta}_{a'} = (\hat{\alpha}, \hat{\beta}'_a)'$. Under the regularity conditions given in the Appendix, we have the following two theorems.

Theorem 1 *($\sqrt{n}$-Consistency) If $\sqrt{n}\lambda = O_p(1)$, then $\sqrt{n}||\hat{\theta} - \theta_0|| = O_p(1)$.*

Theorem 2 *If $\sqrt{n}\lambda \to 0$ and $n\lambda \to \infty$, then as $n \to \infty$*
(i) (Selection-Consistency) $P\left(\hat{\beta}_b = 0\right) \to 1$.
(ii) (Asymptotic Normality) $\sqrt{n}(\hat{\beta}_{a'} - \beta_{a'0})$ converges in distribution to a normal random vector with mean 0 and variance-covariance matrix $\Sigma_{a'}^{-1} V_{a'} \Sigma_{a'}^{-1}$.

The definitions of $\Sigma_{a'}$ and $V_{a'}$, and the proofs of Theorems 1 and 2 are given in Appendix A.

2.3 Computational Algorithm

2.3.1 Augmentation method

To compute the SCMR estimator defined by the minimizer of (2.3), we first represent it as a least absolute deviation problem. To be specific, define

$$(X^*)' = \begin{pmatrix} \frac{\delta_1}{\hat{G}(\tilde{T}_1)} & \frac{\delta_1}{\hat{G}(\tilde{T}_1)}X_{11} & \cdots & \frac{\delta_1}{\hat{G}(\tilde{T}_1)}X_{1p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\delta_n}{\hat{G}(\tilde{T}_n)} & \frac{\delta_n}{\hat{G}(\tilde{T}_n)}X_{n1} & \cdots & \frac{\delta_n}{\hat{G}(\tilde{T}_n)}X_{np} \\ 0 & \frac{n\lambda}{|\tilde{\beta}_1|} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{n\lambda}{|\tilde{\beta}_p|} \end{pmatrix}_{(n+p)\times(p+1)}$$

and

$$Y^* = \left(\frac{\delta_1}{\hat{G}(\tilde{T}_1)} \log(\tilde{T}_1), \cdots, \frac{\delta_n}{\hat{G}(\tilde{T}_n)} \log(\tilde{T}_n), 0, \cdots, 0 \right)'_{(n+p)\times 1}.$$

Then, for any given λ, minimizing (2.3) is equivalent to minimizing $\sum_{i=1}^{n} |Y^* - \theta' X_i^*|$, which can be easily computed using any statistical software for linear programming, for example, the rq function in R. The estimates are obtained by the line of code $rq(Y^* \sim X^*)\$coeff$. We obtained solutions for each value of λ on a grid incremented by .001. The solutions given by this code may not be exactly zero for unimportant coefficients, so we set a threshhold value of 10^{-6}; that is, if the absolute value of an estimated coefficient is less than 10^{-6}, then we take the solution to be zero. This numerical problem is resolved in the following solution path algorithm.

2.3.2 Solution path algorithm

Recently, Li and Zhu (2008) derived the solution path packages for sparse quantile regression with uncensored data. It turns out the problem (2.3) can nicely fit in their framework and therefore, its solution path can be easily obtained by modifying their algorithm to incorporate the inverse-censoring-probability weights in the penalized median regression. Suppose that the observed failure times are given by $T_{(1)}, \ldots, T_{(m)}$ with corresponding covariate vectors $Z_{(1)}, \ldots, Z_{(m)}$, so that the corresponding censoring indicators $\delta_{(i)}$ are all 1

for $i = 1, \ldots, m$. Then for each value of λ, our problem is equivalent to minimizing

$$\sum_{i=1}^{m} b_{(i)} |\log(T_{(i)}) - \alpha - \sum_{j=1}^{p} \beta_j'^* Z_{(i)j}| + n\lambda \sum_{j=1}^{p} |\beta_j^*|,$$

where $b_{(i)} = \frac{1}{\hat{G}(T_{(i)})}$ and $\beta_j^* = \frac{\beta_j}{|\tilde{\beta}_j|}$. It has been pointed out by numerous authors, including Tibshirani (1996), that this is equivalent to minimizing $\sum_{i=1}^{m} b_{(i)} |\log(T_{(i)}) - \alpha - \sum_{j=1}^{p} \beta_j^* Z_{(i)j}|$ subject to $\sum_{j=1}^{p} |\beta_j^*| \leq s$, and since there is a one-to-one correspondence between λ and s, the solutions are identical. Li and Zhu (2008) show that for any given s, the solution $\hat{\beta}^*(s)$ is a piecewise linear function of s, and they give an efficient algorithm for computing the entire solution path $\{\hat{\beta}^*(s), 0 \leq s < \infty\}$. For simplicity, we will specifically consider median regression and the case when m is odd and give a description of the algorithm in this case. Since the absolute value function is convex, we can rewrite the constrained optimization problem as minimizing

$$\frac{1}{2} \sum_{i=1}^{m} (\xi_i + \zeta_i)$$

subject to $\sum_{j=1}^{p} |\beta_j^*| \leq s$, where $-\zeta_i \leq b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} \leq \xi_i$ for $\zeta_i, \xi_i \geq 0$, $i = 1, \ldots, m$, and $f(Z_{(i)}) = \alpha + \sum_{j=1}^{p} \beta_j^* Z_{(i)j}$. For the non-negative Lagrange multipliers $\lambda^*, \omega_i, \gamma_i, \kappa_i$, and η_i, this is equivalent to minimizing the Lagrangian function

$$\begin{aligned} L_p \;=\; & \tfrac{1}{2} \textstyle\sum_{i=1}^{m} (\xi_i + \zeta_i) + \lambda^* \left(\sum_{j=1}^{p} |\beta_j^*| - s \right) + \sum_{i=1}^{m} \omega_i [b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} - \xi_i] \\ & - \textstyle\sum_{i=1}^{m} \gamma_i [b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} + \zeta_i] - \sum_{i=1}^{m} \kappa_i \xi_i - \sum_{i=1}^{m} \eta_i \zeta_i. \end{aligned}$$

Setting the partial derivatives with respect to β^*, α, ξ_i, and ζ_i equal to zero, we have

$$\begin{aligned} \tfrac{\partial}{\partial \beta^*} : \lambda^* \cdot sign(\beta_j^*) &= \textstyle\sum_{i=1}^{m} b_{(i)} (\omega_i - \gamma_i) Z_{(i)j}, \text{ for all } j \text{ with } \beta_j \neq 0 \\ \tfrac{\partial}{\partial \alpha} : \textstyle\sum_{i=1}^{m} b_{(i)} (\omega_i - \gamma_i) &= 0 \\ \tfrac{\partial}{\partial \xi_i} : \omega_i + \kappa_i &= \tfrac{1}{2} \\ \tfrac{\partial}{\partial \zeta_i} : \gamma_i + \eta_i &= \tfrac{1}{2} \end{aligned}$$

In order for the solution to be unique, the Karush-Kuhn-Tucker (KKT) conditions

$$\begin{aligned} \omega_i [b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} - \xi_i] &= 0 \\ \gamma_i [b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} + \zeta_i] &= 0 \\ \kappa_i \xi_i &= 0 \\ \eta_i \zeta_i &= 0 \end{aligned}$$

must be met for all i. Based on these equations and conditions, Li and Zhu (2008) conclude that

$$\begin{array}{lllllll} b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} > 0 & \rightarrow & \omega_i = \frac{1}{2}, & \xi_i > 0, & \gamma_i = 0, & \zeta_i = 0; \\ b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} < 0 & \rightarrow & \omega_i = 0, & \xi_i = 0, & \gamma_i = \frac{1}{2}, & \zeta_i > 0; \\ b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} = 0 & \rightarrow & \omega_i \in [0, \frac{1}{2}], & \xi_i = 0, & \gamma_i \in [0, \frac{1}{2}], & \zeta_i = 0. \end{array}$$

The equations based on the partial derivatives with respect to β and α depend on ω_i and γ_i only through $\mu_i = b_{(i)}(\omega_i - \gamma_i)$. Using this, they define the sets

$$\begin{array}{lcl} \varepsilon & = & \{i : b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} = 0, -\frac{1}{2} \leq \mu_i \leq \frac{1}{2}\}\text{(elbow)} \\ L & = & \{i : b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} < 0, \mu_i = -\frac{1}{2}\}\text{(left of the elbow)} \\ R & = & \{i : b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\} > 0, \mu_i = \frac{1}{2}\}\text{(right of the elbow)} \\ \upsilon & = & \left\{j : \hat{\beta}_j^* \neq 0\right\}\text{ (active set)} \end{array}$$

Li and Zhu (2008) examine how the KKT conditions change as s increases in order to compute the solution path $\beta^*(s)$. As s increases, they define an *event* to be

1. Type I event: A data point reaches (or leaves) the elbow set (a residual $b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\}$ changes from nonzero to zero or vice versa)

2. Type II event: A variable leaves (or enters) the model (a coefficient estimate changes from nonzero to zero or vice versa)

The time that a Type I event occurs corresponds to a nonsmooth point of $\sum_{i=1}^{m} b_{(i)}|\log(T_{(i)}) - f(Z_{(i)})|$, and the time that a Type II event occurs corresponds to a nonsmooth point of $\sum_{j=1}^{p} |\beta_j^*|$. Li and Zhu (2008) note that the sets ε, L, R, and υ will not change unless an event occurs. Equivalently, the KKT conditions will not change unless an event happens. For any residual $b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\}$ that is not in the elbow set, the corresponding μ_i is known, while μ_i is not known if the residual is in the elbow set, so there are only $|\varepsilon|$ values of μ_i that are unknown. So if the KKT conditions do not change, there are $|\varepsilon|$ unknowns (λ^* and μ_i for $i \in \varepsilon$). Since $\lambda^* \cdot sign(\beta_j^*) = \sum_{i=1}^{m} \mu_i Z_{(i)j}$, for all j with $\beta_j^* \neq 0$ and $\sum_{i=1}^{m} \mu_i = 0$, there are $|\upsilon| + 1$ equations. Therefore, for the KKT conditions to be met (and the solution to be therefore unique), it must be true that $|\varepsilon| = |\upsilon|$. That is, the number of residual points in the elbow must be equal to the number of variables in the active set. Li and Zhu

(2008) also note that as s increases, unless a Type I or Type II event occurs, a residual in the elbow will stay there. This means that

$$b_{(i)}\{\log(T_{(i)}) - (\alpha + \sum_{j\in\upsilon} \beta_j Z_{(i)j})\} = 0$$

for all i in the elbow set. There are $|\varepsilon|$ equations in this set, with $|\upsilon| + 1$ unknowns. Since $|\varepsilon| = |\upsilon|$, there is only one free unknown, and therefore the solution for β changes linearly in s (until an event occurs). Li and Zhu (2008) begin with $s = 0$ and keep track of the location of residuals relative to the elbow as s increases. For a point in R to pass through ε, its value for μ_i must change from $\frac{1}{2}$ to $-\frac{1}{2}$, and vice versa for points in L. Li and Zhu (2008) note that by continuity, points in the elbow will linger there, and the set will stay stable until an event occurs.

For $s = 0$, it is necessary that $\beta^* = 0$, and only α is included in the model. The initial estimate $\hat{\alpha} = b_{i_1^*} \log(T_{i_1^*})$ is the weighted sample median (we will consider the case where it is a data point, i.e., m is odd). In this case, $(Z_{i_1^*}, Y_{i_1^*})$ is in the elbow set. For the KKT conditions to be met, we must have that $|\varepsilon| = |\upsilon|$. For $s > 0$, $|\varepsilon| = 1$ and $|\upsilon| = 0$, so a variable must be added to the model. This variable is $\beta^*_{j_1^\star}$, where $j_1^\star = \arg\max_j |\sum_{i=1}^n \mu_i Z_{(i)j}|$, where $\mu_i = \frac{1}{2}$ if $b_{(i)}\{\log(T_{(i)}) - f(Z_{(i)})\}$ is to the left of the elbow, $-\frac{1}{2}$ if it is to the right, and 0 if the point is in the elbow. Because of the initial restriction $\sum_{j=1}^p |\beta_j^*| \leq s$, we have that $|\hat{\beta}^*_{j_1^\star}| = s$. Since $sign(\beta_j^*) = sign(\sum_{i=1}^m \mu_i Z_{(i)j})$, this implies that $\hat{\beta}^*_{j_1^\star}(s) = s \cdot sign(\sum_{i=1}^m \mu_i Z_{(i)j})$ and $\hat{\beta}^\star_j(s) = 0$ for $j \neq j_1^\star$. For a small enough s, the sets will not change, so we have

$$\hat{f}(Z) = \hat{\alpha}(s) + s \cdot sign(\sum_{i=1}^m \mu_i Z_{(i)j^\star}) Z_{j^\star}$$

where

$$\hat{\alpha}(s) = b_{i_1^*} \log(T_{i_1^*}) - s \cdot sign\left(\sum_{i=1}^m \mu_i Z_{(i)j^\star}\right) Z_{i_1^* j^\star}.$$

The solution will continue to change linearly in s and the sets will remain the same until an event occurs. Suppose that the ℓth event occurs immediately before s^ℓ, and that the superscript ℓ indexes the sets, function, and parameter values immediately after the event occurs. For $s^\ell < s < s^{\ell+1}$, $\hat{\beta}^*(s)$ changes linearly in s. The time that the next event occurs, $s^{\ell+1}$, can be calculated in the following way: Since

$$v_0 + \sum_{j\in\upsilon^\ell} v_j b_{(i)} Z_{(i)j} = 0, \text{ for } i \in \varepsilon^\ell$$

and

$$\sum_{j \in v^\ell} v_j \cdot sign(\beta_j^{*\ell}) = 1$$

where $v_0 = (\alpha - \alpha^\ell)/(s - s^\ell)$ and $v_j = (\beta_j^* - \beta_j^{*\ell})/(s - s^\ell)$ these equations can be solved for v and these values can be used to calculate $s^{\ell+1}$. An event occurs either when

$$\beta_j^* = \beta_j^{*\ell} + (s - s^\ell)v_j = 0 \text{ for all } j \in v^\ell$$

or

$$b_{(i)} f(Z_{(i)}) = (s - s^\ell)(v_0 + \sum_{j \in v^\ell} v_j Z_j) + b_{(i)} f^\ell(Z_{(i)}) = b_{(i)} \log(T_{(i)}) \text{ for all } i \notin v^\ell.$$

It is not possible for $\hat{\beta}_j^* = 0$ if $j \in v^\ell$, so we are concerned when a point enters the elbow. When this happens, either one point must be taken out of the elbow or another variable must be added to the model in order to maintain the KKT conditions. The decision is made by finding the smallest (negative) $\frac{\Delta \text{loss}}{\Delta s}$, which is given by

$$\begin{aligned} \frac{\Delta loss}{\Delta s} &= \frac{\sum_i b_{(i)} |\log(T_{(i)}) - f(Z_{(i)})| - \sum_i b_{(i)} |\log(T_{(i)}) - f^\ell(Z_{(i)})|}{s - s^l} \\ &= \frac{1}{2} \sum_{i \in L} \left(v_0 + \sum_{j \in v} b_{(i)} v_j Z_{(i)j} \right) - \frac{1}{2} \sum_{i \in R} \left(v_0 + \sum_{j \in v} b_{(i)} v_j Z_{(i)j} \right). \end{aligned}$$

The algorithm ends when all the $\frac{\Delta \text{loss}}{\Delta s}$ are non-negative. For each value of s, the solution given by the algorithm is $\hat{\beta}_j^*(s)$, and we obtain our estimate as $\hat{\beta}_j(s) = |\tilde{\beta}_j| \hat{\beta}_j^*(s)$.

2.4 Parameter Tuning

Our estimator makes use of the adaptive Lasso penalty, i.e., a penalty of the form $\lambda \sum_{j=1}^p \frac{|\beta_j|}{|\tilde{\beta}_j|}$, where $\tilde{\beta}$ is the unpenalized LAD estimator. We propose to tune the parameter λ based on a BIC-type criteria. Note that if there is no censoring, the least absolute deviation loss is closely related to linear regression with double exponential error. To derive the BIC tuning procedure, we first assume that the error term ϵ_i's are i.i.d. double exponential variables with the scale parameter σ. When there is no censoring, up to a constant the negative log likelihood function can be written as $\sum_{i=1}^n |\log(T_i) - \theta' X_i|/\sigma$ and the maximum likelihood estimator of σ is given by $\sum_{i=1}^n |\log(T_i) - \tilde{\theta}' X_i|/n$ with $\tilde{\theta}$ being the LAD estimator. Motivated by this observation, we propose the following tuning

procedure for sparse censored median regression:

$$BIC(\lambda) = \frac{2}{\tilde{\sigma}} \sum_{i=1}^{n} \frac{\delta_i}{\hat{G}(\tilde{T}_i)} |\log(\tilde{T}_i) - X_i'\hat{\theta}(\lambda)| + \log(n) \cdot r,$$

where $\tilde{\sigma} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\hat{G}(\tilde{T}_i)} |\log(\tilde{T}_i) - X_i'\tilde{\theta}|$ and r is the number of non-zero elements in $\hat{\beta}(\lambda)$. Then we minimize the above criteria over a range of values of λ for choosing the best tuning parameter. Let $\hat{\lambda}_A$ denote the resulting selected value. In the section for numerical studies, we will compare the performance of $\hat{\lambda}_A$ and $\hat{\lambda}_L$ in terms of variable selection and parameter estimation, where $\hat{\lambda}_L$ is the selected value when the standard Lasso penalty ($\mathrm{n}\lambda \sum_{j=1}^{p} |\beta_j|$) is used along with the BIC tuning procedure. Initial simulation studies showed that the performance of the BIC-type tuning procedure was superior to that of an AIC-type procedure and a theoretical parameter estimate $\lambda = \frac{\log(n)}{n}$ given by Wang et al (2007).

2.5 Variance Estimation

Next, we propose a bootstrap method to estimate the variance of our estimator. We first take a random sample of size n (with replacement) from the observed data to obtain $(\tilde{T}_i^{(1)}, \delta_i^{(1)}, X_i^{(1)})_{i=1}^n$. Then we compute the Kaplan-Meier estimate $\hat{G}^{(1)}(\cdot)$ based on $(\tilde{T}_i^{(1)}, 1 - \delta_i^{(1)})_{i=1}^n$, and the inverse censoring probability weighted estimator $\tilde{\theta}^{(1)}$ by minimizing

$$\sum_{i=1}^{n} \frac{\delta_i^{(1)}}{\hat{G}^{(1)}(\tilde{T}_i^{(1)})} |\log(\tilde{T}_i^{(1)}) - \theta' X_i^{(1)}|.$$

The first bootstrapped estimate is computing by minimizing

$$\sum_{i=1}^{n} \frac{\delta_i^{(1)}}{\hat{G}^{(1)}(\tilde{T}_i^{(1)})} |\log(\tilde{T}_i^{(1)}) - \theta' X_i^{(1)}| + n\lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j^{(1)}|}.$$

Here for saving the computation cost, we fix λ at the optimal value chosen by the tuning method based on the original data. This procedure is repeated to obtain $\hat{\theta}^{(b)}$, for $b = 2, \ldots, B$ where B is a large number. The bootstrap variance estimate for $\hat{\beta}_j$ is given by

$$\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\beta}_j^{(b)} - \bar{\hat{\beta}}_j)^2,$$

where $\bar{\hat{\beta}}_j = \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}_j^{(b)}$.

Chapter 3

SCMR for Multivariate Failure Time Data

3.1 Multivariate failure time data

Until now, we have assumed that all underlying failure times are independent and uncorrelated. In many practical applications, however, it is common to observe correlated failure time data, as subject can potentially experience multiple events or failures or the event times may be naturally or artificially clustered (Lu, 2005). In either case, there is an underlying dependence structure within each cluster. Lin (1994) gives an example of a situation of recurrent failure time data. This randomized trial, described in more detail by Fleming and Harrington (1991), involved patients with a group of inherited disorders that are characterized by possibly fatal recurrent infections. The event of interest then is the time until an infection develops, so that a patient may experience multiple infections. It is natural to assume that the times to an infection for a given patient are correlated. Yin and Cai (2005) give an example of clustered failure time data from a clinical trial described by Le and Lindgren (1996) involving children with inflammation of the middle ear, otisis media (OM). Bacteria and viruses can enter the middle ear through the Eustachian tube, which is a small tunnel between the middle ear and the eardrum. As a result, fluid may fill the middle ear, which can result in loss of hearing. This can have disastrous effects on speech and language development in young children. One of the most common remedies is to insert a ventilating tube into the infected ears. As long as the tubes stay in place and are

working, they have been shown to reduce OM and therefore improve hearing. In the clinical trial, the effect of a medical treatment on the life of the ventilating tubes is examined. In this case, there is a natural clustering structure in that the paired observations from the two ears of each child are not independent. When there is dependence within clusters of failure time data, as in the univariate case, we can collect information on a large number of covariates, some of which may not have an effect on the underlying survival time. It is then important to select the variables that are most meaningful and discard the rest. With multivariate failure time data, there are also the dependence structures within the clusters along with other parameters. The additional challenge is then developing techniques that allow for these dependencies.

3.2 Estimation for multivariate survival time data

In the case of multivariate or clustered failure time data, consider a study of n clusters with K_i observations in the ith cluster for $i = 1, \ldots, n$, so that there are a total of $N = \sum_{i=1}^{n} K_i$ observations. Let T_{ik} and C_{ik} for $i = 1, \ldots, n; k = 1, \ldots, K_i$ be the latent failure and censoring times, respectively, for the kth subject in the ith cluster, and Z_{ik} be a p-dimensional vector of covariates. As in Yin and Cai (2005), assume that C_{ik} is independent of both T_{ik} and Z_{ik}, and that the censoring times have common survival function $G(\cdot)$. Within each cluster, $\{(T_{ik}, C_{ik}, Z_{ik}), k = 1, \ldots, K_i\}$ may not be independent, but for $T_i = (T_{i1}, \ldots, T_{iK_i})'$, $C_i = (C_{i1}, \ldots, C_{iK_i})'$, and $Z_i = (Z_{i1}, \ldots, Z_{iK_i})'$, $\{(T_i, C_i, Z_i), i = 1, \ldots, n\}$ are independent and identically distributed. In other words, the observations within a cluster may be dependent, but the clusters themselves are i.i.d. While the dependence of observations in a cluster violates the mutual independence assumption of the Cox proportional hazards model and other models, several authors have studied hazard-based regression and shown that estimating equations based the working independence model can yield estimators with desired asymptotic properties. Under the Cox proportional hazards model, the hazard function for T_{ik} is given by

$$\lambda_{ik}(t) = \lambda_0(t) \exp\{\beta_0' Z_{ik}\},\ t > 0,$$

where $\lambda_0(t)$ is unspecified baseline hazard function and β_0' is the true value of the regression ceofficients. If it is assumed that the observations within each cluster are independent, the

log of the partial likelihood is

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{n}\sum_{k=1}^{K_i}\delta_{ik}\beta' Z_{ik} \\ &\quad -\sum_{i=1}^{n}\sum_{m=1}^{K_i}\delta_{im}\log\left\{\sum_{j=1}^{n}\sum_{k=1}^{K_i} I(\tilde{T}_{jk}\geq\tilde{T}_{im})\exp(\beta' Z_{jk})\right\}.\end{aligned}$$

Let $\hat{\beta}$ be the maximizer of $\ell(\beta)$. Lee, Wei, and Amato (1992) show that under some mild conditions, $\hat{\beta}$ is consistent for β_0 even if the observations within a cluster are correlated. Cai, Wei, and Wilcox (2000) note that the Cox models may not fit the data well, and study a class of linear transformation models (of which the proportional hazards model is a special case). Under the Cox model, we have that $\log[-\log\{S(t|Z_{ik})\}] = \lambda_0(t) + \beta_0' Z_{ik}$, where $S(\cdot|Z_{ik})$ is the conditional survival function of T_{ik}. Cai et al (2000) generalize this model by assuming the conditional survival function has the form

$$S(t|Z_{ik}) = g\{\lambda_0(t) + \beta_0' Z_{ik}\},$$

where $g(\cdot)$ is a known, continuous, and strictly decreasing function. They derive estimating equations based on the working independence model assumption that allow for simultaneous estimation of β_0 and $\lambda_0(\cdot)$. Lee, Wei, and Ying (1993) also use the working independence model assumption in their application of the AFT model to clustered censored data. Under the AFT model,

$$\log(T_{ik}) = \beta_0' Z_{ik} + \epsilon_{ik},$$

where $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iK_i})'$ are i.i.d. for $i = 1, \ldots, n$, while the dependence structure of the ϵ's within each cluster is completely unspecified. Lee et al (1993) construct estimating equations based on the assumption that all observations are independent, and show that the resulting estimators can be consistent for β_0.

3.3 SCMR estimator for multivariate failure time data

Censored median regression is attractive for multivariate failure time data for the same reasons as it is for independent data, in that it is robust against outliers and heavy-tailed or heteroscedastic error distributions. In the context of multivariate failure time data, the median regression model, conditional on Z_{ik}, is given by

$$\log(T_{ik}) = \theta_0' X_{ik} + \epsilon_{ik}, \quad i = 1, \ldots, n;\ k = 1, \ldots, K_i,$$

where $X_{ik} = (1, Z'_{ik})', \theta_0 = (\alpha_0, \beta'_0)'$, and the conditional median of ϵ_{ik} is 0. Define $\tilde{T}_{ik} = \min(T_{ik}, C_{ik})$ and $\delta_{ik} = I(T_{ik} \leq C_{ik})$ so that the observed data consists of $\{(\tilde{T}_{ik}, \delta_{ik}, Z_{ik}), i = 1, \ldots, n;\ k = 1, \ldots, K_i\}$. Also suppose that there are a total of $N = \sum_{i=1}^{n} K_i$ observations. We will also use the working independence assumption to derive estimating equations. As in the independent case, we consider the estimating equation of Bang and Tsiatis (2002) and Zhou (2006), the solution of which is the minimizer of

$$\sum_{i=1}^{n} \sum_{k=1}^{K_i} \frac{\delta_{ik}}{\hat{G}(\tilde{T}_{ik})} |\log(\tilde{T}_{ik}) - \theta' X_{ik}|,$$

where $\hat{G}(\cdot)$ is the Kaplan-Meier estimator based on $\{(\tilde{T}_{ik}, 1 - \delta_{ik}), k = 1, \ldots, K_i;\ i = 1, \ldots, n\}$. In the case of independent survival data, it is well known that the Kaplan-Meier estimator is consistent and asymptotically normal. Under some mild conditions, Ying and Wei (1994) show that the Kaplan-Meier estimator is also consistent in the case of dependent data. Denote the solution as $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta}')'$. For variable selection, we will use the adaptive Lasso penalty by minimizing

$$\sum_{i=1}^{n} \sum_{k=1}^{K_i} \frac{\delta_{ik}}{\hat{G}(\tilde{T}_{ik})} |\log(\tilde{T}_{ik}) - \theta' X_{ik}| + N\lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|},$$

where $\lambda > 0$ is the tuning parameter and $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)'$. Denote the solution for any given λ as $\hat{\theta}(\lambda)$.

3.4 Computational algorithms

The solution is obtained in the same way as in the dependent case, with a total of N observations, using an augmented form of the data. For any λ, the solutions are obtained

by minimizing $\sum_{i=1}^{n}\sum_{k=1}^{K_i}|Y_i^* - \theta' X_i^*|$, where

$$(X^*)' = \begin{pmatrix} \frac{\delta_{11}}{\hat{G}(\tilde{T}_{11})} & \frac{\delta_{11}}{\hat{G}(\tilde{T}_{11})}X_{111} & \cdots & \frac{\delta_{11}}{\hat{G}(\tilde{T}_{11})}X_{11p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\delta_{1k_1}}{\hat{G}(\tilde{T}_{1k_1})} & \frac{\delta_{1k_1}}{\hat{G}(\tilde{T}_{1k_1})}X_{1k_11} & \cdots & \frac{\delta_{1k_1}}{\hat{G}(\tilde{T}_{1k_1})}X_{1k_1p} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\delta_{n1}}{\hat{G}(\tilde{T}_{n1})} & \frac{\delta_{n1}}{\hat{G}(\tilde{T}_{n1})}X_{n11} & \cdots & \frac{\delta_{n1}}{\hat{G}(\tilde{T}_{n1})}X_{1np} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\delta_{1k_n}}{\hat{G}(\tilde{T}_{1k_n})} & \frac{\delta_{1k_n}}{\hat{G}(\tilde{T}_{1k_n})}X_{1k_n1} & \cdots & \frac{\delta_{1k_n}}{\hat{G}(\tilde{T}_{1k_n})}X_{1k_np} \\ 0 & \frac{N\lambda}{|\tilde{\beta}_1|} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{N\lambda}{|\tilde{\beta}_p|} \end{pmatrix}_{(N+p)\times(p+1)}$$

and

$$Y^* = \left(\frac{\delta_{11}}{\hat{G}(\tilde{T}_{11})}\log(\tilde{T}_{11}), \cdots, \frac{\delta_{nk_n}}{\hat{G}(\tilde{T}_{nk_n})}\log(\tilde{T}_{nk_n}), 0, \cdots, 0\right)'$$

is a vector of length $N+p$. We can also obtain the solution path using the algorithm of Li and Zhu (2008) by letting $\beta_j^* = \frac{\beta_j}{|\tilde{\beta}_j|}$. Suppose the observed failure times across all clusters are given by $T_{(1)}, \ldots, T_{(m)}$ with corresponding covariate vectors $(Z_{(1)}, \ldots, Z_{(m)})$, so that corresponding censoring indicators $\delta_{(i)}$ are all 1 for $i = 1, \ldots, m$. We can than obtain the solution path by minimizing

$$\sum_{i=1}^{m} b_{(i)}|\log(T_{(i)}) - \alpha - \sum_{j=1}^{p}\beta_j^* Z_{(i)j}|$$

subject to $\sum_{j=1}^{p}|\beta_j^*| \le s$, where $b_{(i)} = \frac{1}{\hat{G}(T_{(i)})}$. For each value of s, the solution given by the algorithm is $\hat{\beta}_j^*(s)$, and we obtain our estimate as $\hat{\beta}(s) = |\tilde{\beta}_j|\hat{\beta}_j^*(s)$.

3.5 Parameter tuning

For parameter tuning, we again use the BIC-type criteria based on uncensored data with i.i.d. double exponential error. For a given λ, define

$$BIC(\lambda) = \frac{2}{\tilde{\sigma}}\sum_{i=1}^{n}\sum_{k=1}^{K_i}\frac{\delta_{ik}}{\hat{G}(\tilde{T}_{ik})}|\log(\tilde{T}_{ik}) - X_{ik}'\hat{\theta}(\lambda)| + \log(N)\cdot r,$$

where $\tilde{\sigma} = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K_i}\frac{\delta_{ik}}{\hat{G}(\tilde{T}_{ik})}|\log(\tilde{T}_{ik}) - X'_{ik}\tilde{\theta}|$ and r is the number of non-zero elements in $\hat{\beta}(\lambda)$. Then we minimize the above criteria over a range of values of λ for choosing the best tuning parameter. Let $\hat{\lambda}_A$ denote the resulting selected value. In the section for numerical studies, we will compare the performance of $\hat{\lambda}_A$ and $\hat{\lambda}_L$ in terms of variable selection and parameter estimation, where $\hat{\lambda}_L$ is the selected value when the standard Lasso penalty ($n\lambda\sum_{j=1}^{p}|\beta_j|$) is used along with the BIC tuning procedure. Initial simulation studies showed that the performance of the BIC-type tuning procedure was superior to that of an AIC-type procedure and a theoretical parameter estimate $\lambda = \frac{\log(N)}{N}$ given by Wang et al (2007).

3.6 Variance estimation

Next, we propose a bootstrap method to estimate the variance of our estimator by a bootstrap resampling method. Instead of resampling individual responses, we resample entire clusters. Since the clusters are assumed to be independent, we will preserve the dependence structure by taking all the observations within a selected cluster. We first take a random sample (with replacement) of clusters of size n from the observed data and use all the observations in each of the selected clusters. We then have clusters $(\tilde{T}_i^{(1)}, \delta_i^{(1)}, X_i^{(1)}; i = 1, \ldots, n)$, where $\tilde{T}_i^{(1)} = (\tilde{T}_{i1}^{(1)}, \ldots, \tilde{T}_{iK_i^{(1)}}^{(1)})'$, $\delta_i^{(1)} = (\delta_{i1}^{(1)}, \ldots, \delta_{iK_i^{(1)}}^{(1)})'$, and $X_i^{(1)} = (X_{i1}^{(1)}, \ldots, X_{iK_i^{(1)}}^{(1)})'$ for $i = 1, \ldots, n$. Then we compute the Kaplan-Meier estimate $\hat{G}^{(1)}(\cdot)$ based on $(\tilde{T}_{ik}^{(1)}, 1 - \delta_{ik}^{(1)}, i = 1, \ldots, n; k = 1, \ldots, K_i^{(1)})$ and the inverse censoring probability weighted estimator $\tilde{\theta}^{(1)}$ by minimizing

$$\sum_{i=1}^{n}\sum_{k=1}^{K_i^{(1)}}\frac{\delta^{(1)}}{\hat{G}^{(1)}(\tilde{T}_{ik}^{(1)})}|\log(\tilde{T}_{ik}^{(1)}) - \theta' X_{ik}^{(1)}|.$$

The first bootstrapped estimate $\hat{\theta}^{(1)}$ is computing by minimizing

$$\sum_{i=1}^{n}\sum_{k=1}^{K_i^{(1)}}\frac{\delta_{ik}^{(1)}}{\hat{G}^{(1)}(\tilde{T}_{ik}^{(1)})}|\log(\tilde{T}_{ik}^{(1)}) - \theta' X_{ik}^{(1)}| + N^{(1)}\lambda\sum_{j=1}^{p}\frac{|\beta_j|}{|\tilde{\beta}_j^{(1)}|},$$

where $N^{(1)} = \sum_{i=1}^{n} K_i^{(1)}$. Here for saving the computation cost, we fix λ at the optimal value chosen by the tuning method based on the original data. This procedure is repeated

to obtain $\hat{\theta}^{(b)}$, for $b = 2, \ldots, B$ where B is a large number. The bootstrap variance estimate for $\hat{\beta}_j$ is given by

$$\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\beta}_j^{(b)} - \bar{\hat{\beta}}_j)^2,$$

where $\bar{\hat{\beta}}_j = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_j^{(b)}$.

Chapter 4

Simulation Studies

4.1 Introduction

In this chapter, we examine the finite sample performance of the proposed SCMR estimator in terms of variable selection and model estimation. In addition, we conduct a series of sensitivity analyses to check the performance of our estimator when the random censoring assumption is violated. Also a case is examined where the error terms have heterogeneous variances. Finally, a series of simulations of multivariate data is conducted. In the univariate case, the failure times are generated from the median regression model $\log(T_i) = \theta_0' X_i + \epsilon_i, \quad i = 1, \cdots, n.$

4.2 IID error distributions

For the first scenario, we consider four error distributions: $t(5)$ distribution, double exponential distribution with scale parameter 1, standard extreme value distribution, and standard logistic distribution. Since the standard extreme value distribution has median $\log(\log(2))$ rather than 0, we subtract its median. We consider eight covariates, which are i.i.d. standard normal random variables. The regression parameter vector $\theta_0 = (\alpha_0, \beta_0)$ is chosen as $\beta_0 = (0.5, 1, 1.5, 2, 0, 0, 0, 0)$ with intercept $\alpha_0 = 1$. The censoring times are generated from a uniform distribution on $(0, c)$, where c is a constant to obtain the desired censoring level. We consider two censoring rates: 20% and 40%. For each censoring rate, we consider samples of sizes 50, 100, and 200. For each setting, we conduct 100 runs of

simulations.

We compare four estimators, namely the full model estimator $\tilde{\beta}$ (called "Full" in the Tables), our SCMR estimators with the adaptive Lasso penalty ("ALasso") and with the Lasso penalty ("Lasso"), and the oracle estimator assuming the true model is known ("Oracle"), with regard to their overall mean absolute deviation for model error MAD ME, point estimation accuracy, and the variable selection performance. Here the MAD ME of an estimator $\hat{\theta}$ is given by

$$MAD\,ME = \frac{1}{n}\sum_{i=1}^{n} |\hat{\theta}' X_i - \theta_0' X_i|.$$

The tuning for the ALasso and Lasso estimates is done by minimizing the BIC-type criteria described in Section 2.5.

The estimation results of the first three non-zero coefficients are summarized in Tables B.1, B.3, B.5, and B.7. The first column for each coefficient reports the average bias, which is given by the average, across 100 runs, of the estimates minus the true parameter value. The second column gives the sample standard deviation (SD) of the 100 estimates. The third column reports the average of 100 bootstrap standard errors. On each run, 500 bootstrap resamples are taken and the standard deviation of the 500 resulting estimators are calculated. The average of these values across 100 runs gives the standard error (SE) in the third column. The estimated standard errors averaged over all rows of the table (Avg. SE) are given in the bottom row of each table. For each average bias, the estimated standard error is given by $SD/\sqrt{100}$. Since the estimators are asymptotically normal, the standard error of each SD can be estimated by $SD/\sqrt{2 \cdot 100}$. For each SE, the standard error is the sample standard deviation of the 100 bootstrap standard errors divided by $\sqrt{100}$. Based on the results in these tables, the biases of all estimates are relatively small particularly when the sample size is large and the censoring proportion is low. In addition, the estimated standard errors obtained using the proposed bootstrap method are reasonably close to the sample standard deviations in all scenarios. The variable selection results of the different estimators are summarized in Tables B.2, B.4, B.6, and B.8. We compare the MAD ME, the selection frequency of each of the first six variables, the frequency of selecting the exact true model, and the mean number of incorrect and correct zeros selected over 100 simulation runs. The standard error averaged over all rows of the table for MAD ME is given in the bottom row in each table. Since each MAD ME is an average across 100 runs,

the standard error is the sample standard deviation of the 100 values divided by $\sqrt{100}$. Overall, the MAD ME of the Alasso is smaller than those of the Lasso and the Full in all the settings, and is also close to the oracle when the sample size is large and the censoring proportion is low. With regard to variable selection, compared with the Full and the Lasso, the Alasso produces more sparse models and selects the exact true model more frequently. For example, when $n = 200$, censoring rate is 20% and the error distribution is $t(5)$, the frequencies of selecting true model (SF) are: Oracle (100), Alasso (65), Lasso (19), Full (0); the mean numbers of correct/incorrect zeros selected (Cor./Inc.) are: Oracle (4/0), Alasso (3.62/0.01), Lasso (2.52/0.01), Full (0/0).

4.3 Sensivity analyses

Next, we conduct a series of sensitivity analyses to check the performance of our estimator when the random censoring assumption is violated. More specifically, the censoring times are now generated as $C_i = e^{\gamma X_{1i}} C_i^*, i = 1, \cdots, n$, where C_i^*'s are from a uniform distribution on $(0, c)$ as before. We consider two values of γ: $\gamma = 0.1$ or 0.2, two censoring rates: 20% and 40%, and three sample sizes: $n = 50, 100$, or 200. The error terms are i.i.d. from a $t(5)$ distribution. Other settings remain the same as before. The estimation and variable selection results are summarized in Tables B.9-B.12. Based on these results, we observe similar findings as before. The Alasso shows much better performance in terms of variable selection compared with the Lasso and the Full, although all the methods may produce certain biases in point estimation as expected. The estimation and selection results are very similar for both values of γ. These findings suggest that our method may be robust against the random censoring assumption violation.

4.4 Heteroscedastic errors

We now examine the case of heteroscedastic errors. Now the responses are generated from the model

$$\log(T_i) = \alpha_0 + \beta_0' Z_i + e^{\gamma Z_{4i}} \epsilon_i.$$

where the ϵ's are i.i.d. from a $t(5)$ distribution. We consider two values for γ: $\gamma = 0.1$ or 0.2, two censoring rates: 20% and 40%, and three sample sizes: $n = 50, 100$, or 200. The

censoring times are generated from a uniform distribution on $(0, c)$, where c is a constant to obtain the desired censoring level. Other settings remain the same as before. The results for estimation are summarized in Tables B.13-B.16. We observe the same patterns as in the i.i.d. case and sensitivity analyses. The biases are relatively small, especially for large sample sizes. While the bootstrap standard error estimates are larger than the sample standard deviations, these differences become smaller as n increases. One difference is that while the MAD ME for the ALasso is almost always smaller than that of the Lasso, it is slightly larger than the Full model MAD ME when the censoring proportion is 40%. However, it becomes closer to the MAD ME of the Oracle estimator as the sample size gets larger. The ALasso selects the true model more frequently than the Lasso and produces more sparse models. The selection results are similar for both $\gamma = 0.1$ and 0.2.

4.5 Multivariate data

Next, we examine the case of multivariate, or clustered, failure time data. The failure times are generated by the model $\log(T_{ik}) = \alpha_0 + \beta_0' X_{ik} + \epsilon_{ik}$, for $i = 1, \ldots, n$ and $k = 1, \ldots K$, where $\alpha_0 = 1$, $\beta_0 = (.5, 1, 1.5, 2, 0, 0, 0, 0)$. Essentially, there are n clusters each of size K, for a total of $N = nK$ failure times. Within cluster i, for $i = 1, \ldots, n$, the errors $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iK})'$ are generated from a K-variate normal distribution with mean vector 0 and variance-covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

The error vectors for each cluster are generated independently. The covariates X_{ik} for each failure time are multivariate standard normal. The N censoring times are i.i.d. Uniform$(0, c)$, where c is chosen to achieve the desired censoring proportion. Cluster sizes $K = 2$ and 5 are considered, along with 20% and 40% censoring, $\rho = .2$ and .5, and sample sizes $n = 50, 100$, and 200. For each combination, there are 100 runs, with 500 bootstrap resamples for each run. Resampling is done on the clusters, so that all observations in a

resampled cluster are used. The mean absolute deviation for model error is given by

$$MAD\,ME = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{K}|\hat{\theta}'X_{ik} - \theta_0'X_{ik}|.$$

In order to tune λ for the ALasso and Lasso estimates, we minimize the BIC-type criteria given in Section 3.5. The estimation results of the first three non-zero coefficients are summarized in Tables B.17, B.19, B.21, and B.23. The first column for each coefficient reports the average bias, which is given by the average, across 100 runs, of the estimates minus the true parameter value. The second column gives the sample standard deviation (SD) of the 100 estimates. The third column reports the average of 100 bootstrap standard errors. On each run, 500 bootstrap resamples are taken and the standard deviation of the 500 resulting estimators are calculated. The average of these values across 100 runs gives the standard error (SE) in the third column. The estimated standard errors averaged over all rows of the table (Avg. SE) are given in the bottom row of each table. For each average bias, the estimated standard error is given by $SD/\sqrt{100}$. Since the estimators are asymptotically normal, the standard error of each SD can be estimated by $SD/\sqrt{2 \cdot 100}$. For each SE, the standard error is the sample standard deviation of the 100 bootstrap standard errors divided by $\sqrt{100}$. The variable selection results of different estimates are summarized in Tables B.18, B.20, B.22, and B.24. We compare the MAD ME, the selection frequency of each of the first six variables, the frequency of selecting the exact true model, and the mean number of incorrect and correct zeros selected over 100 simulation runs. The standard error averaged over all rows of the table for MAD ME is given in the bottom row in each table. Since each MAD ME is an average across 100 runs, the standard error is the sample standard deviation of the 100 values divided by $\sqrt{100}$. Overall, the findings are similar to that of the univariate (independent) cases. The ALasso outperforms the Lasso in terms of estimation accuracy and variable selection, and its performance becomes more similar to that of the Oracle as the sample size increases.

Being that our estimating equations and tuning procedure are developed under the assumption of independence, it is of major interest to compare the results for different values of the intracluster variance $\rho = 0.2$ and $\rho = 0.5$. The biases for the non-zero coefficients seem to be very similar for both values of ρ when the sample size, cluster size, and censoring proportion are the same. The same holds true for the MAD ME as well as the selection frequency.

Chapter 5

Data Analyses

5.1 PBC data

The primary biliary cirrhosis (PBC) data was collected at the Mayo clinic between 1974 and 1984. This data is given in Therneau and Grambsch (2000). In this study, 312 patients from a total of 424 patients who agreed to participate in the randomized trial are eligible for the analysis. Of those, 125 patients died before the end of follow-up. We study the dependence of the survival time on the following selected covariates: (1) continuous variables: age (in years), alb (albumin in g/dl), alk (alkaline phosphatase in U/liter), bil (serum bilirunbin in mg/dl), chol (serum cholesterol in mg/dl), cop (urine copper in μg/day), plat (platelets per cubic ml/1000), prot (prothrombin time in seconds), sgot (liver enzyme in U/ml), trig (triglycerides in mg/dl); (2) categorical variables: asc (0, absence of ascites; 1, presence of ascites), ede (0 no edema; 0.5 untreated or successfully treated; 1 unsuccessfully treated edema), hep (0, absence of hepatomegaly; 1, presence of hepatomegaly), sex (0 male; 1 female), spid (0, absence of spiders; 1, presence of spiders), stage (histological stage of disease, graded 1, 2, 3 or 4), trt (1 control, 2 treatment).

The PBC data has been previously analyzed by a number of authors using various estimation and variable selection methods. For example, Tibshirani (1997) fitted the proportional hazards model with the stepwise selection and with the LASSO penalty. Zhang and Lu (2007) further studied the PBC data using the penalized partial likelihood estimation method with the SCAD and the adaptive Lasso penalty. Here, we apply the proposed SCMR method to the PBC data. As in Tibshirani (1997) and Zhang and Lu (2007), we

restrict our attention to the 276 observations without missing values. Among these 276 patients, there are 111 deaths, about 60% of censoring. Table 5.1 summarizes the estimated coefficients and the standard errors based on 2000 bootstrap resamples for the Full, the Lasso and the Alasso. We found that the Alasso selects 9 variables: *age, asc, oed, bil, alb, cop, alk, plat* and *prot* and the Lasso selects 13 variables which contain the 9 variables selected by the Alasso. Moreover, the 9 variables selected using the proposed Alasso method in censored median regression shared 6 variables out of 8 selected by the penalized partial likelihood estimation method with the adaptive Lasso penalty of Zhang and Lu (2007) in the proportional hazards model. We also plot the solution path of our SCMR estimator with the adaptive Lasso penalty in Figure 5.1 based on a modified algorithm of Li and Zhu (2008).

5.2 DLBCL microarray data

We also apply the SCMR method to the high dimensional microarray gene expression data of Rosenwald et al. (2002). The data consists of survival times of 240 diffuse large B-cell lymphoma (DLBCL) patients, and the expressions of $7,399$ genes for each patient. Among them, 138 patients died during the follow-up method. The main goals of the study are to identify the important genes that can predict patients's survival and to study their effects on survival. This data was analyzed by Li and Luan (2005). To handle such high dimensional data, a common practice is to first conduct a preliminary gene filtering based on some univariate analysis, and then apply a more sophisticated model-based analysis. Following Li and Luan (2005), we concentrate on the top 50 genes selected using the univariate log-rank test. To evaluate the performance of the proposed SCMR method, the data are randomly divided into two sets: the training set (160 patients) and the testing set (80 patients).

The SCMR estimator with the adaptive Lasso penalty is then computed based on the training data and the proposed BIC method is used for parameter tuning. Our SCMR method selects totally 25 genes. To evaluate the prediction performance of the resulting SCMR estimator built with the training set, we plot, in Figure 5.2, the Kaplan-Meier estimates of survival functions for the high-risk and low-risk groups of patients, defined by the estimated conditional medians of failure times. The cut-off value was determined by

the median failure time of the baseline group from the training set, and the same cutoff was applied to the testing data. It is seen that the separation of the two-risk groups is reasonably good in both the training and the testing data, suggesting a satisfactory prediction performance of the fitted survival model. The log-rank test of differences between two survival curves gives p-values of 0 and 0.031 for the training and testing data, respectively.

Table 5.1: Estimation and variable selection for PBC data with censored median regression.

	Full	ALasso	Lasso
Intercept	7.62 (0.44)	7.72 (0.40)	7.83 (0.45)
trt	0.04 (0.14)	0 (-)	0 (-)
age	-3.29 (1.50)	-2.77 (1.36)	-0.93 (0.94)
sex	0.03 (0.28)	0 (-)	0.04 (0.24)
asc	-0.57 (0.83)	-0.31 (0.83)	-0.31 (0.81)
hep	-0.05 (0.17)	0 (-)	0.02 (0.17)
spid	-0.09 (0.20)	0 (-)	0 (-)
oed	-0.75 (0.63)	-0.70 (0.61)	-0.70 (0.65)
bil	-1.71 (3.24)	-2.09 (3.25)	-1.56 (2.43)
chol	-0.87 (3.50)	0 (-)	-0.64 (1.23)
alb	2.96 (1.33)	3.16 (1.28)	3.54 (1.17)
cop	-4.00 (1.97)	-3.89 (1.87)	-2.61 (1.57)
alk	2.16 (1.02)	2.19 (0.97)	2.11 (0.87)
sgot	-0.20 (1.84)	0 (-)	0 (-)
trig	1.16 (2.11)	0 (-)	0 (-)
plat	-1.61 (1.11)	-1.25 (0.90)	-1.12 (0.67)
prot	2.58 (2.31)	1.62 (2.23)	0.93 (1.85)
stage	0.03 (0.10)	0 (-)	-0.05 (0.09)

Standard errors given in parentheses are based on 2000 bootstrap resamples.

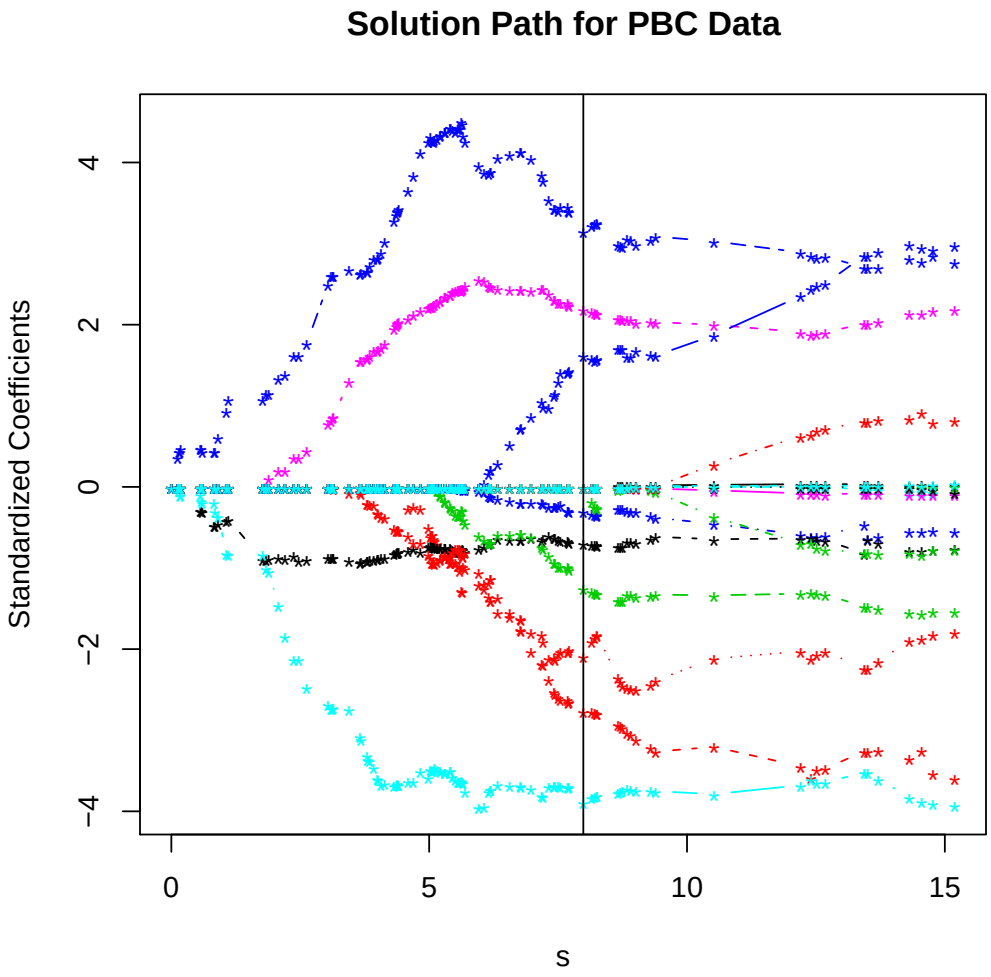

Figure 5.1: The solution path of our SCMR estimator with the adaptive Lasso penalty for PBC data. The solid vertical line denotes the resulting estimator tuned with the proposed BIC criterion.

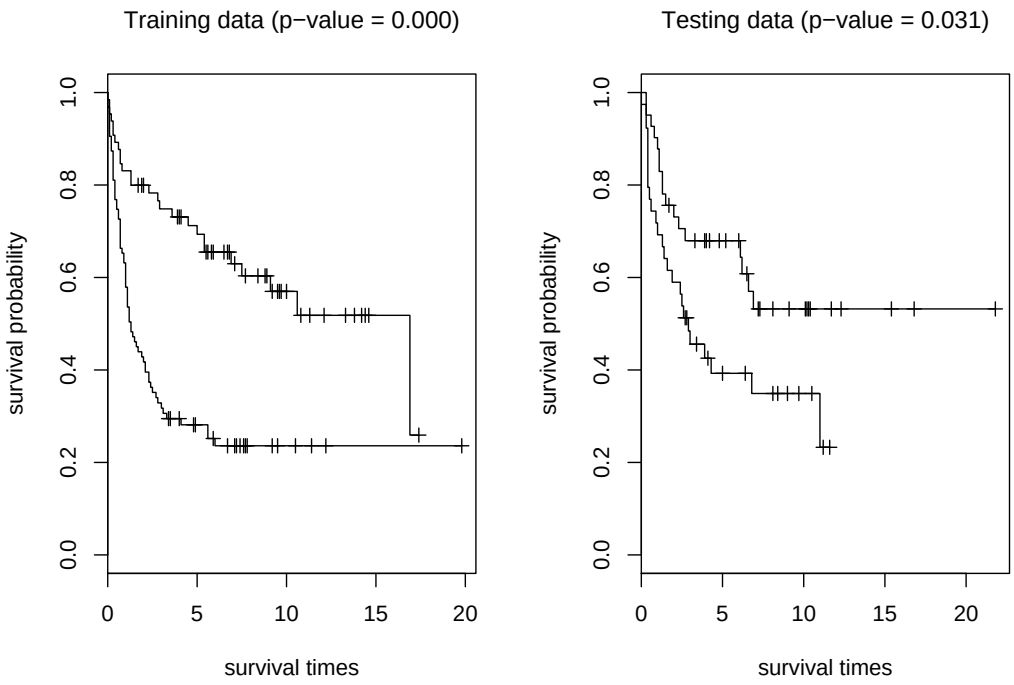

Figure 5.2: Kaplan-Meier estimates of survival curves for high-risk and low-risk groups of patients using the selected genes by the SCMR.

Bibliography

[1] Akaike, H. (1973). Maximum likelihood indentification of Gaussian autoregressive moving average models, *Biometrika*, **60**: 255-265.

[2] Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study, *The Annals of Statistics*, **10**: 1100-1120.

[3] Bang, H. and Tsiatis, A. (2002). Median regression with censored cost data, *Biometrics*, **58**: 643-649.

[4] Bennett, S. (1983a). Analysis of survival data by the proportional odds model, *Statistics in Medicine*, **2**: 273-277.

[5] Bennett, S. (1983b). Log-logistic regression models for survival data, *Applied Statistics*, **32**: 165-171.

[6] Boos, D., Stefanski, L., and Wu, Y. (2008). Fast FSR variable selection with applications to clinical trials, to appear in *Biometrics*.

[7] Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24**: 2350-2383.

[8] Buckley, J. and James, I. (1979). Linear regression with censored data, *Biometrika*, **66**: 429-436.

[9] Cai, T., Wei, L., and Wilcox, M. (2000). Semiparametric regression analysis for clustered failure time data, *Biometrika*, **87**: 867-878.

[10] Cai, T., Huang, J., and Tian, L. (2008). Regularized estimation for the accelerated failure time model, to appear in *Biometrics*.

[11] Cox, D. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society*, **34**: 187-220.

[12] Cox, D. (1975). Partial likelihood, *Biometrika*, **62**: 269-276.

[13] Cox, D. and Oakes (1984). *Analysis of Survival Data*, London: Chapman and Hall.

[14] Davis, R., Knight, K., and Liu, J. (1992). M-Estimation for autoregressions with infinite variance, *Stochastic Process and Their Applications*, **40**: 145-180.

[15] Dickson, E., Grambsch, P., Fleming, T., Fisher, L., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: model for decision making, *Hepatology*, **10**: 1-7.

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

[17] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model, *The Annals of Statistics*, **30**: 74-99.

[18] Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*, New York: Wiley.

[19] Hocking, R. (1976). The analysis and selection of variables in linear regressio, *Biometrics*, **32**: 1-49.

[20] Jin, Z., Lin, D., Wei, L.J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model, *Biometrika*, **90**: 341-353.

[21] Johnson, B. (2008). Variable selection in semi-parametric linear regression with censored data, *Journal of the Royal Statistical Society (B)*, **70**: 351-370.

[22] Johnson, B., Lin, D., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models, *Journal of the American Statistical Association*, **103**: 672-680.

[23] Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data, 2nd ed.*, Hoboken, NJ: Wiley.

[24] Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**: 457-481.

[25] Klein, J. and Moeschberger (2003). *Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.*, New York: Springer.

[26] Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions, *The Annals of Statistics*, **26**: 755-770.

[27] Koenker, R. and Geling, L. (2001). Reappraising medfly longevity: A quantile regression survival analysis, *Journal of the American Statistical Society*, **96**: 458-468.

[28] Koenker, R. and Zhao, Q. (1996). Conditional quantile estimation and inference for ARCH models, *Econometric Theory*, **12**: 793-813.

[29] Koenker, R. and D'Orey, V. (1987). Computing regression quantiles, *Applied Statistics*, **36**: 383-393.

[30] Lam, K. and Leung, T. (2001). Marginal likelihood estimation for proportional odds model with right censored data, *Lifetime Data Analysis*, **7**: 39-54.

[31] Le, C. and Lindgren, B. (1996). Duration of ventilating tubes: a test for comparing two clustered samples of censored data, *Biometrics*, **52**: 328-334.

[32] Lee, E., Wei, L., and Amato, D. (1992). Cox-type regression analysis for large numbers of small groups correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds), 237-247. Dordrecht: Kluwer Academic Publishers.

[33] Lee, E., Wei, L., and Ying, Z. (1993). Linear regression analysis for highly stratefied failure time data, *Journal of the American Statistical Association*, **88**: 557-565.

[34] Li, H. and Luan, Y. (2005). Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data, *Bioinformatics*, **21**: 2403-2409.

[35] Li, Y. and Zhu, J. (2008). L1-norm quantile regression, *Journal of Computational and Graphical Statistics*, **17**: 163-185.

[36] Lin, D. (1994). Cox regression analysis of multivariate failure time data: the marginal approach, *Statistics in Medicine*, **13**: 2233-2247.

[37] Lu, W. (2005). Marginal regression of multivariate event times based on linear transformation models, *Lifetime Data Analysis*, **11**: 389-404.

[38] Lu, W. and Zhang, H. (2007). Variable selection for proportional odds model, *Statistics in Medicine*, **26**: 3771-3781.

[39] Luo, X., Stefanski, L., and Boos, D. (2006). Tuning variable selection procedures by adding noise, *Technometrics*, **48**: 165-175.

[40] Mallows, C. (1973). Some comments on C_p, *Technometrics*, **15**: 661-675.

[41] Montgomery, D. and Peck, E. (1991). *Introduction to Linear Regression Analysis, 2nd ed.*, New York: Wiley.

[42] Pettitt, A. (1984). Proportional odds model for survival data and estimates using ranks, *Applied Statistics*, **33**: 169-175.

[43] Portnoy, S. (2003). Censored regression quantiles, *Journal of the American Statistical Association*, **98**: 1001-1012.

[44] Prentice, R.L. (1978). Linear rank tests with right censored data, *Biometrika*, **65**: 167-179.

[45] Reid, N. (1994). A conversation with Sir David Cox, *Statistical Science*, **9**: 439-455.

[46] Ritov, Y. (1990). Estimation in a linear regression model with censored data, *Annals of Statistics*, **18**: 303-328.

[47] Rosenwald, A., Wright, G., Chan, W., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma, *New England Journal of Medicine*, **346**: 1937-1947.

[48] Scharfstein, D., Tsiatis, A., and Gilbert, P. (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data, *Lifetime Data Analysis*, **4**: 355-391.

[49] Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**: 461-464.

[50] Therneau, T., and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*, New York: Springer-Verlag Inc.

[51] Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model, *Statistics in Medicine*, **16**: 385-395.

[52] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society*, Ser. B, **58**: 267-288.

[53] Tsiatis, A. (1990). Estimating regression parameters using linear rank tests for censored data, *Annals of Statistics*, **18**: 354-372.

[54] Tsiatis, A. (1981). A large sample study of Cox's regression model, *The Annals of Statistics*, **9**: 93-108.

[55] Wang, H., Li., G, and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso, *Journal of Business and Economic Statistics*, **25**: 347-355.

[56] Wu, Y., Boos, D., and Stefanski, L. (2007). Controlling variable selection by the addition of pseudovariables, *Journal of the American Statistical Association*, **102**: 235-243.

[57] Yin, G. and Cai, J. (2005). Quantile regression models with multivariate failure time data, *Biometrics*, **61**: 151-161.

[58] Ying, Z., Jung, S., and Wei, L. (1995). Survival analysis with median regression models, *Journal of the American Statistical Association*, **90**: 178-184.

[59] Ying, Z. and Wei, L. (1994). The Kaplan-Meier estimate for dependent failure time observations, *Journal of Multivariate Analysis*, **50**: 17-29.

[60] Ying, Z. (1993). A large sample study of rank estimation for censored regression data, *Annals of Statistics*, **21**: 76-99.

[61] Zeng, D. and Lin, D. (2007). Maximum likelihood estimation in semiparametric regression models with censored data, *Journal of the Royal Statistical Society, Series B*, **69**: 1-30.

[62] Zhang, H., and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model, *Biometrika*, **94**: 691-703.

[63] Zhou, L. (2006). A simple censored median regression estimator, *Statistica Sinica*, **16**: 1043-1058.

[64] Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**: 1418-1429.

[65] Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model, *Biometrika*, **95**: 241-247.

Appendices

Appendix A: Proofs of theorems

To prove the asymptotic results established in Theorems 1 and 2, we need the following regularity conditions:

1. The error term ϵ has a continuous conditional density function $f(\cdot|Z=z)$ such that $f(0|Z=z) \geq b_0 > 0$, $|\dot{f}(0|Z=z)| \leq B_0$ and $\sup_s f(s|Z=z) \leq B_0$ for all possible values z of Z, where (b_0, B_0) are two positive constants and $\dot{f}$ is the derivative of f.

2. The covariate vector Z are of compact support and the parameter β_0 belongs to the interior of a known compact set $\mathcal{B}_0$.

3. $P(t \leq T \leq C) \geq \zeta_0 > 0$ for any $t \in [0, \tau]$, where τ is the maximum follow-up and ζ_0 is a positive constant.

Proof of Theorem 1: To establish the result given in Theorem 1, it is equivalent to show that for any $\eta > 0$, there is a constant M such that $P(\sqrt{n}||\hat{\theta} - \theta|| \leq M) \geq 1 - \eta$. Let $u = (u_0, u_1, \ldots, u_p)' \in \Re^{p+1}$, and $A_M = \{\theta_0 + \frac{u}{\sqrt{n}} : ||u|| \leq M\}$ be the ball in $\Re^{p+1}$ centered at θ_0 with radius $\frac{M}{\sqrt{n}}$. If, for any $\eta > 0$, we can choose M large enough so that $P(\hat{\theta} \in A_M) \geq 1 - \eta$, then the result is proved. Define

$$Q(G, \theta) = \sum_{i=1}^{n} \frac{\delta_i}{\hat{G}(\tilde{T}_i)} |\log(\tilde{T}_i) - \theta' X_i| + n\lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|}.$$

$Q(\cdot)$ is a convex function in θ, so M satisfying

$$P\left\{ inf_{||u||=M} Q(\hat{G}, \theta_0 + \frac{u}{\sqrt{n}}) > Q(\hat{G}, \theta_0) \right\} \geq 1 - \eta$$

will be sufficient to show that $\hat{\theta}$ is $\sqrt{n}$-consistent. This condition implies that

$$P\left\{ inf_{||u||\geq M} Q(\hat{G}, \theta_0 + \frac{u}{\sqrt{n}}) > Q(\hat{G}, \hat{\theta}) \right\} \geq 1 - \eta,$$

which in turn implies that $P(\hat{\theta} \in A_M) \geq 1 - \eta$. This is because $\hat{\theta}$ minimizes $Q(\cdot)$, so $Q(\hat{G}, \hat{\theta}) \leq Q(\hat{G}, w)$ for any $w \in \Re^{p+1}$. Since $Q(\cdot)$ is convex, $Q(\hat{G}, w) \to \infty$ as $||w|| \to \infty$. This leads to

$$\begin{aligned} P\left\{inf_{||u||=M} Q(\hat{G}, \theta_0 + \tfrac{u}{\sqrt{n}}) > Q(\hat{G}, \theta_0)\right\} &= P\left\{inf_{||u||\geq M} Q(\hat{G}, \theta_0 + \tfrac{u}{\sqrt{n}}) > Q(\hat{G}, \theta_0)\right\} \\ &\quad \text{since } Q(\cdot) \text{ is convex} \\ &\leq P\left\{inf_{||u||\geq M} Q(\hat{G}, \theta_0 + \tfrac{u}{\sqrt{n}}) > Q(\hat{G}, \hat{\theta})\right\} \\ &\quad \text{since } Q(\hat{\theta}) \leq Q(\theta_0). \end{aligned}$$

So if

$$P\left\{inf_{||u||=M} Q(\hat{G}, \theta_0 + \frac{u}{\sqrt{n}}) > Q(\hat{G}, \theta_0)\right\} \geq 1 - \eta,$$

then

$$P\left\{inf_{||u||\geq M} Q(\hat{G}, \theta_0 + \frac{u}{\sqrt{n}}) > Q(\hat{G}, \hat{\theta})\right\} \geq 1 - \eta.$$

Let $D_n(u) = Q(\hat{G}, \theta_0 + \frac{u}{\sqrt{n}}) - Q(\hat{G}, \theta_0)$, which can be written as

$$\begin{aligned} D_n(u) &= \{Q(G_0, \theta_0 + \tfrac{u}{\sqrt{n}}) - Q\left(G_0, \theta_0\right)\} \\ &+ \{Q(\hat{G}, \theta_0 + \tfrac{u}{\sqrt{n}}) - Q(G_0, \theta_0 + \tfrac{u}{\sqrt{n}})\} \\ &- \{Q(\hat{G}, \theta_0) - Q(G_0, \theta_0)\}, \end{aligned}$$

where $G_0(\cdot)$ is the true survival function of the censoring time.

For the first term in $D_n(u)$, we have that

$$\begin{aligned}
Q(G_0, \theta_0 + \tfrac{u}{\sqrt{n}}) - Q\left(G_0, \theta_0\right) &= \textstyle\sum_{i=1}^{n} \frac{\delta_i}{G_0(\tilde{T}_i)} \left(|\epsilon_i - X_i' \frac{u}{\sqrt{n}}| - |\epsilon_i| \right) \\
&+ n\lambda \textstyle\sum_{j=1}^{p} \frac{|\beta_{j0} + u_j/\sqrt{n}| - |\beta_{j0}|}{|\tilde{\beta}_j|} \\
&= \textstyle\sum_{i=1}^{n} \frac{\delta_i}{G_0(\tilde{T}_i)} \left(|\epsilon_i - X_i' \frac{u}{\sqrt{n}}| - |\epsilon_i| \right) \\
&+ n\lambda \textstyle\sum_{j=1}^{q} \frac{|\beta_{j0} + u_j/\sqrt{n}| - |\beta_{j0}|}{|\tilde{\beta}_j|} + n\lambda \sum_{j=q+1}^{p} \frac{|u_j/\sqrt{n}|}{|\tilde{\beta}_j|} \\
&\quad \text{since } \beta_j = 0 \text{ for } j > q. \\
&\geq \textstyle\sum_{i=1}^{n} \frac{\delta_i}{G_0(\tilde{T}_i)} \left(\epsilon_i - X_i' \frac{u}{\sqrt{n}}| - |\epsilon_i| \right) \\
&+ n\lambda \textstyle\sum_{j=1}^{q} \frac{|\beta_{j0} + u_j/\sqrt{n}| - |\beta_{j0}|}{|\tilde{\beta}_j|} \\
&\quad \text{since } n\lambda \textstyle\sum_{j=q+1}^{p} \frac{|u_j/\sqrt{n}|}{|\tilde{\beta}_j|} \geq 0. \\
&\geq \textstyle\sum_{i=1}^{n} \frac{\delta_i}{G_0(\tilde{T}_i)} \left(|\epsilon_i - X_i' \frac{u}{\sqrt{n}}| - |\epsilon_i| \right) \\
&- n\lambda \textstyle\sum_{j=1}^{q} \frac{|u_j|/\sqrt{n}}{|\tilde{\beta}_j|} \\
&= \textstyle\sum_{i=1}^{n} \frac{\delta_i}{G_0(\tilde{T}_i)} \left(|\epsilon_i - X_i' \frac{u}{\sqrt{n}}| - |\epsilon_i| \right) \\
&\quad -\sqrt{n}\lambda \textstyle\sum_{j=1}^{q} \frac{|u_j|}{|\tilde{\beta}_j|} \\
&= L_n(u) - \sqrt{n}\lambda \textstyle\sum_{j=1}^{q} \frac{|u_j|}{|\tilde{\beta}_j|} \\
&\geq L_n(u) - \kappa_1 O_p(||u||),
\end{aligned}$$

where κ_1 is a positive finite constant. The last inequality in the above expression is because $\sqrt{n}\lambda = O_p(1)$, and $\tilde{\beta}_j$, $j = 1, \ldots, q$, converges to β_{j0} that is bounded away from zero.

By making use of the result from Knight (1998), that for $|x| \neq 0$,

$$|x - y| - |x| = -y\left[I(x > 0) - I(x < 0)\right] + 2\int_0^y \left[I(x \leq s) - I(x \leq 0)\right] ds,$$

we let $x = \epsilon_i$ and $y = \frac{u'X_i}{\sqrt{n}}$, to obtain

$$\begin{aligned}
L_n(u) &= \tfrac{u'}{\sqrt{n}} \textstyle\sum_{i=1}^{n} \frac{\delta_i}{G_0(\tilde{T}_i)} X_i \left\{ I(\epsilon_i > 0) - I(\epsilon_i < 0) \right\} \\
&+ 2\textstyle\sum_{i=1}^{n} \frac{\delta_i}{G_0(\tilde{T}_i)} \int_0^{u'X_i/\sqrt{n}} \left[I(\epsilon_i \leq s) - I(\epsilon_i \leq 0) \right] ds \\
&= \tfrac{u'}{\sqrt{n}} \textstyle\sum_{i=1}^{n} \frac{\delta_i}{G_0(\tilde{T}_i)} X_i \, sign(\epsilon_i) + 2\textstyle\sum_{i=1}^{n} \frac{\delta_i}{G_0(\tilde{T}_i)} \int_0^{u'X_i/\sqrt{n}} \left[I(\epsilon_i \leq s) - I(\epsilon_i \leq 0) \right] ds
\end{aligned}$$

We will show that $\frac{u'}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{G_0(\tilde{T}_i)}X_i\,sign(\epsilon_i)$ converges in distribution, and that $2\sum_{i=1}^{n}\frac{\delta_i}{G_0(\tilde{T}_i)}\int_0^{u'X_i/\sqrt{n}}[I(\epsilon_i\le s)-I(\epsilon_i\le 0)]\,ds$ converges to a constant. To begin, we have

$$\begin{aligned}
E\left[\frac{\delta_i}{G_0(\tilde{T}_i)}X_i\,sign(\epsilon_i)\right] &= E_{X,T}\left[E\left\{\frac{\delta_i}{G_0(\tilde{T}_i)}X_i\,sign(\epsilon_i)\right\}|X_i,T_i\right]\\
&= E_{X,T}\left[\frac{X_i}{G_0(\tilde{T}_i)}E\left\{\delta_i\,sign(\epsilon_i)|X_i,T_i\right\}\right]\\
&= E\left[\frac{X_i}{G_0(\tilde{T}_i)}E\left\{\delta_i|X_i,T_i\right\}E\left\{sign(\epsilon_i)|X_i,T_i\right\}\right]\\
&\quad \text{since } \delta_i \text{ and } \epsilon_i \text{ are independent given } X_i,T_i\\
&= E_{X,T}\left[\frac{X_i}{G_0(\tilde{T}_i)}E\{\delta_i|X_i,T_i\}\cdot 0\right]=0,
\end{aligned}$$

since ϵ_i has median 0. So we know that $E\left[\frac{u'}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{G_0(\tilde{T}_i)}X_i\,sign(\epsilon_i)\right]=0$. To find its variance, we get

$$\begin{aligned}
Var\left[\frac{\delta_i}{G_0(\tilde{T}_i)}X_i\,sign(\epsilon_i)\right] &= E\left[\frac{\delta_i^2}{G_0^2(\tilde{T}_i)}X_iX_i'\left\{sign(\epsilon_i)\right\}^2\right]-0\\
&= E\left[\frac{X_iX_i'}{G_0^2(\tilde{T}_i)}E\left\{\delta_i|X_i,T_i\right\}\right]\\
&= E\left[\frac{X_iX_i'}{G_0(\tilde{T}_i)}\right].
\end{aligned}$$

By the Central Limit Theorem, $\frac{u'}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{G_0(\tilde{T}_i)}X_i\,sign(\epsilon_i)$ converges in distribution to $u'W_1$, where W_1 is a $(p+1)$-dimensional Normal with mean 0 and covariance matrix $\Sigma_1=\left[\frac{X_1X_1'}{G_0(\tilde{T}_i)}\right]$. It implies that the first term in $L_n(u)$ can be written as $O_p(||u||)$.

For the second term in $L_n(u)$, let $A_{ni}(u)=\frac{\delta_i}{G_0(\tilde{T}_i)}\int_0^{u'X_i/\sqrt{n}}[I(\epsilon_i\le s)-I(\epsilon_i\le 0)]\,ds$. We will show that $\sum_{i=1}^{n}A_{ni}(u)$ converges in probability to a quadratic function of u. More specifically, for any $\psi>0$, write

$$A_{ni}^2(u)=A_{ni}^2(u)I\left(\frac{|u'X_i|}{\sqrt{n}}\ge\psi\right)+A_{ni}^2(u)I\left(\frac{|u'X_i|}{\sqrt{n}}<\psi\right).$$

The term $\sum_{i=1}^{n}A_{ni}(u)$ will therefore dominate $\frac{u'}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{G_0(\tilde{T}_i)}X_i\,sign(\epsilon_i)$ and $\sqrt{n}\lambda\sum_{j=1}^{q}\frac{|u_j|}{|\tilde{\beta}_j|}$ uniformly in M, since both are only linear functions of u. We must first show that

$$Var\left[\sum_{i=1}^{n}A_{ni}(u)\right]=\sum_{i=1}^{n}Var[A_{ni}(u)]\le nE[A_{ni}^2(u)]\to 0,$$

so that

$$\sum_{i=1}^{n}\left[A_{ni}(u)-E\{A_{ni}(u)\}\right]=o_p(1).$$

First, we have that

$$
\begin{aligned}
& nE\left\{A_{ni}^2(u)\cdot I\left(|u'X_i|\geq\psi\sqrt{n}\right)\right\} \\
&= nE\left\{\frac{\delta_i}{G_0(\tilde{T}_i)}\int_0^{\frac{u'X_i}{\sqrt{n}}}\left[I(\epsilon_i\leq s)-I(\epsilon\leq 0)\right]ds\cdot I\left(|u'X_i|\geq\psi\sqrt{n}\right)\right\}^2 \\
&= E\left\{\frac{\delta_i}{G_0^2(\tilde{T}_i)}\left(\int_0^{\frac{u'X_i}{\sqrt{n}}}\left[I(\epsilon_i\leq s)-I(\epsilon\leq 0)\right]ds\right)^2\cdot I\left(|u'X_i|\geq\psi\sqrt{n}\right)\right\} \\
&\leq nE_{X,T}\left\{\frac{\delta_i}{G_0^2(\tilde{T}_i)}\left(\int_0^{u'X_i/\sqrt{n}}2\,ds\right)^2\cdot I\left(|u'X_i|\geq\psi\sqrt{n}\right)\right\} \\
&= 4nE_{X,T}\left\{\frac{\delta_i}{G_0^2(\tilde{T}_i)}\left(\frac{|u'X_i|}{\sqrt{n}}\right)^2\cdot I\left(|u'X_i|\geq\psi\sqrt{n}\right)\right\} \\
&= 4E_{X,T}\left\{\frac{|u'X_i|^2}{G_0^2(\tilde{T}_i)}E\left[\delta_i|X_i,T_i\right]\cdot I\left(|u'X_i|\geq\psi\sqrt{n}\right)\right\} \\
&= 4E_{X,T}\left\{\frac{|u'X_i|^2}{G_0(\tilde{T}_i)}\cdot I\left(|u'X_i|\geq\psi\sqrt{n}\right)\right\} \\
&= \tfrac{4}{\zeta_0}E\left\{|u'X_i|^2 I(|u'X_i|\geq\psi\sqrt{n})\right\}\to 0,\text{ as } n\to\infty
\end{aligned}
$$

since $I(|u'X_i|\geq\psi\sqrt{n})\to 0$ as $n\to\infty$. So then we have that $nE\left\{A_{ni}^2(u)\cdot I(|u'X_i|\geq\psi\sqrt{n})\right\}\to 0$ in probability.

Next, we show that $nE\left\{A_{ni}^2(u)\cdot I(|u'X_i|<\psi\sqrt{n})\right\}$ converges to a quadratic form of u. We have that

$$
\begin{aligned}
&nE\left\{A_{ni}^2(u)\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&=nE\left\{\tfrac{\delta_i}{G_0^2(\tilde{T}_i)}\left(\textstyle\int_0^{u^TX_i/\sqrt{n}}(I(\epsilon_i\le s)-I(\epsilon_i\le 0)]ds\right)^2\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&\le nE\left\{\tfrac{\delta_i}{G_0^2(\tilde{T}_i)}2\textstyle\int_0^{|u'X_i|/\sqrt{n}}ds\cdot\int_0^{|u'X_i|/\sqrt{n}}\left[I(\epsilon_i\le s)-I(\epsilon_i\le 0)\right]ds\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&\le nE\left\{\tfrac{\delta_i}{G_0^2(\tilde{T}_i)}2\textstyle\int_0^{\psi}ds\cdot\int_0^{|u'X_i|/\sqrt{n}}\left[I(\epsilon_i\le s)-I(\epsilon_i\le 0)\right]ds\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&\text{since }\tfrac{|u'X_i|}{\sqrt{n}}<\psi\text{ must be satisfied for the integral to be non-zero}\\
&\\
&=2n\psi E\left\{\tfrac{\delta_i}{G_0^2(\tilde{T}_i)}\textstyle\int_0^{|u'X_i|/\sqrt{n}}\left[I(\epsilon_i\le s)-I(\epsilon_i\le 0)\right]ds\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&=2n\psi E\left\{E\left(\tfrac{\delta_i}{G_0^2(\tilde{T}_i)}\textstyle\int_0^{|u'X_i|/\sqrt{n}}\left[I(\epsilon_i\le s)-I(\epsilon_i\le 0)\right]ds|X_i,T_i,\epsilon_i\right)\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&=2n\psi E\left\{\tfrac{1}{G_0^2(\tilde{T}_i)}\textstyle\int_0^{|u'X_i|/\sqrt{n}}\left[I(\epsilon\le s)-I(\epsilon_i\le 0)\right]ds\cdot E\left[\delta_i|X_i,T_i,\epsilon_i\right]\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&=2n\psi E\left\{\tfrac{1}{G_0(\tilde{T}_i)}\textstyle\int_0^{|u'X_i|/\sqrt{n}}\left[I(\epsilon_i\le s)-I(\epsilon_i\le 0)\right]ds\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&=2n\psi E_Z\left\{\tfrac{1}{G_0(\tilde{T}_i)}\textstyle\int_0^{|u'X_i|/\sqrt{n}}E_\epsilon\left[I(\epsilon_i\le s)-I(\epsilon_i\le 0)\right]ds\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&=2n\psi E_Z\left\{\tfrac{1}{G_0(\tilde{T}_i)}\textstyle\int_0^{|u'X_i|/\sqrt{n}}\left[F(s|Z)-F(0|Z)\right]ds\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&=2n\psi E_Z\left\{\tfrac{1}{G_0(\tilde{T}_i)}\textstyle\int_0^{|u'X_i|/\sqrt{n}}f(s^*|Z)s\,ds\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&\text{for some}\,0<s^*<s\,\text{by the Mean Value Theorem}\\
&\\
&\le 2n\phi B_0E_Z\left\{\tfrac{1}{G_0(\tilde{T}_i)}\textstyle\int_0^{|u'X_i|/\sqrt{n}}s\,ds\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&\text{since }\sup_s f(s|Z=z)\le B_0\\
&\\
&=\psi B_0E_Z\left\{\tfrac{|u'X_i|^2}{G_0(\tilde{T}_i)}\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&\le\tfrac{\psi B_0}{\zeta_0}E_Z\left\{|u'X_i|^2\cdot I(|u'X_i|<\psi\sqrt{n})\right\}\\
&\le\tfrac{\psi B_0}{\zeta_0}E_Z\left(|u'X_i|^2\right)
\end{aligned}
$$

Since $E_Z\left(|u'X_i|^2\right)$ is bounded, and ψ can be made arbitrarily small, it follows that $\frac{\psi B_0}{\zeta_0}E_Z\left(|u'X_i|^2\right)\to 0$ as $\psi\to 0$. Therefore, we have that, as $n\to\infty$,

$$Var\left\{\sum_{i=1}^n A_{ni}(u)\right\}=\sum_{i=1}^n Var\{A_{ni}(u)\}\le nE\{A_{ni}^2(u)\}\to 0,$$

which implies that $\sum_{i=1}^n[A_{ni}(u)-E\{A_{ni}(u)\}]=o_p(1)$. Further, we have that

$$
\begin{aligned}
E\left\{\sum_{i=1}^n A_{ni}(u)\right\} &= nE\{A_{n1}(u)\} \\
&= nE\left[\frac{\delta_1}{G_0(\tilde{T}_1)}\int_0^{u'X_1/\sqrt{n}}\{I(\epsilon_1 \le s) - I(\epsilon_1 \le 0)\}\,ds\right] \\
&= nE\left[\frac{1}{G_0(\tilde{T}_1)}\int_0^{u'X_1/\sqrt{n}}\{I(\epsilon_1 \le s) - I(\epsilon_1 \le 0)\}\,ds \cdot E\left\{\delta_1 | X_1, T_1\right\}\right] \\
&= nE_Z\left[\int_0^{u'X_1/\sqrt{n}}\{I(\epsilon_1 \le s) - I(\epsilon_1 \le 0)\}\,ds\right] \\
&= nE_Z\left[\int_0^{u'X_1/\sqrt{n}}\{F(s|Z) - F(0|Z)\}\,ds\right] \\
&= nE_Z\left\{\int_0^{u'X_1/\sqrt{n}} sf(0|Z)\,ds\right\} + o_p(1) \\
&= \tfrac{1}{2}u'\Sigma u + o_p(1),
\end{aligned}
$$

where $\Sigma = E_Z\left\{f(0|X)XX'\right\}$ is positive and finite. Thus, we have that

$$
Q(G_0, \theta_0 + \frac{u}{\sqrt{n}}) - Q(G_0, \theta_0) \ge \frac{1}{2}u'\Sigma u + u'W_1 - \kappa_1 O_p(||u||) + o_p(1).
$$

For the other terms in $D_n(u)$, by the Taylor expansion, we have

$$
\begin{aligned}
\sqrt{n}\{\tfrac{1}{\hat{G}(\tilde{T}_i)} - \tfrac{1}{G_0(\tilde{T}_i)}\} &= -\tfrac{\sqrt{n}\{\hat{G}(\tilde{T}_i) - G_0(\tilde{T}_i)\}}{G_0^2(\tilde{T}_i)} + o_p(1) \\
&= \tfrac{1}{G_0(\tilde{T}_i)}\tfrac{1}{\sqrt{n}}\sum_{j=1}^n \int_0^\tau I(\tilde{T}_i \ge s)\tfrac{dM_j^C(s)}{y(s)} + o_p(1),
\end{aligned}
$$

where $y(s) = \lim_{n\to\infty}(1/n)\sum_{i=1}^n I(\tilde{T}_i \ge s)$, $M_i^C(t) = (1-\delta_i)I(\tilde{T}_i \le t) - \int_0^t I(\tilde{T}_i \ge s)d\Lambda_C(s)$, and $\Lambda_C(\cdot)$ is the cumulative hazard function of the censoring time C. This leads to

$$
\begin{aligned}
Q\left(\hat{G}, \theta_0 + \tfrac{u}{\sqrt{n}}\right) - Q\left(G_0, \theta_0 + \tfrac{u}{\sqrt{n}}\right) &= \textstyle\sum_{i=1}^n \tfrac{\delta_i}{\hat{G}(\tilde{T}_i)}|\epsilon_i - X_i'\tfrac{u}{\sqrt{n}}| + n\lambda\sum_{j=1}^p \tfrac{|\beta_{j0}+u_j/\sqrt{n}|}{|\tilde{\beta}_j|} \\
&- \left(\textstyle\sum_{i=1}^n \tfrac{\delta_i}{G_0(\tilde{T}_i)}|\epsilon_i - X_i'\tfrac{u}{\sqrt{n}}| + n\lambda\sum_{j=1}^p \tfrac{|\beta_{j0}+u_j/\sqrt{n}|}{|\tilde{\beta}_j|}\right) \\
&= \textstyle\sum_{i=1}^n \delta_i|\epsilon_i - X_i'\tfrac{u}{\sqrt{n}}|\left(\tfrac{1}{\hat{G}(\tilde{T}_i)} - \tfrac{1}{G_0(\tilde{T}_i)}\right) \\
&= \textstyle\tfrac{1}{n}\sum_{i=1}^n \tfrac{\delta_i}{G_0(\tilde{T}_i)}|\epsilon_i - X_i'\tfrac{u}{\sqrt{n}}|\sum_{j=1}^n \int_0^\tau \tfrac{I(\tilde{T}_i \ge s)}{y(s)}dM_j^C(s) \\
&+ o_p(1)
\end{aligned}
$$

.

Similarly,

$$
Q(\hat{G}, \theta_0) - Q(G_0, \theta_0) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{G_0(\tilde{T}_i)}|\epsilon_i|\sum_{j=1}^n \int_0^\tau \frac{I(\tilde{T}_i \ge s)}{y(s)}dM_j^C(s) + o_p(1).
$$

Therefore,

$$\begin{aligned} & Q(\hat{G},\theta_0+\tfrac{u}{\sqrt{n}})-Q(G_0,\theta_0+\tfrac{u}{\sqrt{n}})-\{Q(\hat{G},\theta_0)-Q(G_0,\theta_0)\} \\ = \;& \tfrac{1}{n}\textstyle\sum_{i=1}^n \frac{\delta_i}{G_0(\tilde{T}_i)}(|\epsilon_i - X_i'\frac{u}{\sqrt{n}}| - |\epsilon_i|)\int_0^\tau \sum_{j=1}^n \frac{I(\tilde{T}_i\geq s)}{y(s)}dM_j^C(s)+o_p(1). \\ = \;& \tfrac{1}{n}\textstyle\sum_{j=1}^n \int_0^\tau L_n^*(u)\frac{dM_j^C(s)}{y(s)}+o_p(1), \end{aligned}$$

where $L_n^*(u)=\sum_{i=1}^n I(\tilde{T}_i\geq s)\frac{\delta_i}{G_0(\tilde{T}_i)}\left(|\epsilon_i - X_i'\frac{u}{\sqrt{n}}| - |\epsilon_i|\right)$. Following the method used for $L_n(u)$, we get

$$L_n^*(u) = -\frac{u'}{\sqrt{n}}\sum_{i=1}^n \frac{\delta_i}{G_0(\tilde{T}_i)}X_i I(\tilde{T}_i\geq s)sign(\epsilon_i)+2\sum_{i=1}^n A_{ni}^*(u),$$

where $A_{ni}^*(u)=\frac{\delta_i}{G_0(\tilde{T}_i)}I(\tilde{T}_i\geq s)\int_0^{u'X_i/\sqrt{n}}\left[I(\epsilon_i\leq t)-I(\epsilon_i\leq 0)\right]dt$. For the first term in $L_n^*(u)$, we have that

$$\begin{aligned} \tfrac{u'}{\sqrt{n}}\textstyle\sum_{j=1}^n\left(\int_0^\tau \frac{1}{n}\sum_{i=1}^n\left[\frac{\delta_i}{G_0(\tilde{T}_i)}X_i I(\tilde{T}_i\geq s)sign(\epsilon_i)\right]\frac{dM_j^C(s)}{y(s)}\right) &= \tfrac{u'}{\sqrt{n}}\textstyle\sum_{j=1}^n\int_0^\tau h(s)\frac{dM_j^C(s)}{y(s)} \\ &+ o_p(1), \end{aligned}$$

by the Law of Large Numbers, where $h(s)=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\frac{\delta_i I(\tilde{T}_i\geq s)}{G_0(\tilde{T}_i)}X_i sign(\epsilon_i)$ is a bounded function on $[0,\tau]$. Then by the Martingale Central Limit Theorem, we have that $\frac{u'}{\sqrt{n}}\sum_{j=1}^n\int_0^\tau h(s)\frac{dM_j^C(s)}{y(s)}$ converges in distribution to $u'W_2$, where W_2 is $(p+1)$-dimensional Normal with mean 0 and covariance matrix $\Sigma_2=\int_0^\tau\frac{h^2(s)}{y(s)}d\Lambda_C(s)$.
For $2\sum_{i=1}^n A_{ni}^*(u)$, notice that $\{A_{ni}^*(u)\}^2\leq A_{ni}^2(u)$, so that

$$\begin{aligned} Var\left[\textstyle\sum_{i=1}^n A_{ni}^*(u)\right] &= n\cdot Var[A_{n1}^*(u)] \\ &\leq n\cdot E\left[A_{n1}^*(u)\right]^2 \\ &\leq n\cdot E\left[Z_{n1}^2(u)\right] \end{aligned}$$

which converges to 0. Therefore, $\sum_{i=1}^n A_{ni}^*(u)$ converges to $n\cdot E\left[Z_{n1}^*(u)\right]$, where

$$\begin{aligned} n\cdot E\left[A_{n1}^*(u)\right] &= n\cdot E\left\{I(\tilde{T}_1\geq s)\tfrac{\delta_i}{G_0(\tilde{T}_1)}\int_0^{u'X_1/\sqrt{n}}\left[I(\epsilon_1\leq t)-I(\epsilon_1\leq 0)\right]dt\right] \\ &= n\cdot E\left\{\tfrac{I(\tilde{T}_i\geq s)}{G_0(\tilde{T}_1)}\int_0^{u'X_1/\sqrt{n}}\left[I(\epsilon_1\leq t)-I(\epsilon_1\leq 0)\right]dt\cdot E\left[\delta_1|X_1,T_i,\epsilon_1\right]\right\} \\ &= n\cdot E\left\{I(\tilde{T}_1\geq s)\int_0^{u'X_1/\sqrt{n}}\left[I(\epsilon_1\leq t)-I(\epsilon_1\leq 0)\right]dt\right\} \\ &= n\cdot E\left\{I(\tilde{T}_1\geq s)\cdot E\left(\int_0^{u'X_1/\sqrt{n}}\left[I(\epsilon_1\leq t)-I(\epsilon_1\leq 0)\right)dt|T_1\right)\right\} \\ &= n\cdot E\left\{I(\tilde{T}_1\geq s)\int_0^{u'X_1/\sqrt{n}}\left(F_{\epsilon_1|T_1}(t)-F_{\epsilon_1|T_1}(0)\right)dt\right\} \\ &= n\cdot E\left\{I(\tilde{T}_1\geq s)\int_0^{u'X_1/\sqrt{n}}t\, f_{\epsilon_1|T_1}(0)dt\right\}+o_p(1) \\ &= n\cdot E\left\{I(\tilde{T}_1\geq s)f_{\epsilon_1|T_1}(0)\tfrac{1}{2}\tfrac{u'X_1X_1'u}{n}\right\}+o_p(1) \\ &= \tfrac{1}{2}u'\left[q(s)\right]u+o_p(1), \end{aligned}$$

where $q(s) = E\left[I(\tilde{T}_1 \geq s) f_{\epsilon_1|T_1}(0) X_i X_i'\right]$ is a bounded function. This results in

$$\frac{1}{n}\sum_{j=1}^{n}\left[\int_0^\tau \left(2\sum_{i=1}^{n} A_{ni}^*(u)\right)\frac{dM_j^C(s)}{y(s)}\right] = u'\left[\frac{1}{n}\sum_{j=1}^{n}\left(\int_0^\tau q(s)\frac{dM_j^C(s)}{y(s)}\right)\right]u + o_p(1),$$

which is $o_p(1)$ by the Law of Large Numbers, since $\frac{1}{n}\sum_{j=1}^{n}\left(\int_0^\tau q(s)\frac{dM_j^C(s)}{y(s)}\right)$ converges to 0.

In summary, we showed that

$$D_n(u) \geq \frac{1}{2}u'\Sigma u + u'(W_1 + W_2) - \kappa_1 O_p(||u||) + o_p(1).$$

For the right-hand side in the above expression, the first term dominates the remain terms if $M = ||u||$ is large enough. So for any $\eta > 0$, as n gets large, we have

$$P\left\{\inf_{||u||=M} D_n(u) > 0\right\} \geq 1 - \eta,$$

which implies that $\hat{\theta}$ is $\sqrt{n}$-consistent. □

Proof of Theorem 2: (i) Proof of selection-consistency. We will first take the derivative of $Q(\hat{G}, \theta)$ with respect to β_j for $j = q+1, \cdots, p$, at any differentiable point $\theta = (\alpha, \beta_a', \beta_b')'$. Then we will show that for $\sqrt{n}||\alpha - \alpha_0|| \leq M$, $\sqrt{n}||\beta_a - \beta_{a0}|| \leq M$, and $||\beta_b - \beta_{b0}|| = ||\beta_b|| \leq \frac{M}{\sqrt{n}} \equiv \varepsilon_n$, when n is large, $\frac{1}{\sqrt{n}}\frac{\delta}{\delta\beta_j}Q(\hat{G}, \theta)$, for $j = q+1, \cdots, p$, is negative if $-\varepsilon_n < \beta_j < 0$ and positive if $0 < \beta_j < \varepsilon_n$. Since $Q(\hat{G}, \theta)$ is a piecewise linear function of θ, it achieve its minimum at some breaking point. Moreover, based on Theorem 1, the minimizer $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_a', \hat{\beta}_b')'$ of $Q(\hat{G}, \theta)$ is $\sqrt{n}$-consistent. Thus, each component of $\hat{\beta}_b$ must be contained in the interval $(-\varepsilon_n, \varepsilon_n)$ for all large n. Then as $n \to \infty$, $P(\hat{\beta}_b = 0) \to 1$.

To do this, we have, for $j = q+1, \cdots, p$,

$$\begin{aligned}
\tfrac{1}{\sqrt{n}}\tfrac{\partial}{\partial\beta_j}Q(\hat{G},\theta) &= -\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}\tfrac{\delta_i}{\hat{G}(\tilde{T}_i)}Z_{ij}\,sign(\log(\tilde{T}_i)-X_i'\theta)+\tfrac{n\lambda\,sign(\beta_j)}{|\tilde{\beta}_j|}.\\
&= -\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}\tfrac{\delta_i}{G(\tilde{T}_i)}Z_{ij}sign\left(\log(\tilde{T}_i)-X_i'\theta\right)\\
&+ \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\tfrac{\delta_i}{G_0^2(\tilde{T}_i)}\sqrt{n}\left[\hat{G}(\tilde{T}_i)-G_0(\tilde{T}_i)\right]Z_{ij}sign\left(\log(\tilde{T}_i)-X_i'\theta\right)\\
&+ \sqrt{n}\lambda\tfrac{sign(\beta_j)}{|\tilde{\beta}_j|}+o_p\left(\tfrac{1}{n}\right)\\
&= -\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}\tfrac{\delta_i}{G_0(\tilde{T}_i)}Z_{ij}sign(\log(\tilde{T}_i)-X_i'\theta)\\
&- \tfrac{1}{\sqrt{n}}\textstyle\sum_{k=1}^{n}\int_0^\tau\left[\tfrac{1}{n}\sum_{i=1}^{n}\tfrac{\delta_i}{G_0(\tilde{T}_i)}I(\tilde{T}_i\geq s)Z_{ij}sign(\log(\tilde{T}_i)-X_i'\theta)\right]\tfrac{dM_k^C(s)}{y(s)}\\
&+ \sqrt{n}\lambda\tfrac{sign(\beta_j)}{|\tilde{\beta}_j|}+o_p(1).
\end{aligned}$$

using the Taylor expansion of $\frac{1}{\hat{G}(\tilde{T}_i)}$ around $G_0(\tilde{T}_i)$.

Let $\Delta=\sqrt{n}(\theta-\theta_0)$ and define

$$V_n(\Delta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i\frac{\delta_i}{G_0(\tilde{T}_i)}\text{sign}(\epsilon_i-X_i'\Delta/\sqrt{n}).$$

Write $V_n(\Delta)=\{V_{n,0}(\Delta),V_{n,1}(\Delta),\cdots,V_{n,p}(\Delta)\}'$. Then the first term at the right-hand side of $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\beta_j}Q(\hat{G},\theta)$ can be rewritten as $V_{n,j}(\Delta)$. As shown in Theorem 1, $V_n(0)$ converges in distribution to a $(p+1)$-dimensional normal with mean 0 and variance-covariance matrix Σ_1. Since $\sqrt{n}\lambda\to 0$, we know that $\sqrt{n}(\beta_a-\beta_{a0})=O_p(1)$ from Theorem 1, we can bound $\sqrt{n}||\beta_a-\beta_{a0}||$. Then using the result from Koenker and Zhao (1996), we get

$$\begin{aligned}
V_n(\Delta)-V_n(0)+\Sigma_1\Delta &= \tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}\tfrac{\delta_i}{G_0(\tilde{T}_i)}X_i\,sign(\log(\tilde{T}_i)-X_i'\theta)\\
&- \tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}\tfrac{\delta_i}{G_0(\tilde{T}_i)}X_i\,sign(\epsilon_i)+\Sigma_1\Delta\\
&\leq sup_{||\Delta||\leq M}||V_n(\Delta)-V_n(0)+\Sigma_1\Delta||=o_p(1).
\end{aligned}$$

Result from Koenker and Zhao (1996): Suppose u_i is such that $\left(E\left[|u_i|^{r_u}\right]\right)^{1/r_u}<\infty$ for some $1\leq r_u<\infty$, and u_i are IID with distribution function F, and g_i are such that $\left(E\left[|g_i|^{r_g}\right]\right)^{1/r_g}<\infty$ for some $1\leq r_g<\infty$. Also, H_i is a p-dimensional random vector such that $E\left[||H_i||\right]^{2+\delta}\leq S<\infty$, g_i and H_i are independent of u_i, and $\frac{1}{n}\sum_{i=1}^{n}g_iH_i^T\to G$ in probability. Then

$$V(\Delta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_i\psi_\tau\left(u_i-F^{-1}(\tau)-\frac{1}{\sqrt{n}}H_i^T\Delta\right)$$

satisfies

$$sup_{||\Delta||\leq M}|V(\Delta) - V(0) + f(F^{-1}(\tau))G\Delta| = o_p(1).$$

We let $u_i = \epsilon_i$ with $r_u = 1$, so that $E[|\epsilon_i|] < \infty$. Let $g_i = \frac{\delta_i}{G_0(\tilde{T}_i)}X_i$, with $r_g = 2$, so that $\left(E\left[\frac{\delta_i}{G_0^2(\tilde{T}_i)}X_iX_i'\right]\right)^{1/2} < \infty$ since Σ_1 is finite. Let $H_i = X_i$, so that $E[||X_i||^{2+\delta}]$ is bounded since Σ_1 is finite. Also, X_i is independent of ϵ_i. Let $G = \Sigma_1$, since $\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{G_0(\tilde{T}_i)}X_iX_i'$ converges in probability to Σ_1. Let $\tau = \frac{1}{2}$ and $\psi_\tau(z) = sign(z)$. Then

$$V_n(\Delta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\delta_i}{G_0(\tilde{T}_i)}X_i\, sign(\epsilon_i - \frac{1}{\sqrt{n}}X_i'\Delta)$$

satisfies

$$sup_{||\Delta||\leq M}|V_n(\Delta) - V_n(0) + \Sigma_1\Delta| = o_p(1).$$

So we have that $V_n(\Delta) - V_n(0) + \Sigma_1\Delta = o_p(1)$. We have already shown that $V_n(0)$ converges in distribution to a normal vector. By assumption, Σ_1 is finite, and the conditions placed on α, β_a, and β_b ensure that Δ is bounded. This implies that $V_{n,j}(\Delta) = O_p(1)$.

Next, since $\sqrt{n}||\theta - \theta_0||$ is bounded, by the law of large numbers, we have that $\frac{1}{n}\sum_{i=1}^n Z_{ij}\frac{\delta_i}{G_0(\tilde{T}_i)}I(\tilde{T}_i \geq s)\text{sign}\{\log(\tilde{T}_i) - X_i'\theta\}$ converges to
$b_j(s,\theta) \equiv E\left\{Z_{ij}\frac{\delta_i}{G_0(\tilde{T}_i)}I(\tilde{T}_i \geq s)\text{sign}(\epsilon_i)\right\}$. Thus, the second term at the right-hand side of $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\beta_j}Q(\hat{G},\theta)$ can be written as

$$\frac{1}{\sqrt{n}}\sum_{k=1}^n \int_0^\tau \frac{b_j(s,\theta)}{y(s)}dM_k^C(s) + o_p(1).$$

By a Taylor expansion of $b_j(s,\theta)$ around θ_0, we get

$$b_j(s,\theta) = b_j(s,\theta_0) + b_j'(s,\theta_0)||\theta - \theta_0|| + o_p(1),$$

leading to

$$\frac{1}{\sqrt{n}}\sum_{k=1}^n \int_0^\tau b_j'(s,\theta_0)||\theta - \theta_0||\frac{dM_k^C(s)}{y(s)} = \sqrt{n}||\theta - \theta_0||\frac{1}{n}\sum_{k=1}^n \int_0^\tau b_j'(s,\theta_0) = o_p(1),$$

since $\sqrt{n}||\theta - \theta_0|| = O_p(1)$ by assumption and $\frac{1}{n}\sum_{k=1}^n \int_0^\tau b_j'(s,\theta_0)\frac{dM_k^C(s)}{y(s)} = o_p(1)$ by the law of large numbers.

So then we have that

$$\frac{1}{\sqrt{n}}\sum_{k=1}^n \int_0^\tau b_j(s,\theta)\frac{dM_k^C(s)}{y(s)} = \frac{1}{\sqrt{n}}\sum_{k=1}^n \int_0^\tau b_j(s,\theta_0)\frac{dM_k^C(s)}{y(s)} + o_p(1),$$

which is $O_p(1)$ since, by the Martingale Central Limit Theorem, $\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\int_0^\tau b_j(s,\theta_0)\frac{dM_k^C(s)}{y(s)}$ converges to a normal variable with mean 0.

Thus, for $j = q+1,\ldots,p$,

$$\frac{1}{\sqrt{n}}\frac{\partial}{\partial\beta_j}Q(\hat{G},\theta) = O_p(1) + \frac{n\lambda \text{sign}(\beta_j)}{\sqrt{n}|\tilde{\beta}_j|}.$$

Since the LAD estimator $\tilde{\theta}$ is $\sqrt{n}$-consistent, we have, for $j = q+1,\ldots,p$, $\sqrt{n}|\tilde{\beta}_j| = O_p(1)$. Then based on the assumption $n\lambda \to \infty$, when n is large, the sign of $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\beta_j}Q(\hat{G},\theta)$ is determined by the sign of β_j. So as n gets large, $\frac{1}{\sqrt{n}}\frac{\delta}{\delta\beta_j}Q(\hat{G},\theta)$, for $j = q+1,\cdots,p$, is negative if $-\varepsilon_n < \beta_j < 0$ and positive if $0 < \beta_j < \varepsilon_n$, which implies $P(\hat{\beta}_b = 0) \to 1$ as $n \to \infty$. □

(ii) Proof of asymptotic normality. Based on the results established in Theorem 1 and (i) of Theorem 2, we have the minimizer $\hat{\theta}$ is $\sqrt{n}$-consistent and $P(\hat{\beta}_b = 0) \to 1$ as $n \to \infty$. Thus to derive the asymptotic distribution for the estimators of non-zero coefficients, we only need to establish the asymptotic representation for the following function:

$$S_n(\hat{G},v) = Q\{\hat{G},(\beta'_{a'0} + v'/\sqrt{n},0')'\} - Q\{\hat{G},(\beta'_{a'0},0')'\},$$

where $\beta_{a'0} = (\alpha_0,\beta'_{a0})'$ and v is a $(q+1)$-dimensional vector with bounded norm. Define $X_{a'i} = (1,Z_{i1},\cdots,Z_{iq})'$, $i = 1,\cdots,n$. Following the similar derivations as those in the proof of Theorem 1, we can show that

$$\begin{aligned} S_n(\hat{G},v) &= \frac{1}{2}v'\Sigma_{a'}v + \frac{v'}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{G_0(\tilde{T}_i)}X_{a'i}\{I(\epsilon_i<0) - I(\epsilon_i>0)\} \\ &\quad + \frac{v'}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \frac{h_{a'}(s)}{y(s)}dM_i^C(s) + o_p(1), \end{aligned} \tag{A.1}$$

where $\Sigma_{a'} = E\{f(0|X)X_{a'}X'_{a'}\}$ and $h_{a'}(s) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i I(\tilde{T}_i\geq s)}{G_0(\tilde{T}_i)}X_{a'i}\{I(\epsilon_i<0) - I(\epsilon_i>0)\}$. Define

$$s_i = \frac{\delta_i}{G_0(\tilde{T}_i)}X_{a'i}\{I(\epsilon_i<0) - I(\epsilon_i>0)\} + \int_0^\tau \frac{h_{a'}(s)}{y(s)}dM_i^C(s).$$

Then by the central limit theorem, $(1/\sqrt{n})\sum_{i=1}^{n}s_i$ converges in distribution to a $(q+1)$-dimensional normal vector $W_{a'}$ with mean 0 and variance-covariance matrix $V_{a'} = E(s_1s'_1)$. By the lemma given in Davis et al. (1992), $\sqrt{n}(\hat{\beta}_{a'} - \beta_{a'0})$ converges to v_0 in distribution as $n \to \infty$. The lemma given in Davis et al. (1992) is as follows: Suppose $V_n(\cdot) \to V$ on

$C(\Re^p)$, ϵ_n minimizes $V_n(\cdot)$, and ϵ minimizes $V(\cdot)$. If $V_n(\cdot)$ is convex for each n and ϵ is unique with probability 1, then ϵ_n converges to ϵ in distribution.

We take $V_n(v) = S_n(\hat{G}, v)$, which is convex since $Q\left(\hat{G}, \beta_{a'0} + \frac{v}{\sqrt{n}}, 0'\right)$ is convex in v, and $Q(\hat{G}, \beta_0)$ is constant with respect to v. The minimizer of $Q(\cdot)$ is $(\hat{\beta}_{a'}, \hat{\beta}_b)$, where $\hat{\beta}_b \to 0$ in probability, so the minimum with respect to v is attained for $v = \sqrt{n}\left(\hat{\beta}_{a'} - \beta_{a'0}\right)$. So, we let $\epsilon_n = \sqrt{n}\left(\hat{\beta}_{a'} - \beta_{a'0}\right)$. Then $V = \frac{1}{2}v'\Sigma_{a'}v + v'W_{a'}$. To find ϵ, we take the derivative of V with respect to v and set it equal to 0. We then get

$$\frac{dV}{dv} = \Sigma_{a'}v' + W_{a'} = 0,$$

which implies that the minimizer of V is then $v_0 = -\Sigma_{a'}^{-1}W_{a'}$. So then we get $E(v_0') = -\Sigma_{a'}^{-1}E(W_{a'}) = 0$, and $Var(v_0') = \Sigma_{a'}^{-1}V_{a'}\Sigma_{a'}^{-1}$. Therefore we have that $\sqrt{n}(\hat{\beta}_{a'} - \beta_{a'0})$ converges in distribution to $(q+1)$-dimensional normal vector with mean 0 and variance-covariance matrix $\Sigma_{a'}^{-1}V_{a'}\Sigma_{a'}^{-1}$. $\square$

Appendix B: Simulation tables

Table B.1: Estimation of non-zero coefficients for t(5) error distribution

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
Prop.	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
20%	50	Full	-0.02	0.22	0.36	-0.04	0.26	0.37	-0.09	0.27	0.37
		ALasso	-0.13	0.25	0.32	-0.10	0.27	0.35	-0.12	0.27	0.36
		Lasso	-0.12	0.23	0.30	-0.12	0.25	0.35	-0.19	0.27	0.37
		Oracle	-0.01	0.21	0.31	-0.04	0.25	0.32	-0.06	0.26	0.32
	100	Full	-0.00	0.16	0.23	-0.07	0.17	0.22	-0.12	0.19	0.22
		ALasso	-0.08	0.20	0.22	-0.10	0.17	0.23	-0.15	0.17	0.22
		Lasso	-0.09	0.20	0.20	-0.17	0.18	0.24	-0.21	0.19	0.23
		Oracle	-0.00	0.17	0.21	-0.07	0.16	0.21	-0.12	0.18	0.21
	200	Full	-0.03	0.12	0.14	-0.05	0.13	0.14	-0.10	0.12	0.15
		ALasso	-0.05	0.13	0.15	-0.07	0.13	0.15	-0.11	0.11	0.15
		Lasso	-0.08	0.13	0.15	-0.12	0.13	0.16	-0.17	0.12	0.15
		Oracle	-0.02	0.12	0.13	-0.04	0.12	0.14	-0.09	0.10	0.14
40%	50	Full	-0.07	0.29	0.45	-0.16	0.30	0.47	-0.17	0.30	0.47
		ALasso	-0.18	0.29	0.37	-0.26	0.33	0.41	-0.25	0.31	0.44
		Lasso	-0.17	0.27	0.35	-0.33	0.34	0.39	-0.34	0.37	0.44
		Oracle	-0.08	0.25	0.37	-0.16	0.32	0.39	-0.15	0.29	0.41
	100	Full	-0.04	0.22	0.29	-0.14	0.19	0.29	-0.18	0.25	0.30
		ALasso	-0.13	0.26	0.26	-0.19	0.21	0.28	-0.23	0.26	0.30
		Lasso	-0.11	0.24	0.23	-0.23	0.22	0.28	-0.31	0.28	0.30
		Oracle	-0.05	0.22	0.27	-0.14	0.19	0.27	-0.21	0.25	0.28
	200	Full	-0.05	0.13	0.18	-0.13	0.15	0.19	-0.18	0.14	0.20
		ALasso	-0.10	0.16	0.17	-0.16	0.15	0.19	-0.21	0.14	0.20
		Lasso	-0.11	0.13	0.16	-0.21	0.16	0.20	-0.27	0.16	0.21
		Oracle	-0.05	0.13	0.18	-0.12	0.15	0.18	-0.18	0.14	0.19
		Avg. SE	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.01

Prop.: censoring proportion; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.2: Variable selection results for t(5) error distribution

				No. of times selected in 100 runs								
Prop.	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
20%	50	Full	0.57	100	100	100	100	100	100	0	0	0
		ALasso	0.52	78	98	100	100	24	24	29	0.24	3.17
		Lasso	0.58	90	99	100	100	51	51	8	0.11	2.14
		Oracle	0.44	100	100	100	100	0	0	100	0	4
	100	Full	0.42	100	100	100	100	100	100	0	0	0
		ALasso	0.38	93	100	100	100	8	16	59	0.07	3.47
		Lasso	0.47	95	100	100	100	39	37	16	0.05	2.37
		Oracle	0.32	100	100	100	100	0	0	100	0	4
	200	Full	0.31	100	100	100	100	100	100	0	0	0
		ALasso	0.29	99	100	100	100	10	7	65	0.01	3.62
		Lasso	0.36	99	100	100	100	35	35	19	0.01	2.52
		Oracle	0.25	100	100	100	100	0	0	100	0	4
40%	50	Full	0.73	100	100	100	100	100	100	0	0	0
		ALasso	0.73	66	92	100	100	28	24	16	0.42	2.86
		Lasso	0.84	74	89	98	99	51	53	7	0.40	1.92
		Oracle	0.57	100	100	100	100	0	0	100	0	4
	100	Full	0.61	100	100	100	100	100	100	0	0	0
		ALasso	0.62	78	99	100	100	17	15	34	0.23	3.25
		Lasso	0.69	87	99	100	100	42	49	4	0.14	2.07
		Oracle	0.53	100	100	100	100	0	0	100	0	4
	200	Full	0.48	100	100	100	100	100	100	0	0	0
		ALasso	0.50	94	100	100	100	13	11	55	0.06	3.45
		Lasso	0.57	99	100	100	100	43	42	13	0.01	2.29
		Oracle	0.44	100	100	100	100	0	0	100	0	4
		Avg. SE	0.02									

Prop.: censoring proportion; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.3: Estimation of non-zero coefficients for double exponential error distribution

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
Prop.	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
20%	50	Full	-0.01	0.26	0.40	-0.08	0.26	0.40	-0.06	0.30	0.39
		ALasso	-0.10	0.30	0.34	-0.16	0.32	0.38	-0.10	0.30	0.39
		Lasso	-0.10	0.27	0.31	-0.21	0.29	0.38	-0.15	0.30	0.39
		Oracle	0.01	0.24	0.32	-0.07	0.27	0.33	-0.04	0.26	0.33
	100	Full	-0.03	0.15	0.22	-0.07	0.17	0.23	-0.11	0.17	0.23
		ALasso	-0.08	0.19	0.21	-0.09	0.17	0.23	-0.12	0.18	0.23
		Lasso	-0.10	0.16	0.20	-0.15	0.18	0.24	-0.20	0.19	0.24
		Oracle	-0.02	0.16	0.20	-0.06	0.16	0.20	-0.10	0.16	0.21
	200	Full	-0.06	0.12	0.14	-0.06	0.11	0.14	-0.08	0.12	0.15
		ALasso	-0.08	0.12	0.15	-0.07	0.10	0.14	-0.09	0.12	0.15
		Lasso	-0.10	0.12	0.15	-0.12	0.11	0.16	-0.14	0.12	0.15
		Oracle	-0.04	0.11	0.13	-0.06	0.10	0.14	-0.07	0.12	0.14
40%	50	Full	-0.08	0.30	0.44	-0.17	0.30	0.44	-0.24	0.28	0.44
		ALasso	-0.19	0.29	0.37	-0.26	0.32	0.40	-0.31	0.30	0.41
		Lasso	-0.16	0.26	0.34	-0.27	0.31	0.38	-0.38	0.33	0.41
		Oracle	-0.07	0.28	0.37	-0.13	0.24	0.37	-0.21	0.27	0.36
	100	Full	-0.05	0.20	0.29	-0.12	0.20	0.29	-0.20	0.24	0.30
		ALasso	-0.15	0.24	0.25	-0.19	0.21	0.28	-0.25	0.24	0.31
		Lasso	-0.14	0.22	0.23	-0.23	0.21	0.29	-0.32	0.24	0.31
		Oracle	-0.06	0.19	0.27	-0.12	0.19	0.27	-0.19	0.22	0.28
	200	Full	-0.04	0.14	0.18	-0.13	0.16	0.19	-0.15	0.15	0.20
		ALasso	-0.09	0.16	0.18	-0.16	0.15	0.19	-0.17	0.15	0.20
		Lasso	-0.10	0.14	0.16	-0.22	0.16	0.20	-0.25	0.17	0.21
		Oracle	-0.03	0.16	0.18	-0.13	0.17	0.18	-0.16	0.15	0.19
		Avg. SE	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01

Prop.: censoring proportion; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.4: Variable selection results for double exponential error distribution

				No. of times selected in 100 runs								
Prop.	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
20%	50	Full	0.57	100	100	100	100	100	100	0	0	0
		ALasso	0.54	75	98	100	100	17	21	35	0.27	3.21
		Lasso	0.58	85	98	100	100	50	50	11	0.17	2.18
		Oracle	0.42	100	100	100	100	0	0	100	0	4
	100	Full	0.40	100	100	100	100	100	100	0	0	0
		ALasso	0.36	92	100	100	100	10	11	53	0.08	3.50
		Lasso	0.44	96	100	100	100	42	39	17	0.04	2.40
		Oracle	0.30	100	100	100	100	0	0	100	0	4
	200	Full	0.29	100	100	100	100	100	100	0	0	0
		ALasso	0.26	99	100	100	100	10	5	76	0.01	3.69
		Lasso	0.33	98	100	100	100	39	42	18	0.02	2.30
		Oracle	0.22	100	100	100	100	0	0	100	0	4
40%	50	Full	0.76	100	100	100	100	100	100	0	0	0
		ALasso	0.77	69	94	100	100	21	26	26	0.37	3.08
		Lasso	0.84	84	96	98	100	43	48	4	0.22	2.05
		Oracle	0.59	100	100	100	100	0	0	100	0	4
	100	Full	0.61	100	100	100	100	100	100	0	0	0
		ALasso	0.62	75	100	100	100	18	18	33	0.25	3.32
		Lasso	0.71	85	100	100	100	55	48	4	0.15	2.00
		Oracle	0.48	100	100	100	100	0	0	100	0	4
	200	Full	0.48	100	100	100	100	100	100	0	0	0
		ALasso	0.47	96	100	100	100	10	13	64	0.04	3.51
		Lasso	0.58	99	100	100	100	42	46	16	0.01	2.29
		Oracle	0.42	100	100	100	100	0	0	100	0	4
		Avg. SE	0.02									

Prop.: censoring proportion; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.5: Estimation of non-zero coefficients for logistic error distribution

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
Prop.	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
20%	50	Full	-0.07	0.37	0.51	-0.04	0.42	0.52	-0.11	0.41	0.50
		ALasso	-0.23	0.32	0.43	-0.19	0.44	0.48	-0.18	0.42	0.48
		Lasso	-0.20	0.32	0.39	-0.26	0.42	0.46	-0.29	0.38	0.48
		Oracle	-0.06	0.35	0.43	-0.07	0.38	0.44	-0.11	0.35	0.42
	100	Full	-0.05	0.25	0.29	-0.08	0.26	0.30	-0.19	0.24	0.29
		ALasso	-0.17	0.29	0.26	-0.15	0.26	0.30	-0.24	0.25	0.29
		Lasso	-0.16	0.26	0.24	-0.21	0.27	0.31	-0.31	0.25	0.30
		Oracle	-0.04	0.25	0.27	-0.07	0.25	0.27	-0.17	0.24	0.27
	200	Full	-0.05	0.16	0.18	-0.12	0.16	0.18	-0.17	0.19	0.19
		ALasso	-0.13	0.19	0.19	-0.16	0.17	0.19	-0.20	0.19	0.19
		Lasso	-0.13	0.16	0.18	-0.20	0.16	0.20	-0.26	0.19	0.20
		Oracle	-0.05	0.15	0.18	-0.13	0.15	0.17	-0.17	0.19	0.18
40%	50	Full	-0.17	0.41	0.49	-0.28	0.34	0.50	-0.36	0.43	0.49
		ALasso	-0.27	0.38	0.40	-0.43	0.39	0.43	-0.46	0.46	0.45
		Lasso	-0.26	0.33	0.35	-0.47	0.37	0.40	-0.55	0.44	0.43
		Oracle	-0.18	0.42	0.40	-0.28	0.35	0.42	-0.35	0.38	0.41
	100	Full	-0.07	0.31	0.33	-0.21	0.27	0.34	-0.38	0.29	0.34
		ALasso	-0.19	0.32	0.27	-0.30	0.33	0.31	-0.47	0.31	0.34
		Lasso	-0.20	0.28	0.25	-0.35	0.30	0.30	-0.54	0.29	0.34
		Oracle	-0.08	0.28	0.30	-0.19	0.27	0.32	-0.37	0.30	0.32
	200	Full	-0.10	0.17	0.21	-0.23	0.22	0.21	-0.31	0.23	0.22
		ALasso	-0.20	0.21	0.19	-0.29	0.22	0.22	-0.35	0.22	0.23
		Lasso	-0.19	0.19	0.17	-0.33	0.22	0.22	-0.43	0.26	0.23
		Oracle	-0.09	0.17	0.20	-0.22	0.19	0.21	-0.30	0.21	0.21
		Avg. SE	0.03	0.02	0.01	0.03	0.02	0.01	0.03	0.02	0.01

Prop.: censoring proportion; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.6: Variable selection results for logistic error distribution

				No. of times selected in 100 runs								
Prop.	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
20%	50	Full	0.86	100	100	100	100	100	100	0	0	0
		ALasso	0.82	48	90	99	99	22	23	15	0.64	3.13
		Lasso	0.88	60	90	99	99	34	42	8	0.52	2.41
		Oracle	0.67	100	100	100	100	0	0	100	0	4
	100	Full	0.63	100	100	100	100	100	100	0	0	0
		ALasso	0.63	67	98	100	100	14	9	34	0.35	3.38
		Lasso	0.70	80	98	100	100	39	35	11	0.22	2.40
		Oracle	0.51	100	100	100	100	0	0	100	0	4
	200	Full	0.48	100	100	100	100	100	100	0	0	0
		ALasso	0.48	88	100	100	100	6	10	55	0.12	3.51
		Lasso	0.54	99	100	100	100	38	29	27	0.01	2.50
		Oracle	0.41	100	100	100	100	0	0	100	0	4
40%	50	Full	1.13	100	100	100	100	100	100	0	0	0
		ALasso	1.18	49	81	97	100	20	29	12	0.73	2.98
		Lasso	1.26	64	80	94	98	41	50	3	0.64	2.13
		Oracle	0.97	100	100	100	100	0	0	100	0	4
	100	Full	0.96	100	100	100	100	100	100	0	0	0
		ALasso	1.03	58	93	99	99	16	14	27	0.51	3.28
		Lasso	1.13	74	93	99	100	36	32	9	0.34	2.39
		Oracle	0.85	100	100	100	100	0	0	100	0	4
	200	Full	0.80	100	100	100	100	100	100	0	0	0
		ALasso	0.83	77	99	100	100	16	17	36	0.24	3.32
		Lasso	0.92	87	100	100	100	41	42	14	0.13	2.28
		Oracle	0.72	100	100	100	100	0	0	100	0	4
		Avg. SE	0.02									

Prop.: censoring proportion; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.7: Estimation of non-zero coefficients for extreme value error distribution

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
Prop.	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
20%	50	Full	-0.02	0.31	0.39	0.00	0.30	0.39	-0.07	0.30	0.38
		ALasso	-0.13	0.30	0.33	-0.10	0.30	0.38	-0.13	0.30	0.37
		Lasso	-0.11	0.29	0.31	-0.15	0.29	0.38	-0.22	0.30	0.38
		Oracle	-0.03	0.32	0.33	-0.06	0.28	0.33	-0.07	0.28	0.33
	100	Full	-0.05	0.19	0.22	-0.07	0.21	0.23	-0.10	0.19	0.23
		ALasso	-0.12	0.22	0.21	-0.10	0.20	0.23	-0.13	0.19	0.23
		Lasso	-0.13	0.21	0.19	-0.15	0.20	0.24	-0.19	0.19	0.23
		Oracle	-0.02	0.19	0.21	-0.05	0.20	0.21	-0.08	0.20	0.21
	200	Full	-0.03	0.12	0.14	-0.08	0.13	0.15	-0.10	0.13	0.15
		ALasso	-0.08	0.14	0.15	-0.11	0.13	0.15	-0.13	0.13	0.15
		Lasso	-0.09	0.12	0.16	-0.15	0.13	0.16	-0.18	0.13	0.15
		Oracle	-0.04	0.12	0.14	-0.09	0.12	0.14	-0.11	0.13	0.14
40%	50	Full	-0.10	0.30	0.42	-0.19	0.28	0.44	-0.26	0.34	0.43
		ALasso	-0.22	0.29	0.34	-0.29	0.33	0.39	-0.33	0.39	0.40
		Lasso	-0.21	0.28	0.31	-0.33	0.29	0.37	-0.42	0.36	0.40
		Oracle	-0.12	0.31	0.36	-0.16	0.29	0.37	-0.22	0.34	0.37
	100	Full	-0.05	0.23	0.28	-0.10	0.23	0.29	-0.26	0.24	0.29
		ALasso	-0.17	0.27	0.24	-0.17	0.24	0.28	-0.30	0.24	0.29
		Lasso	-0.16	0.23	0.22	-0.22	0.23	0.27	-0.38	0.25	0.29
		Oracle	-0.04	0.22	0.26	-0.10	0.24	0.26	-0.24	0.21	0.27
	200	Full	-0.10	0.16	0.18	-0.16	0.14	0.18	-0.21	0.17	0.19
		ALasso	-0.17	0.18	0.17	-0.20	0.16	0.19	-0.25	0.17	0.19
		Lasso	-0.16	0.16	0.16	-0.25	0.17	0.20	-0.30	0.18	0.20
		Oracle	-0.09	0.16	0.18	-0.17	0.15	0.18	-0.20	0.17	0.18
		Avg. SE	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02

Prop.: censoring proportion; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.8: Variable selection results for extreme value error distribution

				No. of times selected in 100 runs								
Prop.	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
20%	50	Full	0.67	100	100	100	100	100	100	0	0	0
		ALasso	0.62	71	100	99	100	14	26	32	0.30	3.19
		Lasso	0.69	80	100	99	100	43	52	6	0.21	2.13
		Oracle	0.52	100	100	100	100	0	0	100	0	4
	100	Full	0.47	100	100	100	100	100	100	0	0	0
		ALasso	0.45	86	100	100	100	15	12	40	0.14	3.34
		Lasso	0.50	89	100	100	100	39	37	11	0.11	2.36
		Oracle	0.35	100	100	100	100	0	0	100	0	4
	200	Full	0.34	100	100	100	100	100	100	0	0	0
		ALasso	0.33	99	100	100	100	9	8	72	0.01	3.62
		Lasso	0.39	100	100	100	100	35	32	28	0	2.61
		Oracle	0.29	100	100	100	100	0	0	100	0	4
40%	50	Full	0.84	100	100	100	100	100	100	0	0	0
		ALasso	0.87	59	92	98	100	25	21	16	0.51	3.09
		Lasso	0.95	68	95	97	100	55	57	6	0.40	1.98
		Oracle	0.68	100	100	100	100	0	0	100	0	4
	100	Full	0.71	100	100	100	100	100	100	0	0	0
		ALasso	0.74	71	100	100	100	14	18	32	0.29	3.31
		Lasso	0.82	85	100	100	100	41	39	7	0.15	2.16
		Oracle	0.61	100	100	100	100	0	0	100	0	4
	200	Full	0.59	100	100	100	100	100	100	0	0	0
		ALasso	0.62	87	100	100	100	18	16	45	0.13	3.29
		Lasso	0.69	93	100	100	100	58	53	8	0.07	1.96
		Oracle	0.53	100	100	100	100	0	0	100	0	4
		Avg. SE	0.02									

Prop.: censoring proportion; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.9: Estimation of non-zero coefficients for sensitivity analysis, $\gamma = 0.1$

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
Prop.	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
20%	50	Full	0.01	0.22	0.36	-0.05	0.26	0.37	-0.09	0.29	0.37
		ALasso	-0.11	0.27	0.32	-0.11	0.27	0.35	-0.12	0.28	0.36
		Lasso	-0.09	0.23	0.31	-0.14	0.25	0.35	-0.19	0.28	0.37
		Oracle	0.01	0.22	0.31	-0.04	0.24	0.31	-0.06	0.26	0.32
	100	Full	0.02	0.17	0.22	-0.07	0.17	0.22	-0.11	0.20	0.22
		ALasso	-0.05	0.20	0.22	-0.11	0.17	0.23	-0.13	0.19	0.22
		Lasso	-0.08	0.19	0.20	-0.17	0.18	0.24	-0.21	0.20	0.23
		Oracle	0.02	0.18	0.20	-0.07	0.17	0.20	-0.11	0.20	0.21
	200	Full	0.01	0.14	0.14	-0.06	0.12	0.15	-0.11	0.12	0.15
		ALasso	-0.02	0.14	0.14	-0.08	0.11	0.15	-0.12	0.12	0.15
		Lasso	-0.05	0.14	0.15	-0.13	0.12	0.16	-0.18	0.13	0.15
		Oracle	0.01	0.14	0.13	-0.05	0.11	0.14	-0.11	0.12	0.14
40%	50	Full	-0.04	0.27	0.39	-0.12	0.27	0.39	-0.19	0.30	0.40
		ALasso	-0.17	0.30	0.33	-0.19	0.29	0.35	-0.26	0.29	0.39
		Lasso	-0.13	0.26	0.32	-0.23	0.30	0.34	-0.32	0.32	0.39
		Oracle	-0.08	0.24	0.33	-0.11	0.27	0.34	-0.15	0.25	0.36
	100	Full	0.00	0.23	0.28	-0.15	0.22	0.28	-0.20	0.24	0.29
		ALasso	-0.06	0.28	0.26	-0.20	0.24	0.27	-0.25	0.25	0.29
		Lasso	-0.06	0.26	0.24	-0.23	0.23	0.27	-0.32	0.25	0.29
		Oracle	0.01	0.25	0.26	-0.14	0.20	0.26	-0.21	0.24	0.27
	200	Full	-0.02	0.14	0.17	-0.12	0.15	0.18	-0.17	0.14	0.19
		ALasso	-0.06	0.17	0.18	-0.16	0.15	0.19	-0.20	0.13	0.20
		Lasso	-0.07	0.15	0.18	-0.20	0.16	0.19	-0.27	0.15	0.20
		Oracle	-0.01	0.14	0.17	-0.13	0.15	0.18	-0.18	0.13	0.19
		Avg. SE	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01

Prop.: censoring proportion; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.10: Variable selection results for sensitivity analysis, $\gamma = 0.1$

				No. of times selected in 100 runs								
Prop.	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
20%	50	Full	0.57	100	100	100	100	100	100	0	0	0
		ALasso	0.53	79	97	100	100	22	25	31	0.24	3.18
		Lasso	0.58	92	99	100	100	47	48	11	0.09	2.18
		Oracle	0.44	100	100	100	100	0	0	100	0	4
	100	Full	0.42	100	100	100	100	100	100	0	0	0
		ALasso	0.38	96	100	100	100	9	13	65	0.04	3.51
		Lasso	0.46	97	100	100	100	34	35	18	0.03	2.47
		Oracle	0.33	100	100	100	100	0	0	100	0	4
	200	Full	0.32	100	100	100	100	100	100	0	0	0
		ALasso	0.30	100	100	100	100	11	10	65	0	3.53
		Lasso	0.36	100	100	100	100	34	35	24	0	2.55
		Oracle	0.26	100	100	100	100	0	0	100	0	4
40%	50	Full	0.71	100	100	100	100	100	100	0	0	0
		ALasso	0.70	65	98	99	100	25	22	17	0.38	3.06
		Lasso	0.78	83	96	99	100	55	52	7	0.22	1.89
		Oracle	0.57	100	100	100	100	0	0	100	0	4
	100	Full	0.62	100	100	100	100	100	100	0	0	0
		ALasso	0.62	86	98	100	100	22	14	37	0.16	3.28
		Lasso	0.69	90	98	100	100	40	43	6	0.12	2.15
		Oracle	0.54	100	100	100	100	0	0	100	0	4
	200	Full	0.48	100	100	100	100	100	100	0	0	0
		ALasso	0.49	96	100	100	100	11	7	58	0.04	3.51
		Lasso	0.57	99	100	100	100	42	42	13	0.01	2.32
		Oracle	0.44	100	100	100	100	0	0	100	0	4
		Avg. SE	0.02									

Prop.: censoring proportion; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.11: Estimation of non-zero coefficients for sensitivity analysis, $\gamma = 0.2$

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
Prop.	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
20%	50	Full	0.01	0.23	0.36	-0.06	0.27	0.37	-0.11	0.27	0.37
		ALasso	-0.09	0.27	0.33	-0.11	0.28	0.35	-0.14	0.27	0.36
		Lasso	-0.08	0.23	0.31	-0.14	0.25	0.35	-0.20	0.27	0.36
		Oracle	0.01	0.21	0.31	-0.05	0.24	0.31	-0.08	0.24	0.32
	100	Full	0.04	0.17	0.22	-0.06	0.17	0.22	-0.11	0.21	0.22
		ALasso	-0.03	0.20	0.22	-0.10	0.17	0.23	-0.14	0.20	0.22
		Lasso	-0.04	0.20	0.21	-0.15	0.18	0.24	-0.20	0.21	0.23
		Oracle	0.05	0.18	0.21	-0.06	0.16	0.21	-0.10	0.20	0.21
	200	Full	0.02	0.12	0.14	-0.06	0.12	0.14	-0.10	0.12	0.15
		ALasso	-0.02	0.13	0.14	-0.08	0.12	0.15	-0.11	0.12	0.14
		Lasso	-0.03	0.12	0.15	-0.12	0.13	0.16	-0.16	0.13	0.15
		Oracle	0.02	0.12	0.13	-0.06	0.12	0.14	-0.11	0.12	0.14
40%	50	Full	-0.04	0.28	0.38	-0.11	0.26	0.39	-0.22	0.31	0.40
		ALasso	-0.13	0.30	0.33	-0.19	0.28	0.35	-0.26	0.30	0.39
		Lasso	-0.10	0.28	0.32	-0.23	0.28	0.34	-0.33	0.34	0.39
		Oracle	-0.05	0.26	0.33	-0.11	0.25	0.34	-0.17	0.26	0.36
	100	Full	0.02	0.22	0.29	-0.14	0.21	0.28	-0.21	0.23	0.29
		ALasso	-0.05	0.26	0.27	-0.19	0.22	0.28	-0.25	0.23	0.29
		Lasso	-0.04	0.23	0.26	-0.24	0.22	0.27	-0.32	0.25	0.29
		Oracle	0.03	0.21	0.26	-0.14	0.20	0.26	-0.21	0.22	0.27
	200	Full	0.01	0.13	0.17	-0.12	0.14	0.18	-0.18	0.14	0.19
		ALasso	-0.02	0.15	0.18	-0.15	0.14	0.19	-0.21	0.14	0.20
		Lasso	-0.03	0.14	0.18	-0.20	0.14	0.19	-0.27	0.16	0.21
		Oracle	0.02	0.14	0.17	-0.12	0.14	0.18	-0.19	0.13	0.19
		Avg. SE	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01

Prop.: censoring proportion; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.12: Variable selection results for sensitivity analysis, $\gamma = 0.2$

				No. of times selected in 100 runs								
Prop.	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
20%	50	Full	0.57	100	100	100	100	100	100	0	0	0
		ALasso	0.53	82	97	100	100	21	30	29	0.21	3.14
		Lasso	0.58	94	99	100	100	47	48	10	0.07	2.17
		Oracle	0.44	100	100	100	100	0	0	100	0	4
	100	Full	0.42	100	100	100	100	100	100	0	0	0
		ALasso	0.37	96	100	100	100	10	13	64	0.04	3.52
		Lasso	0.45	97	100	100	100	36	38	17	0.03	2.38
		Oracle	0.32	100	100	100	100	0	0	100	0	4
	200	Full	0.32	100	100	100	100	100	100	0	0	0
		ALasso	0.30	100	100	100	100	9	10	60	0	3.54
		Lasso	0.35	100	100	100	100	37	37	15	0	2.35
		Oracle	0.26	100	100	100	100	0	0	100	0	4
40%	50	Full	0.71	100	100	100	100	100	100	0	0	0
		ALasso	0.71	71	98	99	100	29	22	20	0.32	3.04
		Lasso	0.78	85	97	99	100	57	57	4	0.19	1.80
		Oracle	0.58	100	100	100	100	0	0	100	0	4
	100	Full	0.63	100	100	100	100	100	100	0	0	0
		ALasso	0.62	86	99	100	100	19	18	38	0.15	3.23
		Lasso	0.69	94	98	100	100	39	48	8	0.08	2.18
		Oracle	0.53	100	100	100	100	0	0	100	0	4
	200	Full	0.49	100	100	100	100	100	100	0	0	0
		ALasso	0.49	99	100	100	100	15	10	58	0.01	3.43
		Lasso	0.57	100	100	100	100	42	38	13	0	2.40
		Oracle	0.45	100	100	100	100	0	0	100	0	4
		Avg. SE	0.02									

Prop.: censoring proportion; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.13: Estimation of non-zero coefficients for heteroscedastic errors, $\gamma = 0.1$

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
Prop.	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
20%	50	Full	-0.03	0.22	0.36	-0.03	0.26	0.37	-0.09	0.26	0.36
		ALasso	-0.14	0.26	0.32	-0.08	0.27	0.35	-0.12	0.27	0.35
		Lasso	-0.13	0.22	0.29	-0.11	0.24	0.35	-0.19	0.28	0.36
		Oracle	-0.01	0.21	0.31	-0.02	0.25	0.31	-0.05	0.26	0.31
	100	Full	-0.01	0.16	0.22	-0.07	0.16	0.22	-0.11	0.18	0.22
		ALasso	-0.08	0.20	0.21	-0.10	0.17	0.22	-0.13	0.18	0.22
		Lasso	-0.09	0.19	0.20	-0.16	0.20	0.23	-0.20	0.19	0.23
		Oracle	0.00	0.18	0.20	-0.07	0.16	0.21	-0.11	0.17	0.20
	200	Full	-0.03	0.11	0.14	-0.06	0.11	0.14	-0.10	0.11	0.15
		ALasso	-0.05	0.12	0.14	-0.08	0.12	0.15	-0.11	0.11	0.15
		Lasso	-0.08	0.12	0.15	-0.13	0.12	0.16	-0.17	0.13	0.15
		Oracle	-0.02	0.11	0.13	-0.06	0.10	0.14	-0.10	0.10	0.14
40%	50	Full	-0.10	0.30	0.41	-0.15	0.31	0.42	-0.19	0.32	0.42
		ALasso	-0.19	0.29	0.35	-0.24	0.33	0.38	-0.24	0.32	0.39
		Lasso	-0.20	0.28	0.31	-0.24	0.32	0.36	-0.33	0.35	0.39
		Oracle	-0.07	0.24	0.35	-0.15	0.28	0.36	-0.19	0.28	0.36
	100	Full	-0.05	0.22	0.29	-0.15	0.21	0.29	-0.18	0.24	0.30
		ALasso	-0.11	0.25	0.26	-0.20	0.24	0.28	-0.21	0.24	0.30
		Lasso	-0.12	0.24	0.23	-0.23	0.23	0.28	-0.28	0.25	0.30
		Oracle	-0.05	0.21	0.27	-0.15	0.20	0.26	-0.19	0.23	0.28
	200	Full	-0.05	0.12	0.18	-0.11	0.14	0.19	-0.17	0.14	0.19
		ALasso	-0.09	0.15	0.17	-0.14	0.14	0.19	-0.19	0.14	0.20
		Lasso	-0.11	0.13	0.16	-0.19	0.17	0.20	-0.26	0.16	0.21
		Oracle	-0.04	0.12	0.18	-0.12	0.14	0.18	-0.17	0.13	0.19
		Avg. SE	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01

Prop.: censoring proportion; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.14: Variable selection results for heteroscedastic errors, $\gamma = 0.1$

				No. of times selected in 100 runs								
Prop.	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
20%	50	Full	0.57	100	100	100	100	100	100	0	0	0
		ALasso	0.54	77	98	100	100	22	28	31	0.25	3.17
		Lasso	0.58	91	99	100	100	48	48	12	0.10	2.24
		Oracle	0.45	100	100	100	100	0	0	100	0	4
	100	Full	0.43	100	100	100	100	100	100	0	0	0
		ALasso	0.39	90	100	100	100	13	19	58	0.10	3.42
		Lasso	0.48	93	100	100	100	37	39	16	0.07	2.39
		Oracle	0.33	100	100	100	100	0	0	100	0	4
	200	Full	0.32	100	100	100	100	100	100	0	0	0
		ALasso	0.30	100	100	100	100	8	7	67	0	3.60
		Lasso	0.37	100	100	100	100	31	26	27	0	2.64
		Oracle	0.26	100	100	100	100	0	0	100	0	4
40%	50	Full	0.72	100	100	100	100	100	100	0	0	0
		ALasso	0.73	67	94	100	100	20	21	25	0.39	3.07
		Lasso	0.81	72	95	99	100	51	53	3	0.34	1.98
		Oracle	0.60	100	100	100	100	0	0	100	0	4
	100	Full	0.62	100	100	100	100	100	100	0	0	0
		ALasso	0.63	81	98	100	100	22	15	35	0.21	3.23
		Lasso	0.69	87	100	100	100	47	50	7	0.13	2.06
		Oracle	0.53	100	100	100	100	0	0	100	0	4
	200	Full	0.48	100	100	100	100	100	100	0	0	0
		ALasso	0.49	95	100	100	100	11	10	60	0.05	3.46
		Lasso	0.58	99	100	100	100	44	37	15	0.01	2.32
		Oracle	0.43	100	100	100	100	0	0	100	0	4
		Avg. SE	0.02									

Prop.: censoring proportion; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.15: Estimation of non-zero coefficients for heteroscedastic errors, $\gamma = 0.2$

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
Prop.	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
20%	50	Full	-0.04	0.20	0.36	-0.04	0.24	0.36	-0.09	0.26	0.36
		ALasso	-0.15	0.25	0.32	-0.07	0.25	0.35	-0.13	0.29	0.35
		Lasso	-0.13	0.23	0.29	-0.11	0.25	0.34	-0.19	0.28	0.36
		Oracle	-0.02	0.21	0.31	-0.03	0.25	0.31	-0.06	0.27	0.31
	100	Full	-0.01	0.16	0.22	-0.06	0.17	0.22	-0.09	0.18	0.22
		ALasso	-0.08	0.19	0.22	-0.10	0.17	0.22	-0.12	0.17	0.22
		Lasso	-0.09	0.18	0.20	-0.15	0.19	0.23	-0.19	0.18	0.23
		Oracle	-0.01	0.17	0.20	-0.07	0.15	0.21	-0.10	0.17	0.20
	200	Full	-0.03	0.11	0.13	-0.06	0.11	0.14	-0.09	0.11	0.15
		ALasso	-0.05	0.13	0.14	-0.07	0.11	0.15	-0.10	0.11	0.15
		Lasso	-0.07	0.12	0.15	-0.11	0.12	0.16	-0.15	0.13	0.15
		Oracle	-0.02	0.12	0.13	-0.05	0.11	0.14	-0.09	0.10	0.14
40%	50	Full	-0.09	0.23	0.38	-0.11	0.24	0.38	-0.16	0.26	0.39
		ALasso	-0.21	0.27	0.32	-0.19	0.25	0.35	-0.24	0.27	0.38
		Lasso	-0.16	0.23	0.30	-0.23	0.26	0.34	-0.31	0.30	0.38
		Oracle	-0.12	0.20	0.33	-0.10	0.23	0.33	-0.17	0.21	0.35
	100	Full	-0.04	0.22	0.28	-0.12	0.20	0.28	-0.16	0.23	0.30
		ALasso	-0.11	0.24	0.25	-0.19	0.23	0.28	-0.21	0.22	0.30
		Lasso	-0.11	0.22	0.23	-0.23	0.22	0.27	-0.27	0.23	0.30
		Oracle	-0.04	0.19	0.26	-0.15	0.19	0.26	-0.18	0.21	0.28
	200	Full	-0.06	0.12	0.18	-0.12	0.13	0.18	-0.17	0.14	0.19
		ALasso	-0.10	0.14	0.17	-0.16	0.13	0.19	-0.20	0.13	0.20
		Lasso	-0.12	0.12	0.16	-0.20	0.15	0.20	-0.26	0.15	0.21
		Oracle	-0.05	0.12	0.17	-0.12	0.13	0.18	-0.17	0.13	0.19
		Avg. SE	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01

Prop.: censoring proportion; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.16: Variable selection results for heteroscedastic errors, $\gamma = 0.2$

				No. of times selected in 100 runs								
Prop.	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
20%	50	Full	0.58	100	100	100	100	100	100	0	0	0
		ALasso	0.56	79	100	100	100	15	25	37	0.21	3.30
		Lasso	0.60	89	100	100	100	47	45	10	0.11	2.20
		Oracle	0.46	100	100	100	100	0	0	100	0	4
	100	Full	0.42	100	100	100	100	100	100	0	0	0
		ALasso	0.39	91	100	100	100	12	17	59	0.09	3.42
		Lasso	0.47	96	100	100	100	35	41	15	0.04	2.33
		Oracle	0.33	100	100	100	100	0	0	100	0	4
	200	Full	0.32	100	100	100	100	100	100	0	0	0
		ALasso	0.30	99	100	100	100	5	8	67	0.01	3.64
		Lasso	0.36	100	100	100	100	41	38	17	0	2.39
		Oracle	0.26	100	100	100	100	0	0	100	0	4
40%	50	Full	0.70	100	100	100	100	100	100	0	0	0
		ALasso	0.72	64	100	100	100	22	16	26	0.36	3.17
		Lasso	0.81	82	98	99	100	47	43	7	0.21	2.03
		Oracle	0.60	100	100	100	100	0	0	100	0	4
	100	Full	0.62	100	100	100	100	100	100	0	0	0
		ALasso	0.64	84	98	100	100	18	22	39	0.18	3.27
		Lasso	0.71	89	100	100	100	45	48	11	0.11	2.13
		Oracle	0.57	100	100	100	100	0	0	100	0	4
	200	Full	0.51	100	100	100	100	100	100	0	0	0
		ALasso	0.53	98	100	100	100	10	13	64	0.02	3.49
		Lasso	0.62	100	100	100	100	40	35	22	0	2.49
		Oracle	0.47	100	100	100	100	0	0	100	0	4
		Avg. SE	0.02									

Prop.: censoring proportion; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.17: Estimation of non-zero coefficients for 20% Censoring and cluster size $K = 2$

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
ρ	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
.2	50	Full	-0.01	0.16	0.20	-0.10	0.16	0.20	-0.09	0.16	0.21
		ALasso	-0.06	0.17	0.20	-0.12	0.17	0.21	-0.11	0.16	0.21
		Lasso	-0.09	0.17	0.18	-0.17	0.17	0.20	-0.18	0.17	0.21
		Oracle	-0.01	0.15	0.19	-0.09	0.16	0.19	-0.09	0.16	0.20
	100	Full	-0.03	0.11	0.13	-0.05	0.11	0.13	-0.06	0.12	0.14
		ALasso	-0.06	0.12	0.14	-0.07	0.12	0.13	-0.08	0.12	0.14
		Lasso	-0.08	0.12	0.13	-0.10	0.12	0.13	-0.13	0.13	0.14
		Oracle	-0.03	0.10	0.13	-0.05	0.12	0.13	-0.06	0.11	0.13
	200	Full	-0.03	0.07	0.09	-0.06	0.08	0.09	-0.07	0.07	0.09
		ALasso	-0.05	0.07	0.09	-0.08	0.08	0.09	-0.07	0.08	0.09
		Lasso	-0.07	0.07	0.09	-0.11	0.09	0.09	-0.12	0.08	0.10
		Oracle	-0.03	0.07	0.09	-0.06	0.08	0.09	-0.07	0.07	0.09
.5	50	Full	-0.01	0.17	0.20	-0.10	0.15	0.20	-0.08	0.15	0.21
		ALasso	-0.06	0.18	0.20	-0.13	0.15	0.20	-0.11	0.15	0.20
		Lasso	-0.08	0.17	0.19	-0.16	0.15	0.20	-0.16	0.16	0.20
		Oracle	-0.02	0.16	0.19	-0.10	0.16	0.19	-0.08	0.14	0.19
	100	Full	-0.02	0.11	0.13	-0.05	0.11	0.13	-0.07	0.10	0.14
		ALasso	-0.05	0.12	0.13	-0.07	0.11	0.13	-0.07	0.10	0.13
		Lasso	-0.07	0.11	0.13	-0.11	0.11	0.13	-0.13	0.12	0.13
		Oracle	-0.03	0.11	0.12	-0.06	0.11	0.13	-0.06	0.10	0.13
	200	Full	-0.03	0.08	0.09	-0.06	0.08	0.09	-0.07	0.08	0.09
		ALasso	-0.05	0.08	0.09	-0.07	0.08	0.09	-0.07	0.09	0.09
		Lasso	-0.07	0.08	0.09	-0.11	0.08	0.09	-0.12	0.09	0.09
		Oracle	-0.02	0.08	0.08	-0.06	0.08	0.09	-0.06	0.09	0.09
		Avg. SE	0.01	0.01	<.005	0.01	0.01	<.005	0.01	0.01	<.005

ρ: intracluster covariance; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.18: Variable selection results for 20% Censoring and cluster size $K = 2$

				No. of times selected in 100 runs								
ρ	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
.2	50	Full	0.38	100	100	100	100	100	100	0	0	0
		ALasso	0.35	96	100	100	100	19	18	56	0.04	3.41
		Lasso	0.43	97	100	100	100	39	42	13	0.03	2.37
		Oracle	0.30	100	100	100	100	0	0	100	0	4
	100	Full	0.30	100	100	100	100	100	100	0	0	0
		ALasso	0.28	99	100	100	100	14	16	61	0.01	3.48
		Lasso	0.33	100	100	100	100	43	40	15	0	2.35
		Oracle	0.24	100	100	100	100	0	0	100	0	4
	200	Full	0.23	100	100	100	100	100	100	0	0	0
		ALasso	0.22	100	100	100	100	6	10	77	0	3.66
		Lasso	0.26	100	100	100	100	47	45	9	0	2.28
		Oracle	0.20	100	100	100	100	0	0	100	0	4
.5	50	Full	0.38	100	100	100	100	100	100	0	0	0
		ALasso	0.35	96	100	100	100	18	15	58	0.04	3.41
		Lasso	0.41	96	100	100	100	42	41	18	0.04	2.36
		Oracle	0.31	100	100	100	100	0	0	100	0	4
	100	Full	0.29	100	100	100	100	100	100	0	0	0
		ALasso	0.26	100	100	100	100	10	15	68	0	3.57
		Lasso	0.32	100	100	100	100	31	38	14	0	2.44
		Oracle	0.24	100	100	100	100	0	0	100	0	4
	200	Full	0.22	100	100	100	100	100	100	0	0	0
		ALasso	0.21	100	100	100	100	5	11	72	0	3.65
		Lasso	0.26	100	100	100	100	42	39	25	0	2.49
		Oracle	0.19	100	100	100	100	0	0	100	0	4
		Avg. SE	0.02									

ρ: intracluster covariance; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.19: Estimation of non-zero coefficients for 40% Censoring and cluster size $K = 2$

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
ρ	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
.2	50	Full	-0.05	0.20	0.28	-0.13	0.20	0.28	-0.18	0.19	0.29
		ALasso	-0.11	0.21	0.25	-0.19	0.19	0.27	-0.21	0.20	0.29
		Lasso	-0.13	0.21	0.23	-0.22	0.18	0.25	-0.29	0.22	0.28
		Oracle	-0.05	0.18	0.26	-0.15	0.18	0.25	-0.18	0.19	0.27
	100	Full	-0.06	0.12	0.17	-0.10	0.14	0.17	-0.13	0.15	0.18
		ALasso	-0.10	0.14	0.16	-0.13	0.13	0.18	-0.15	0.15	0.18
		Lasso	-0.11	0.12	0.15	-0.16	0.13	0.17	-0.20	0.14	0.18
		Oracle	-0.06	0.13	0.17	-0.11	0.12	0.17	-0.13	0.16	0.18
	200	Full	-0.06	0.09	0.12	-0.11	0.08	0.12	-0.15	0.10	0.13
		ALasso	-0.08	0.09	0.12	-0.13	0.09	0.13	-0.17	0.10	0.13
		Lasso	-0.10	0.09	0.11	-0.16	0.09	0.12	-0.22	0.11	0.13
		Oracle	-0.06	0.09	0.12	-0.11	0.08	0.12	-0.15	0.10	0.13
.5	50	Full	-0.03	0.18	0.27	-0.15	0.18	0.28	-0.16	0.20	0.29
		ALasso	-0.12	0.21	0.24	-0.19	0.19	0.27	-0.19	0.20	0.29
		Lasso	-0.12	0.20	0.23	-0.22	0.19	0.25	-0.25	0.23	0.27
		Oracle	-0.04	0.18	0.26	-0.14	0.18	0.25	-0.16	0.18	0.27
	100	Full	-0.06	0.13	0.17	-0.12	0.13	0.17	-0.13	0.16	0.19
		ALasso	-0.10	0.14	0.17	-0.14	0.13	0.18	-0.14	0.16	0.19
		Lasso	-0.12	0.13	0.16	-0.19	0.14	0.17	-0.22	0.16	0.18
		Oracle	-0.06	0.12	0.16	-0.13	0.12	0.17	-0.13	0.16	0.18
	200	Full	-0.05	0.09	0.12	-0.12	0.09	0.13	-0.14	0.10	0.13
		ALasso	-0.07	0.10	0.12	-0.13	0.09	0.13	-0.15	0.10	0.13
		Lasso	-0.09	0.09	0.11	-0.17	0.09	0.13	-0.21	0.11	0.13
		Oracle	-0.05	0.09	0.11	-0.11	0.09	0.13	-0.14	0.09	0.13
		Avg. SE	0.01	0.01	<.005	0.01	0.01	<.005	0.02	0.01	<.005

ρ: intracluster covariance; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.20: Variable selection results for 40% Censoring and cluster size $K = 2$

				No. of times selected in 100 runs								
ρ	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
.2	50	Full	0.52	100	100	100	100	100	100	0	0	0
		ALasso	0.54	88	99	100	100	27	14	39	0.13	3.28
		Lasso	0.62	91	99	100	100	55	48	4	0.10	2.09
		Oracle	0.46	100	100	100	100	0	0	100	0	4
	100	Full	0.44	100	100	100	100	100	100	0	0	0
		ALasso	0.44	96	100	100	100	17	19	51	0.04	3.37
		Lasso	0.50	99	100	100	100	50	50	14	0.01	2.09
		Oracle	0.40	100	100	100	100	0	0	100	0	4
	200	Full	0.40	100	100	100	100	100	100	0	0	0
		ALasso	0.41	100	100	100	100	10	10	69	0	3.62
		Lasso	0.47	100	100	100	100	44	48	19	0	2.26
		Oracle	0.38	100	100	100	100	0	0	100	0	4
.5	50	Full	0.53	100	100	100	100	100	100	0	0	0
		ALasso	0.53	89	99	100	100	17	12	48	0.12	3.46
		Lasso	0.61	90	99	100	100	47	53	10	0.11	2.10
		Oracle	0.45	100	100	100	100	0	0	100	0	4
	100	Full	0.45	100	100	100	100	100	100	0	0	0
		ALasso	0.45	97	100	100	100	17	19	52	0.03	3.37
		Lasso	0.54	99	100	100	100	43	47	15	0.01	2.24
		Oracle	0.41	100	100	100	100	0	0	100	0	4
	200	Full	0.38	100	100	100	100	100	100	0	0	0
		ALasso	0.38	100	100	100	100	10	11	69	0	3.56
		Lasso	0.46	100	100	100	100	43	50	18	0	2.19
		Oracle	0.35	100	100	100	100	0	0	100	0	4
		Avg. SE	0.01									

ρ: intracluster covariance; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.21: Estimation of non-zero coefficients for 20% Censoring and cluster size $K = 5$

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
ρ	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
.2	50	Full	-0.03	0.11	0.11	-0.04	0.11	0.12	-0.10	0.11	0.12
		ALasso	-0.06	0.12	0.12	-0.06	0.11	0.12	-0.10	0.11	0.12
		Lasso	-0.07	0.11	0.11	-0.10	0.11	0.11	-0.15	0.12	0.12
		Oracle	-0.03	0.11	0.11	-0.04	0.10	0.11	-0.09	0.11	0.11
	100	Full	-0.02	0.06	0.08	-0.05	0.06	0.08	-0.07	0.08	0.08
		ALasso	-0.04	0.07	0.08	-0.06	0.07	0.08	-0.08	0.08	0.08
		Lasso	-0.06	0.07	0.08	-0.09	0.07	0.08	-0.11	0.08	0.08
		Oracle	-0.03	0.07	0.08	-0.05	0.07	0.08	-0.07	0.08	0.08
	200	Full	-0.03	0.05	0.05	-0.06	0.05	0.06	-0.08	0.05	0.06
		ALasso	-0.04	0.05	0.06	-0.06	0.05	0.06	-0.09	0.05	0.06
		Lasso	-0.06	0.05	0.05	-0.09	0.06	0.06	-0.11	0.06	0.06
		Oracle	-0.03	0.05	0.05	-0.06	0.05	0.05	-0.08	0.05	0.06
.5	50	Full	-0.03	0.11	0.11	-0.04	0.10	0.12	-0.09	0.11	0.12
		ALasso	-0.06	0.12	0.12	-0.05	0.10	0.12	-0.10	0.11	0.12
		Lasso	-0.08	0.12	0.11	-0.09	0.11	0.11	-0.15	0.12	0.12
		Oracle	-0.04	0.12	0.11	-0.03	0.10	0.11	-0.09	0.10	0.11
	100	Full	-0.03	0.07	0.08	-0.06	0.07	0.08	-0.07	0.09	0.08
		ALasso	-0.04	0.07	0.08	-0.07	0.07	0.08	-0.08	0.09	0.08
		Lasso	-0.07	0.07	0.08	-0.10	0.07	0.08	-0.12	0.09	0.08
		Oracle	-0.03	0.06	0.08	-0.06	0.07	0.08	-0.08	0.08	0.08
	200	Full	-0.03	0.05	0.05	-0.06	0.05	0.06	-0.08	0.05	0.06
		ALasso	-0.04	0.05	0.05	-0.07	0.05	0.06	-0.09	0.05	0.06
		Lasso	-0.05	0.05	0.05	-0.09	0.05	0.06	-0.12	0.05	0.06
		Oracle	-0.03	0.05	0.05	-0.06	0.05	0.05	-0.08	0.05	0.06
		Avg. SE	0.01	0.01	<.005	0.01	0.01	<.005	0.01	0.01	<.005

ρ: intracluster covariance; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.22: Variable selection results for 20% Censoring and cluster size $K = 5$

				No. of times selected in 100 runs								
ρ	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
.2	50	Full	0.27	100	100	100	100	100	100	0	0	0
		ALasso	0.25	100	100	100	100	12	11	69	0	3.59
		Lasso	0.30	100	100	100	100	49	34	19	0	2.41
		Oracle	0.22	100	100	100	100	0	0	100	0	4
	100	Full	0.21	100	100	100	100	100	100	0	0	0
		ALasso	0.20	100	100	100	100	3	5	86	0	3.82
		Lasso	0.24	100	100	100	100	39	41	23	0	2.41
		Oracle	0.18	100	100	100	100	0	0	100	0	4
	200	Full	0.19	100	100	100	100	100	100	0	0	0
		ALasso	0.18	100	100	100	100	7	3	83	0	3.80
		Lasso	0.22	100	100	100	100	44	34	23	0	2.51
		Oracle	0.17	100	100	100	100	0	0	100	0	4
.5	50	Full	0.27	100	100	100	100	100	100	0	0	0
		ALasso	0.25	100	100	100	100	10	10	70	0	3.62
		Lasso	0.30	100	100	100	100	48	42	15	0	2.31
		Oracle	0.22	100	100	100	100	0	0	100	0	4
	100	Full	0.22	100	100	100	100	100	100	0	0	0
		ALasso	0.21	100	100	100	100	4	9	79	0	3.75
		Lasso	0.26	100	100	100	100	24	36	29	0	2.76
		Oracle	0.20	100	100	100	100	0	0	100	0	4
	200	Full	0.19	100	100	100	100	100	100	0	0	0
		ALasso	0.18	100	100	100	100	4	3	87	0	3.85
		Lasso	0.22	100	100	100	100	32	31	28	0	2.74
		Oracle	0.17	100	100	100	100	0	0	100	0	4
		Avg. SE	0.02									

ρ: intracluster covariance; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.

Table B.23: Estimation of non-zero coefficients for 40% Censoring and cluster size $K = 5$

			$\beta_1 = 0.5$			$\beta_2 = 1$			$\beta_3 = 1.5$		
ρ	n	Methods	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
.2	50	Full	-0.05	0.12	0.15	-0.09	0.10	0.15	-0.18	0.13	0.16
		ALasso	-0.09	0.14	0.15	-0.12	0.11	0.16	-0.20	0.14	0.17
		Lasso	-0.10	0.12	0.14	-0.15	0.11	0.15	-0.25	0.14	0.16
		Oracle	-0.05	0.13	0.15	-0.09	0.11	0.15	-0.19	0.13	0.16
	100	Full	-0.07	0.08	0.11	-0.10	0.09	0.11	-0.15	0.09	0.12
		ALasso	-0.09	0.09	0.11	-0.12	0.09	0.11	-0.16	0.09	0.12
		Lasso	-0.11	0.08	0.10	-0.16	0.10	0.11	-0.21	0.10	0.11
		Oracle	-0.06	0.08	0.11	-0.11	0.09	0.11	-0.15	0.09	0.12
	200	Full	-0.06	0.05	0.07	-0.11	0.06	0.08	-0.16	0.07	0.08
		ALasso	-0.07	0.06	0.08	-0.12	0.06	0.08	-0.16	0.07	0.08
		Lasso	-0.09	0.05	0.07	-0.15	0.06	0.08	-0.20	0.07	0.08
		Oracle	-0.06	0.06	0.07	-0.12	0.06	0.08	-0.15	0.07	0.08
.5	50	Full	-0.06	0.11	0.15	-0.10	0.12	0.15	-0.18	0.12	0.17
		ALasso	-0.10	0.13	0.15	-0.12	0.13	0.16	-0.21	0.13	0.17
		Lasso	-0.12	0.12	0.14	-0.16	0.13	0.15	-0.26	0.14	0.17
		Oracle	-0.06	0.11	0.15	-0.10	0.13	0.15	-0.19	0.12	0.16
	100	Full	-0.05	0.08	0.11	-0.11	0.09	0.11	-0.14	0.09	0.12
		ALasso	-0.08	0.08	0.11	-0.12	0.08	0.12	-0.16	0.09	0.12
		Lasso	-0.10	0.08	0.11	-0.16	0.09	0.11	-0.20	0.10	0.12
		Oracle	-0.06	0.08	0.11	-0.11	0.08	0.11	-0.15	0.09	0.12
	200	Full	-0.06	0.05	0.07	-0.12	0.06	0.08	-0.17	0.06	0.08
		ALasso	-0.07	0.06	0.08	-0.13	0.06	0.08	-0.17	0.06	0.08
		Lasso	-0.09	0.06	0.07	-0.16	0.07	0.08	-0.21	0.07	0.08
		Oracle	-0.06	0.06	0.07	-0.12	0.05	0.08	-0.16	0.07	0.08
		Avg. SE	0.01	0.01	<.005	0.01	0.01	<.005	0.01	0.01	<.005

ρ: intracluster covariance; Bias: average bias over 100 runs; SD: sample standard deviation of estimates; SE: mean of estimated standard errors computed based on 500 bootstraps per run; Avg. SE: estimated standard error averaged over all rows of table.

Table B.24: Variable selection results for 40% Censoring and cluster size $K = 5$

				No. of times selected in 100 runs								
ρ	n	Methods	MAD ME	β_1	β_2	β_3	β_4	β_5	β_6	SF	Inc.	Cor.
.2	50	Full	0.41	100	100	100	100	100	100	0	0	0
		ALasso	0.42	96	100	100	100	16	11	63	0.04	3.54
		Lasso	0.48	98	100	100	100	52	45	9	0.02	2.12
		Oracle	0.38	100	100	100	100	0	0	100	0	4
	100	Full	0.38	100	100	100	100	100	100	0	0	0
		ALasso	0.38	100	100	100	100	10	13	67	0	3.62
		Lasso	0.45	100	100	100	100	43	46	16	0	2.32
		Oracle	0.36	100	100	100	100	0	0	100	0	4
	200	Full	0.36	100	100	100	100	100	100	0	0	0
		ALasso	0.36	100	100	100	100	7	5	73	0	3.64
		Lasso	0.42	100	100	100	100	45	45	14	0	2.19
		Oracle	0.35	100	100	100	100	0	0	100	0	4
.5	50	Full	0.42	100	100	100	100	100	100	0	0	0
		ALasso	0.43	99	100	100	100	12	14	63	0.01	3.45
		Lasso	0.50	100	100	100	100	47	48	12	0	2.13
		Oracle	0.37	100	100	100	100	0	0	100	0	4
	100	Full	0.38	100	100	100	100	100	100	0	0	0
		ALasso	0.39	100	100	100	100	13	10	72	0	3.56
		Lasso	0.44	100	100	100	100	44	39	16	0	2.29
		Oracle	0.37	100	100	100	100	0	0	100	0	4
	200	Full	0.38	100	100	100	100	100	100	0	0	0
		ALasso	0.38	100	100	100	100	9	9	71	0	3.63
		Lasso	0.43	100	100	100	100	35	50	13	0	2.22
		Oracle	0.36	100	100	100	100	0	0	100	0	4
		Avg. SE	0.01									

ρ: intracluster covariance; MAD ME: average of mean absolute deviation of model error across 100 runs; SF: number of times the true model is chosen among 100 runs; Inc. and Cor.: mean number of incorrect and correct 0's selected, respectively; Avg. SE: estimated standard error averaged over all rows of table.