VARIABLE SELECTION FOR CENSORED QUANTILE REGRESION

Huixia Judy Wang[1], Jianhui Zhou[2] and Yi Li[3]

North Carolina State University[1]*, University of Virginia*[2] *and Harvard University*[4]

Abstract: Quantile regression has emerged as a powerful tool in survival analysis as it directly links the quantiles of patients' survival times to their demographic and genomic profiles, facilitating the identification of important prognostic factors. In view of limited work on variable selection in the context, we develop a new adaptive-lasso-based variable selection procedure for quantile regression with censored outcomes. To account for random censoring for data with multivariate covariates, we employ the ideas of redistribution-of-mass and effective dimension reduction. Asymptotically our procedure enjoys the model selection consistency, that is, identifying the true model with probability tending to one. Moreover, as opposed to the existing methods, our new proposal requires fewer assumptions, leading to more accurate variable selection. The analysis of a real cancer clinical trial demonstrates that our procedure can identify and distinguish important factors associated with patient subpopulations characterized by short or long survivals, which is of particular interest to oncologists.

Key words and phrases: Conditional Kaplan-Meier, dimension reduction, kernel, quantile regression, survival analysis, variable selection.

1. Introduction

Quantile regression, as a valuable alternative to the commonly used Cox proportional hazard model and accelerated failure time (AFT) model (Koenker and Geling, 2001; Portnoy, 2003), directly links the quantiles of subjects' survival times to their demographic and genomic profiles, and thus can facilitate the identification of important prognostic factors. Direct applications of this model lie in, for example, cancer studies, where physicians are often interested in identifying effective treatments for more severe cases (with shorter survival times). As might be expected, treatments often cause different impacts among patients that fall within the upper or lower quantiles of the survival distribution. It is well known that both Cox and AFT models restrict the covariates to affect only the location but not the shape of the survival distribution, and thus may overlook interesting forms of heterogeneity. For instance, these two models do not permit the treatment effect to be positive for severe cases while negative for the other cases. In contrast, quantile regression offers a convenient approach to capture the variation caused by heterogeneities by allowing the covariates to exhibit different impacts at different tails of the survival distribution.

Current literature on quantile regression for censored survival outcomes often focus on coefficient estimation (Portnoy, 2003; Peng and Huang, 2008; Huang, 2009; Wang and Wang, 2009), but little effort has been devoted to variable selection. In contrast, various penalization-based variable selection methods have been developed for Cox and AFT models, for instance, Huang et al. (2006), Zhang and Lu (2007), Wang et al. (2008), Engler and Li (2009), among others.

The nature of the existing estimating procedures for censored quantile regression prohibits the direct usage of the popular penalization methods. The point estimation methods of Portnoy (2003) and Peng and Huang (2008) require fitting an entire quantile process, as the estimation at an upper quantile depends on the estimations at all the lower quantiles. Therefore, even if our main interest was to identify variables with strong impacts on the median survival, we would have to estimate and select variables for all the lower quantiles as well. This imposes both computational and theoretical challenges. Alternative estimation methods such as those proposed by Subramanian (2002), Wang and Wang (2009) and Leng and Tong (2011) rely on the kernel-smoothing estimation, and thus are practically feasible only for data with few covariates. Recently, Shows, Lu and Zhang (2010) developed a variable selection approach for censored median regression by using the inverse-probability-weighting scheme of Bang and Tsiatis (2002). However, the method requires the restrictive unconditional independence assumption between survival and censoring times. Moreover, the procedure uses information of only uncensored observations, leading to an efficiency loss in the estimation.

We develop a new and flexible variable selection method based on the adaptive-lasso penalization for censored quantile regression. Our work advances the field in the following three significant ways. First, our method adapts the redistribution-of-mass idea of Efron (1967) to account for the censoring in quantile regression. Different from existing estimation procedures (Portnoy, 2003; Wang and Wang, 2009), the proposed method estimates the masses for redistribution by using the conditional Kaplan-Meier estimation on a reduced data space. Therefore, the new method is more flexible and is able to accommodate high-dimensional covariates. Secondly, our variable selection procedure enjoys computation readiness and requires much fewer stringent assumptions than those in the literature. As a result, our procedure is model selection consistent under mild conditions and gives better finite sample performance than the existing methods. Lastly, the proposed method is able to capture important heterogeneities in the survival population, and to identify effective biomarkers that have impacts at different tails of the survival distribution. Such results will be of particular interest to physicians who are keen on designing effective treatments for targeted patient subpopulations, often characterized by short survivals.

2. Variable Selection for Censored Quantile Regression

2.1. Model setup

Denote T_i as the uncensored survival outcome, $\mathbf{x}_i$ as the observable p-dimensional covariates, and C_i as the censoring variable. Consider the following quantile regression model

$$T_i = \mathbf{x}_i^T \boldsymbol{\beta}_0(\tau) + e_i(\tau), \tag{2.1}$$

where $0 < \tau < 1$ is the given quantile level of interest, $\boldsymbol{\beta}_0(\tau)$ is the p-dimensional unknown quantile coefficient vector, and $e_i(\tau)$ is the random error whose τth conditional quantile given $\mathbf{x}_i$ equals zero. Without loss of generality, we assume that the first element of $\mathbf{x}_i$ is 1 corresponding to the intercept. In practice, we only observe $(\mathbf{x}_i, Y_i, \delta_i)$, where $Y_i = \min(T_i, C_i)$ is the observed response variable and $\delta_i = I(T_i \le C_i)$ is the censoring indicator. Our main objective in this paper is to select important predictors that have nonzero effect on the τth conditional quantile of T in the parametric quantile regression model (2.1).

2.2. Variable selection via redistribution-of-mass

We first briefly review the idea of redistribution-of-mass in censored quantile regression. For censored data without covariates, that is, $p = 1$ in (2.1), Efron (1967) proposed a simple algorithm for deriving the Kaplan-Meier estimator by redistributing the mass of each censored observation uniformly to observations on the right. In regression setup, this means to redistribute the probability masses $P(T_i > C_i | C_i, \mathbf{x}_i)$ of censored cases to observations on the right.

Let $F_0(t|\mathbf{x}) = P(T < t|\mathbf{x})$ denote the conditional distribution function of T given $\mathbf{x}$. Define $\pi_{0i} = F_0(C_i|\mathbf{x}_i)$ as the conditional probability for the ith subject not to be censored. Consider an ideal scenario where π_{0i} are known. Then $\beta_0(\tau)$ can be estimated by minimizing the following weighted quantile objective function with respect to $\boldsymbol{\beta}$,

$$L(\boldsymbol{\beta}, w_0) \quad = \quad \sum_{i=1}^{n} \left\{ w_{0i}\rho_\tau(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + (1 - w_{0i})\rho_\tau(Y^{+\infty} - \mathbf{x}_i^T\boldsymbol{\beta}) \right\}, \quad (2.2)$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$, $Y^{+\infty}$ is any value sufficiently large to exceed $\mathbf{x}_i^T\boldsymbol{\beta}_0(\tau)$ for all i, and

$$w_{0i} = \begin{cases} 1 & \delta_i = 1 \\ 0 & \delta_i = 0 \text{ and } \pi_{0i} > \tau \\ (\tau - \pi_{0i})/(1 - \pi_{0i}) & \delta_i = 0 \text{ and } \pi_{0i} \le \tau \end{cases}. \quad (2.3)$$

It can be shown that a subgradient of the above weighted objective function,

$$nM_n(\boldsymbol{\beta}, w_0) = \sum_{i=1}^{n} \mathbf{x}_i \left\{ 1 - w_{0i} I(Y_i < \mathbf{x}_i^T\boldsymbol{\beta}) \right\}, \quad (2.4)$$

is an unbiased estimating function of $\boldsymbol{\beta}_0(\tau)$. Therefore, minimizing $L(\boldsymbol{\beta}, w_0)$ with respect to $\boldsymbol{\beta}$ leads to a consistent estimator of $\boldsymbol{\beta}_0(\tau)$. More explanation of the intuition behind the above weighting scheme in (2.2) can be found in Wang and Wang (2009), and Portnoy and Lin (2010).

To select variables, we consider the following penalized objective function

$$L_{AL}(\boldsymbol{\beta}, w_0) = L(\boldsymbol{\beta}, w_0) + \lambda_n \sum_{j=1}^{p} \nu_j |\beta_j|, \quad (2.5)$$

where λ_n is the positive penalization parameter and ν_j are the adaptive weights. The type of adaptive lasso penalization was first proposed by Zou (2006) for least squares regression and later extended to quantile regression for uncensored data by Wu and Liu (2009) and to Cox's model by Zhang and Lu (2007). The adaptive lasso assigns heavier penalties to the potentially irrelevant variables, so the corresponding effects are shrunk more towards zero. This approach leads to a sparse coefficient estimation, and thus provides a convenient way to conduct model fitting and variable selection simultaneously. The choice of the adaptive weights ν_j is explained in (2.8) of Section 2.3.

2.3. Estimation of the redistributed mass

In practice, the masses for redistribution, $1-\pi_{0i} = 1-F_0(C_i|\mathbf{x}_i)$, are unknown and a variety of attempts have been made to estimate them. For example, Portnoy (2003) proposed to estimate π_{0i} through fitting an entire quantile regression process under the global linearity assumption of the conditional quantile functions. McKeague et al. (2001) suggested to fit a semiparametric regression model such as Cox proportional hazards model to obtain an approximation of π_{0i}. Lindgren (1997), Subramanian (2002) and Wang and Wang (2009) employed a fully nonparametric approach based on the conditional Kaplan-Meier estimator of $F_0(\cdot|\mathbf{x})$. Though flexible, this nonparametric approach is only feasible when the covariate dimension is small because of the curse of dimensionality. The theoretical results in Subramanian (2002) and Wang and Wang (2009) were developed only for cases with univariate covariate.

We propose an index-based procedure to obtain a nonparametric estimation of π_{0i} for multivariate covariates. The main idea is to summarize the regression information contained in $\mathbf{x}$ by indices through dimension reduction. Specifically, we adopt a global dimension reduction (DR) formulation:

$$T_i \perp\!\!\!\perp x_i|(\mathbf{x}_i^T\boldsymbol{\gamma}_1, \cdots, \mathbf{x}_i^T\boldsymbol{\gamma}_q), \tag{2.6}$$

where $\perp\!\!\!\perp$ stands for independence. This formulation stipulates that the dependence of T_i on the p-dimensional $\mathbf{x}_i$ only comes from q indices, $\mathbf{z}_{i,1} = \mathbf{x}_i^T\boldsymbol{\gamma}_1, \cdots, \mathbf{z}_{i,q} = \mathbf{x}_i^T\boldsymbol{\gamma}_q$, where q is often much smaller than p in practice. For randomly censored data, $\boldsymbol{\gamma}_j$, often referred to as effective dimension reduction (EDR) directions, can be estimated by using the sliced inverse regression (SIR) method of Li et al. (1999) or the hazard-function-based minimum average variance estimation (MAVE) method of Xia et al. (2010). Under some regularity assumptions, both methods lead to estimators $\widehat{\boldsymbol{\gamma}}_j$ that are root-n consistent to $\boldsymbol{\gamma}_{0,j}$, where $\boldsymbol{\gamma}_{0,j} \in R^p$, $j = 1, \cdots, q$, is a set of EDR directions. Hereafter, we denote the estimated indices as $\widehat{\mathbf{z}}_i = (\widehat{z}_{i,1}, \cdots, \widehat{z}_{i,q})$ with $\widehat{z}_{i,j} = \mathbf{x}_i^T\widehat{\boldsymbol{\gamma}}_j$, and denote $\mathbf{z}_{0i} = (z_{0i,1}, \cdots, z_{0i,q})$ with $z_{0i,j} = \mathbf{x}_i^T\boldsymbol{\gamma}_{0,j}$, $j = 1, \cdots, q$.

Under model (2.6), we have $F_0(t|\mathbf{x}_i) = F_0(t|\mathbf{z}_{0i})$ for any t and i. We then proceed to use Beran's local Kaplan-Meier estimator $\widehat{F}(\cdot|\mathbf{z})$ (Beran,1981) to estimate $F(\cdot|\mathbf{z})$. Specifically,

$$\widehat{F}(t|\mathbf{z}) = 1 - \prod_{j=1}^{n} \left\{ 1 - \frac{B_{nj}(\mathbf{z})}{\sum_{k=1}^{n} I(Y_k \geq Y_j) B_{nk}(\mathbf{z})} \right\}^{\eta_j(t)}, \tag{2.7}$$

where $\eta_j(t) = I(Y_j \leq t, \delta_j = 1)$, $B_{nk}(\mathbf{z}) = K_q\left(\frac{\mathbf{z}-\mathbf{z}_k}{h_n}\right) / \sum_{i=1}^{n} K_q\left(\frac{\mathbf{z}-\mathbf{z}_i}{h_n}\right)$, h_n is the bandwidth, and $K_q(\frac{\mathbf{z}-\mathbf{z}_i}{h_n}) = K_q(\frac{z_1-z_{i,1}}{h_n}, \ldots, \frac{z_q-z_{i,q}}{h_n})$. In this paper, we adopt the commonly used product kernel function $K_q(u_1, \ldots, u_q) = \prod_{i=1}^{q} K(u_i)$, where $K(\cdot)$ is a univariate kernel function. Here we opt for Beran's estimator as it is nonparametric and thus flexible, avoiding estimating the entire quantile process assuming global linear models as in Portnoy (2003). Therefore, π_{0i} can be estimated by $\widehat{\pi}_i = \widehat{F}(C_i|\widehat{\mathbf{z}}_i)$, the nonparametric estimation of the conditional distribution of T given the indices with a reduced dimension q.

For variable selection, we define the proposed penalized estimator $\widehat{\boldsymbol{\beta}}(\tau)$ for $\boldsymbol{\beta}_0(\tau)$ in model (2.1) as the minimizer of the following penalized objective function

$$\begin{aligned} L_{AL}(\boldsymbol{\beta}, \widehat{w}) &= \sum_{i=1}^{n} \left\{ \widehat{w}_i \rho_\tau(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + (1 - \widehat{w}_i) \rho_\tau(Y^{+\infty} - \mathbf{x}_i^T \boldsymbol{\beta}) \right\} \\ &\quad + \lambda_n \sum_{j=1}^{p} \nu_j |\beta_j|, \end{aligned} \tag{2.8}$$

where $\widehat{w}_i$ are the weights for redistribution-of-mass as defined by replacing π_{0i} with $\widehat{\pi}_i$ in (2.3), and ν_j are the adaptive weights. We let $\nu_j = |\widetilde{\beta}_j(\tau)|^{-r}$, where $\widetilde{\beta}_j(\tau)$ is the jth element of the initial consistent estimator of $\boldsymbol{\beta}(\tau)$. We define the initial estimator $\widetilde{\boldsymbol{\beta}}(\tau)$ as the unpenalized estimator, that is, the minimizer of $L_{AL}(\boldsymbol{\beta}, \widehat{w})$ with $\lambda_n = 0$. In our implementation, we choose $r = 2$.

We further stress that our aim of the paper is to select the important predictors that have nonzero effect on the τth conditional quantile of T in the parametric quantile regression model (2.1). This variable selection is for the local quantiles of T and thus is different from the global nonparametric dimension reduction in the formulation (2.6). In addition, the proposed penalization procedure does not require the unconditional independence between the survival times T_i and censoring times C_i, which is a significant improvement compared to Shows et al. (2010).

2.4. Computation and tuning

The proposed procedure requires choosing the bandwidth parameter h_n. Our empirical experience suggests that the performance of the proposed procedure is not sensitive to the choice of h_n (see Section 3). In practice, we can use K-fold cross validation to choose the bandwidth. We first divide the data set randomly into K parts with roughly equal size. For the kth part, $k = 1, \cdots, K$, we fit model (2.1) using the rest $K-1$ parts of the data, and then evaluate the quantile loss from predicting the τth conditional quantile of T for the uncensored data that are left out. For computational simplicity, we use the unpenalized estimator $\widetilde{\boldsymbol{\beta}}(\tau)$ with $\lambda_n = 0$ when calculating the quantile loss. We choose the h_n that gives the minimum average quantile loss.

The proposed estimation also involves the penalization parameter λ_n, which determines the sparseness of the resulting estimator. In practice, the penalization parameter is often selected by minimizing some model-selection criterion. One commonly used criterion in variable selection literature is the Bayesian Information Criterion (BIC, Schwarz, 1948), which provides a large-sample approximation to twice the logarithm of the Bayes factor. Specifically, the BIC is defined as

$$\text{BIC} = -2\{\log L(\widehat{\boldsymbol{\beta}}_R) - \log L(\widehat{\boldsymbol{\beta}}_F)\} + (p_R - p)\log n,$$

where $L(\widehat{\boldsymbol{\beta}}_R)$ and $L(\widehat{\boldsymbol{\beta}}_F)$ are the maximized likelihoods under a reduced model with p_R parameters and under the full model with p parameters, respectively. It is known that Rao's score test statistics (Rao, 1948) are asymptotically equivalent to the likelihood ratio statistics under both null and Pitman's alternative hypotheses (Serfling, 1980, page 156). Koenker and Machado (1999) discussed tests based on these two types of statistics in linear quantile regression setup. Motivated by this, we propose to choose λ_n that minimizes the following Score-based Bayesian Information Criterion (SBIC):

$$\text{SBIC}(\lambda_n) = nM_n\{\widehat{\boldsymbol{\beta}}_{\lambda_n}(\tau), \widehat{w}_i\}D_n^{-1}M_n\{\widehat{\boldsymbol{\beta}}_{\lambda_n}(\tau), \widehat{w}_i\} + p_{\lambda_n}\log(n), \tag{2.9}$$

where $\widehat{\boldsymbol{\beta}}_{\lambda_n}(\tau)$ is the penalized estimator with the penalization parameter value of λ_n, p_{λ_n} is the number of non-zero elements in $\widehat{\boldsymbol{\beta}}_{\lambda_n}(\tau)$, and $D_n = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T\{\tau\widehat{w}_i^2 + \tau^2(1-2\widehat{w}_i)\}$ is the asymptotic covariance matrix of the subgradient based on the plugged-in weights $\widehat{w}_i$. Similar score-based information criterion was also employed and justified by Leng (2010) for regularized rank regression.

In addition, to obtain the weights for redistribution-of-mass, we need determine the number of indices q. Li et al. (1999) proposed a chi-squared test for determining the number of significant EDR directions obtained by SIR. Xia et al. (2010) developed an alternative selection criterion. This method was shown to be consistent for selecting q but it is based on cross-validation and thus is computationally more intensive than the chi-squared test of Li et al. (1999).

2.5. Asymptotic Properties

To establish the asymptotic results in this paper, we require the following assumptions.

A1 The random design vector $\mathbf{x}$ is bounded in probability, has a bounded density function, and $E(\mathbf{x}\mathbf{x}^T)$ is a positive definite $p \times p$ matrix.

A2 Let $F_0(t|\mathbf{x})$ and $G(t|\mathbf{x})$ denote the survival functions of T_i and C_i conditional on $\mathbf{x}$, respectively. The first derivatives of $F_0(t|\mathbf{x})$ and $G(t|\mathbf{x})$ with respect to t, denoted as $f_0(t|\mathbf{x})$ and $g(t|\mathbf{x})$ are uniformly bounded with respect to t and $\mathbf{x}$. The $F_0(t|\mathbf{x})$ and $G(t|\mathbf{x})$ have bounded (uniformly in t) second-order partial derivatives with respect to $\mathbf{x}$. In addition, $\sup_t |F_0(t|\mathbf{x}') - F_0(t|\mathbf{x})| = O(\|\mathbf{x}' - \mathbf{x}\|)$, where $\|\cdot\|$ denotes the Euclidean norm.

A3 The true coefficient $\boldsymbol{\beta}_0(\tau)$ is in the interior of a bounded convex region $\mathcal{B}$. For $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0(\tau)$, $E\left\{\mathbf{x}\mathbf{x}^T f_0(\mathbf{x}^T\boldsymbol{\beta}_0|\mathbf{x})\right\}$ is positive definite, and $1 - G(\mathbf{x}^T\boldsymbol{\beta}|\mathbf{x}) = P(C > \mathbf{x}^T\boldsymbol{\beta}|\mathbf{x}) > 0$ with probability one.

A4 There exists an EDR direction $\boldsymbol{\gamma}_{0,j} \in R^q$ such that for any $j = 1, \cdots, q$, (i) $\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{0,j} = O_p(n^{-1/2})$; (ii) $n^{-1/2}(\widehat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k) = n^{-1}\sum_{i=1}^n \mathbf{d}_{ki}$, where $\mathbf{d}_{ki}$ are independent p-dimensional vectors with means zero and finite variances.

A5 The univariate nonnegative kernel function $K(\cdot)$ has a compact support. It is a νth order kernel function satisfying $\int K(u)du = 1$, $\int K^2(u)du < \infty$, $\int u^j K(u)du = 0$ for $j < \nu$ and $\int |u|^v K(u)du < \infty$, and it is Lipschitz continuous of order ν, where $\nu \geq 2$ is an integer.

A6 The bandwidth h_n satisfies $h_n = O(n^{-\alpha})$ with (i) $0 < \alpha < \min(1/\nu, 1/q)$; (ii) $1/(2\nu) < \alpha < 1/(3q)$ and $\nu > 3/2q$.

Remark 1 *The boundedness condition of $\mathbf{x}$ in A1 is posed for technical convenience. It is possible to allow the bound on $\mathbf{x}$ to grow slowly in n but this will complicate the technical proof. Assumption A2 is needed to obtain the asymptotic* *properties of the local Kaplan-Meier estimator. Assumption A3 ensures the identifiability of $\boldsymbol{\beta}_0(\tau)$. Assumption A4(i) states the root-n consistency of the estimated EDR direction, which is needed to establish the root-n consistency of $\widetilde{\boldsymbol{\beta}}(\tau)$. This assumption holds for the modified sliced inverse regression estimation in Li et al. (1999) and for the hazard-function-based minimum average variance estimation in Xia et al. (2010). The linear presentation of $\widehat{\boldsymbol{\gamma}}_k$ is assumed in A4(ii) to help establish the normality of $\widetilde{\boldsymbol{\beta}}(\tau)$, and this condition can be obtained for the dimension reduction methods of Li et al. (1999) and Xia et al. (2010) with more technical endeavors under some higher-level conditions. Assumption A5 requires $K(\cdot)$ to be a νth order kernel, where the requirement of ν depends on the dimensionality of indices q. For larger q, higher order kernel function is needed in order to control the bias; see Hu and Fan (1992) for discussions on the construction of higher order kernel functions. Assumption A6 specifies the conditions on the bandwidth h_n, where the weaker condition (i) is needed for establishing the consistency and the stronger condition (ii) is needed for the normality.*

We first establish the consistency and asymptotic normality of the initial unpenalized estimator $\widetilde{\boldsymbol{\beta}}(\tau)$.

Theorem 1 *Suppose models (2.1), (2.6) and assumptions A1-A4(i), A5, A6(i) hold, then $\widetilde{\boldsymbol{\beta}}(\tau) \to \boldsymbol{\beta}_0(\tau)$ in probability as $n \to \infty$. Furthermore, if A4(ii) and A6(ii) both hold, then we have $n^{1/2}\{\widetilde{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0(\tau)\} \xrightarrow{D} N(0, \Gamma_1^{-1} V \Gamma_1)$, where $\Gamma_1 = E\left[\mathbf{x}\mathbf{x}^T\{1 - G(\mathbf{x}^T\boldsymbol{\beta}_0(\tau)|\mathbf{x})\} f_0\{\mathbf{x}^T\boldsymbol{\beta}_0(\tau)|\mathbf{x}\}\right]$, $V = cov(\mathbf{v}_i)$ with $\mathbf{v}_i$ defined in (6.9) of the Appendix.*

We next establish the property of consistency in variable selection of the proposed penalized estimator $\widehat{\boldsymbol{\beta}}(\tau)$. Denote $\mathcal{A}(\tau) = \{j : \boldsymbol{\beta}_j(\tau) \neq 0\}$ and $\mathcal{A}^c(\tau) = \{j : \boldsymbol{\beta}_j(\tau) = 0\}$.

Theorem 2 *Suppose models (2.1), (2.6) and assumptions A1-A6(ii) hold, $n^{-1/2}\lambda_n \to 0$ and $n^{r/2-1}\lambda_n \to \infty$, then $P\left(\{j : \widehat{\boldsymbol{\beta}}_j(\tau) \neq 0\} = \mathcal{A}(\tau)\right) \to 1$ as $n \to \infty$.*

Remark 2 *Theorem 2 suggests that the proposed procedure is able to select the correct model with probability approaching one. To achieve the same efficiency as the oracle estimator obtained under the true model, we can update the estimates for the non-zero coefficients in $\widehat{\beta}(\tau)$ by minimizing the weighted objective function*

(2.8) using the selected covariates only with $\lambda_n = 0$. By Theorem 1, the updated estimator is asymptotically normal and it has the same efficiency as the oracle estimator. For finite samples, our numerical studies show that the re-estimation helps reduce the estimation bias of non-zero coefficients caused by shrinkage and thus leads to more efficient estimation.

3. Simulation Study

We study two different examples to investigate the performance of the proposed penalized estimator via redistribution of mass, referred to as PROM. The dimension of covariates is 20 in Example 1 and 100 in Example 2. We focus on two quantile levels $\tau = 0.25$ and 0.5, and sample sizes $n = 200$ and 500. For each scenario, the simulation is repeated 500 times.

Example 1. The survival times are generated from the following model

$$T_i = 1 + 1.5x_{i1} + 0.7x_{i2} + x_{i3} - 0.5x_{i4} + (1 + \gamma x_{i4})\epsilon_i,$$

where $i = 1, \cdots, n$, $\epsilon_i \sim N(0, 1)$, and γ measures the heteroscedasticity. Besides the 4 relevant covariates, we include another 16 independent noise variables, $x_{i5}, \cdots, x_{i20}$. For $j = 1, \cdots, 20$, $x_{ij} \sim U(-1, 1)$. We set $\gamma = -0.742$, so that the quantile coefficients are $(0.326, 1.5, 0.7, 1.0, 0, \cdots, 0)$ at $\tau = 0.25$ and $(1.0, 1.5, 0.7, 1.0, -0.5, 0, \cdots, 0)$ at $\tau = 0.5$. Under this heteroscedastic model, the covariate x_{i4} has a negative impact on the median but no impact on the first quartile of the conditional distribution of T_i. The observed responses are obtained by $Y_i = \min(T_i, C_i)$, where $C_i \sim U(-2, 18)$ yielding averaged 15% censoring and $C_i \sim U(-2, 8)$ yielding averaged 30% censoring.

We compare the following estimators. The $PROM_S$ and $PROM_M$ are two variations of the proposed PROM estimator, where the indices are estimated by using the sliced inverse regression (SIR) estimation of Li et al. (1999) and the minimum average variance estimation (MAVE) of Xia et al. (2010), respectively. The SIR estimation is obtained by using the R function implemented by Sun available at `http://www.bios.unc.edu/~wsun/`, and the MAVE estimation is obtained by using the matlab program provided by Xia. The oracle estimator is obtained by using the proposed unpenalized method under the true model, that is, with the first three covariates at $\tau = 0.25$ and the first four covariates at $\tau = 0.50$. The oracle estimator serves as a gold standard. Two variations of the oracle estimator, $Oracle_M$ and $Oracle_S$ are included corresponding to MAVE and SIR indices estimation, respectively. The $PIPW$ is the penalized estimator developed by Shows et al. (2010) by using the inverse-probability-weighting scheme of Bang and Tsiatis (2002).

For both examples, the number of true indices is $q = 2$. Therefore, for the $PROM$ and oracle estimators, we use the fourth-order kernel function (Müller, 1984): $K(x) = (105/64)(1 - 5x^2 + 7x^4 - 3x^6)I(|x| \leq 1)$. The bandwidth h_n is selected by 5-fold cross validation as described in Section 2.4. Our numerical studies suggest that the proposed estimator $PROM$ based on the selected $\hat{q}$ performs very similarly as that based on the the true q; see Table 4 for comparison in Example 2. For computational convenience, in Example 1, we use $q = 2$ for both MAVE and SIR estimation.

Table 1 summarizes the variable selection results of the penalized estimators $PROM_M$, $PROM_S$ and $PIPW$. Overall, the proposed $PROM$ methods with indices by both MAVE and SIR outperform the $PIPW$ method in variable selection. The $PROM$ and $PIPW$ methods perform comparably for selecting the relevant variables (TP), but the $PIPW$ selects irrelevant variables more often and thus has much lower oracle proportions in all scenarios considered. As n increases to 500, the $PROM$ methods have oracle proportions close to 1 at both quantiles.

Table 2 summarizes the mean squared errors of estimates for the non-zero quantile coefficients from all methods. Compared to the other coefficients, $\beta_2(0.25) = \beta_2(0.5) = 0.7$ and $\beta_4(0.5) = -0.5$ have smaller magnitudes. When $n = 200$, the proposed methods $PROM_M$ and $PROM_S$ tend to over shrink these small coefficients to trade for simpler models, which results in larger mean squared errors than $PIPW$. However, when n increases to 500, the $PROM$ estimates become more efficient than $PIPW$, and their mean squared errors become comparable to those of the oracle estimates.

To study the sensitivity of the developed $PROM$ method to the bandwidth h_n, we choose $h_n = cn^{-0.15}$ and apply $PROM$ to the same data sets used in Tables 1 and 2 for $c = 0.2, 0.4, 0.6, ..., 2.0$. The results show that the developed $PROM$ method is very robust to the bandwidth h_n in both variable selection and parameter estimation. To save space, we report in Table 3 part of the results for 15% censoring, $\tau = 0.25$, $n = 500$ and three selected values of c.

[Tables 1-3 are about here.]

As for the methods to obtain the EDR subspace and the corresponding indices in $PROM$, Tables 1 and 2 show that the SIR and MAVE estimators perform comparably. Since SIR is computationally more convenient, we use SIR to estimate the dimension q and the dimension reduction directions γ_i in Example 2.

Example 2. In this study, we increase the number of covariates to 100, and generate data from the following model

$$T_i = 1 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + (1 - 0.5x_{i4})\epsilon_i,$$

where $i = 1, \cdots, n$, $\epsilon_i \sim N(0, 1)$, and $x_{ij} \sim U(-1, 1)$, $j = 1, \cdots, 100$. The true quantile coefficients equal $(0.326, 1, 1, 1, 1.337, 0, \cdots, 0)$ at $\tau = 0.25$ and $(1, 1, 1, 1, 1, 0, \cdots, 0)$ at $\tau = 0.50$. The observed responses are obtained by $Y_i = \min(T_i, C_i)$, where $C_i \sim U(-2, 18)$ yielding averaged 15% censoring, and $C_i \sim U(-2, 8)$ yielding averaged 30% censoring. For the developed PROM method, we use SIR to estimate the indices, as well as the dimension q selected by the chi-square test (Li et al., 1999). We report the results of the proposed method with both the true $q = 2$ and the estimated $\hat{q}$ by SIR in Tables 4 and 5.

[Tables 4-5 are about here.]

Table 4 suggests that the performance of $PIPW$ in variable selection deteriorates with higher dimension of covariates, and $PIPW$ selects many irrelevant variables in all scenarios. In contrast, the $PROM$ methods have much higher accuracy in variable selection, and their oracle proportions approach to 1 as n increases. For those nonzero quantile coefficients, the $PROM$ estimates have mean squared errors smaller than those of $PIPW$ in most cases. For larger sample size $n = 500$, the $PROM$ estimates are almost as efficient as the oracle estimates. In addition, both Tables 4 and 5 show that the $PROM$ method based on the estimated $\hat{q}$ behaves very similarly as that based on the true q.

4. Real Data Analysis

We apply the proposed variable selection procedure to a head and neck cancer clinical trial, conducted by Eastern Cooperative Oncology Group and the Southwest Oncology Group (Adelstein et al., 2003). In addition to the evaluation of the overall effectiveness of three treatments, namely, standard radiotherapy (treatment A), radiotherapy plus simultaneous Cisplatin (treatment B), and split-course radiotherapy plus simultaneous cisplatin and 5-fluorouracil (treatment C), it was of substantial interest to detect the treatment effectiveness over high-risk patient populations (often characterized by the lower quantiles of the survival distribution). Such investigations would potentially lead to more effective next generation therapies for targeted subpopulations.

After excluding the ineligible patients and those with missing data, the data set has 171 subjects, of whom 129 died during the follow-up period. We apply the proposed variable selection method $PROM$ to the data set to study the impacts of the predictors on the τth quantile of the overall survival times (in months). We focus on two quantiles $\tau = 0.25$ and 0.5. The standard radiotherapy treatment (treatment A) is treated as the baseline.

Besides the three treatment arms, we also consider the following continuous or ordinal confounders, including age, height, weight, weight loss, tumor differentiation, size of the primary tumor (cm), and categorical variables including gender (1 = female, 0 = male), race (1 = white, 0 = black), smoking (nonsmoker, light cigarette smoker with less than 20 packs a year, moderate cigarette smoker with 20-40 packs a year, heavy cigarette smoker with more than 40 packs a year), alcohol drinking (light drinker: consuming less than 10oz whiskey a week or equivalent, moderate drinker: consuming 10-32 oz whiskey a week or equivalent, heavy drinker: consuming more than 32 whiskey a week or equivalent) and primary tumor site (oralcavity, orapharynx, hypopharynx, larynx with oralcavity). The continuous and ordinal variables are standardized to have mean zero and standard deviation one. For the categorical variables smoking, alcohol drinking and primary tumor site, we treat nonsmoker, light drinker and oralcavity as the baseline. Therefore, the full model contains total $p = 19$ coefficients including the intercept effect, which presents the τth quantile of survival times for a male, black, nonsmoking and light-alcohol-drinking patient who received standard radiotherapy treatment and had average age, height, weight, weight loss, average tumor size and moderately well differentiated oral cavity tumor.

By using the selection criterion in Xia et al. (2010), the dimension of the central subspace (CS) is selected to be 4 for this data set. The 4 indices are then estimated by the MAVE method of Xia et al. (2010) and the SIR method of Li et al. (1999). Results based on the MAVE indices estimation are very similar to those based on SIR and thus are omitted. The sparse index-based estimation of the coefficients is obtained for $\tau = 0.25$ and 0.5 as in Sections 2.3 and 2.4 with the bandwidth h_n selected by 5-fold cross validation and the tuning parameter λ_n selected by minimizing SBIC(λ_n). We use the eighth-order kernel function (Hu and Fan, 1992): $K(x) = 1/13(1 - x^2)(35 - 385x^2 + 1001x^4 - 715x^6)I(|x| \leq 1)$.

The penalized coefficient estimates from the $PIPW$ method (Shows et al., 2010) and the proposed $PROM$ method are summarized in Table 6. Previous analysis (without accounting for any confounders) pointed that, in terms of effect, treatment A differed from C significantly, while only differring from B marginally (Adelstein et al., 2003). Our quantile regression analysis reveals some interesting heterogeneity in the population. The treatment B tends to be more effective for more severe cases (at the lower quartile), while both treatments B and C show positive effects for the typical cases (at the median). In addition, height shows negative effects at the median, while white patients tend to have longer median survival than black patients. These two effects are not selected for more severe cases. On the other hand, larger size of primary tumor is associated with shorter survivals for more severe cases, but not at the median. The two tumor sites hypopharynx and larynx show positive effects at both quantiles, while oropharynx tends to have no effect. As suggested in the simulation study, $PIPW$ had difficulty identifying the correct model for data sets with larger number of predictors. For this data set, $PIPW$ yields more shrinkage, leading to much more sparse models at both quantiles. More specifically, $PIPW$ suggests that both treatments B and C have no difference than the baseline treatment at both quantiles, and only tumor size, weight and gender have effects on the lower quartile of the survival distribution.

[Tables 6-7 are about here.]

To further compare the results from $PIPW$ and $PROM$, we evaluate the risk prediction accuracy of the models selected by the two methods. For each subject i, we define the risk score as the estimated conditional quantile, $m(\mathbf{x}_i) = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}(\tau)$, where $\mathbf{x}_i$ is the covariate vector and $\widehat{\beta}(\tau)$ is the penalized coefficient estimate from either $PIPW$ or $PROM$ at quantile level τ. In our context, good risk scores are expected to better discriminate among subjects with longer and shorter survivals. In medical studies, the concordance measure $C_{t_0} = P\{m(\mathbf{x}_2) > m(\mathbf{x}_1)|T_2 > T_1, T_1 < t_0\}$ is commonly used to evaluate the overall performance of a risk scoring system, where t_0 is a prespecified follow-up time point. To account for the censoring, we employ the C-statistic of Pencina and D'Agostino (2004):

$$\widehat{C}_{t_0} = \frac{\sum_{i \neq j} \delta_i I(Y_i < Y_j, Y_i < t_0) I\{m(\mathbf{x}_i) < m(\mathbf{x}_j)\}}{\sum_{i \neq j} \delta_i I(Y_i < Y_j, Y_i < t_0)}.$$

The C_{t_0}-statistics with different values of t_0 are summarized in Table 7. Results show that the model selected by $PROM$ has higher risk prediction accuracy than that selected by $PIPW$ at both quantile levels for all values of t_0 examined.

5. Discussion

We have developed a new variable selection approach for censored quantile regression. Our models depict more completely the survival distribution of interest, identifying important factors leading to poor prognosis in survival. This is of particular interest to physicians, who are keen on designing effective treatments for targeted patient sub-population, often characterized by short survival. In contrast, such a small group may be overlooked by using the more popularly used Cox and AFT models. As opposed to the existing methods for censored quantile regression, our developed methods require fewer stringent assumptions, and enjoys computational readiness.

We proposed to estimate the censoring probabilities nonparametrically based on effective dimension reduction indices. The proposed method can be extended to situations where the number of predictors grow in the sample size. However, the theoretical justification requires studying properties of the dimension reduction methods for censored regression data with growing dimensions, and this is beyond the scope of the current paper. To avoid the curse of dimensionality, an alternative way is to estimate the censoring probabilities by fitting a semiparametric regression model (e.g. Cox proportional hazards model). McKeague et al. (2001) showed that this method works reasonably well even when the Cox model is slightly misspecified. However, this approach requires the semiparametric and the quantile regression models to be compatible. We leave the formal investigation of this semiparametric approach to a future study.

Acknowledgement

Wang's research is supported by National Science Foundation (DMS-1007420). Zhou's research is supported by National Science Foundation (DMS-0906665). Li's research is supported by National Institutes of Health (R01CA95747). The authors would like to thank Professor Robert Gray and the Eastern Cooperative Oncology Group for providing the head and neck cancer data set, and Professor Yingcun Xia for providing the matlab program of the MAVE method. We would like to thank two anonymous reviewers, the associate editor and the editor for constructive comments and helpful suggestions.

APPENDIX

To simplify the presentation, we will omit τ in various expressions such as $\boldsymbol{\beta}(\tau)$, $e_i(\tau)$ and $\mathcal{A}(\tau)$, and we focus on the cases with $q=1$. Proof for $q>1$ follows the same line by using the asymptotic properties of the local Kaplan-Meier estimate $\widehat{F}(\cdot|\mathbf{z})$ for general cases with $q \geq 1$, but the notations are more complicated. Define $\widehat{z}_i = \mathbf{x}_i^T\widehat{\boldsymbol{\gamma}}$, $z_i = \mathbf{x}_i^T\boldsymbol{\gamma}$ and $z_{0i} = \mathbf{x}_i^T\boldsymbol{\gamma}_0$. To reflect the dependence of the weights w_i for redistribution of masses on F and γ, we write w_i as $w_i(F,\gamma)$. In addition, we define $\mathbf{M}_n(\boldsymbol{\beta},F,\boldsymbol{\gamma}) = n^{-1}\sum_{i=1}^n \mathbf{m}_i(\boldsymbol{\beta},F,\boldsymbol{\gamma})$ as the subgradient of the weighted quantile objective function $n^{-1}L(\beta,w)$ defined in (2.2), where

$$\begin{aligned}
\mathbf{m}_i(\boldsymbol{\beta},F,\boldsymbol{\gamma}) &= \mathbf{x}_i\left\{\tau - w_i(F,\boldsymbol{\gamma})I(Y_i \leq \mathbf{x}_i^T\boldsymbol{\beta})\right\} \\
&= \mathbf{x}_i\Big[\left\{\tau - I(C_i > \mathbf{x}_i^T\boldsymbol{\beta}, T_i \leq \mathbf{x}_i^T\boldsymbol{\beta}) - I(C_i \leq \mathbf{x}_i^T\boldsymbol{\beta}, T_i \leq C_i)\right\} \\
&\qquad - I(C_i \leq \mathbf{x}_i^T\boldsymbol{\beta}, T_i \geq C_i) I\left\{F(C_i|z_i) < \tau\right\}\left\{\frac{\tau - F(C_i|z_i)}{1-F(C_i|z_i)}\right\}\Big].
\end{aligned}$$

Let $\mathbf{M}(\boldsymbol{\beta},F,\boldsymbol{\gamma}) = E_{\mathbf{x}}\left\{\mathbf{m}(\boldsymbol{\beta},F,\boldsymbol{\gamma})\right\} = E\left[\mathbf{x}\left\{\tau - H(\mathbf{x}^T\boldsymbol{\beta}|\mathbf{x}) - R(\boldsymbol{\beta},F,\gamma|\mathbf{x})\right\}\right]$, where

$$H(t|\mathbf{x}) = \{1-G(t|\mathbf{x})\}F_0(t|\mathbf{x}) + \int_{-\infty}^t F_0(u|\mathbf{x})g(u|\mathbf{x})du,$$

$$R(\boldsymbol{\beta},F,\mathbf{x}) = \int_{-\infty}^{\mathbf{x}^T\boldsymbol{\beta}} \{1-F_0(u|\mathbf{x})\}g(u|\mathbf{x})\frac{\tau - F(u|z_i)}{1-F(u|z_i)} I\{F(u|z_i) < \tau\}du.$$

The proof of Theorem 1 depends on the following Lemma.

Lemma 1 *Suppose assumptions A1-A5 hold, then we have for any $q \geq 1$,*

(i) $\|\widehat{F} - F_0\|_{\mathcal{H}} \doteq \sup_t \sup_{\mathbf{x}} |\widehat{F}(t|\widehat{\mathbf{z}}) - F_0(t|\mathbf{x})| = \sup_t \sup_{\mathbf{x}} |\widehat{F}(t|\widehat{\mathbf{z}}) - F_0(t|z_0)| = O_p(\{\log n/(nh_n^p)\}^{1/2} + h_n^{\nu})$;

(ii) $\hat{F}(t|\mathbf{z}) - F_0(t|\mathbf{z}) = \sum_{j=1}^n B_{nj}(\mathbf{z})\xi(Y_j, \delta_j, t, \mathbf{z}) + O_p\left(\{\log n/(nh_n^p)\}^{3/4} + h_n^\nu\right)$ *a.s.*,

where $\xi(Y, \delta, t, \mathbf{z}) = \{1-F_0(t|\mathbf{z})\}\big[-\int_0^{\min(Y,t)} f_0(s|\mathbf{z})\{1-F_0(s|\mathbf{z})\}^{-2}\{1-G(s|\mathbf{z})\}^{-1}ds$ $+I(Y \le t, \delta = 1)\{1 - F_0(Y|\mathbf{z})\}^{-1}\{1 - G(Y|\mathbf{z})\}^{-1}\big]$.

Proof. Note that

$$\widehat{F}(t|\widehat{z}) - F_0(t|z_0) = \left\{\widehat{F}(t|\mathbf{x}^T\widehat{\boldsymbol{\gamma}}) - F_0(t|\mathbf{x}^T\widehat{\boldsymbol{\gamma}})\right\} + \left\{F_0(t|\mathbf{x}^T\widehat{\boldsymbol{\gamma}}) - F_0(t|\mathbf{x}^T\boldsymbol{\gamma}_0)\right\}. \quad (6.1)$$

By extending Theorem 2.1 of Gonzalez-Manteiga and Cadarso-Suarez (1994) to general cases with $q \ge 1$, we have

$$\sup_t \sup_z |\widehat{F}(t|z) - F_0(t|z)| = O_p(\{\log n/(nh_n^p)\}^{1/2} + h_n^\nu). \quad (6.2)$$

Lemma 1(i) thus follows by combining (6.1), (6.2), assumptions A1, A2 and A4. Lemma 1(ii) gives the linear representation of $\widehat{F}(\cdot)$ for general cases with $q \ge 1$. Its proof is similar to that of Theorem 2.3 in Gonzalez-Manteiga and Cadarso-Suarez (1994). The main difference is that the bias influence and the variance influence are h^ν and $(nh_n^q)^{-1}$ for a νth order kernel function in the q-dimensional context, in contrast to h^2 and $(nh_n)^{-1}$ for a second order Kernel function in the one-dimensional context.

Proof of Theorem 1.

The consistency of $\widetilde{\boldsymbol{\beta}}$ can be easily shown by using Lemma 1(i) and the similar arguments as in the proof of Theorem 1 in Wang and Wang (2009). Therefore, we omit the details.

To establish the asymptotic normality, we first prove the $\sqrt{n}$-consistency of $\widetilde{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}_0$. Define

$$\Gamma_1 = \frac{\partial M(\boldsymbol{\beta}, F_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = -E\left[\mathbf{x}\mathbf{x}^T\left\{1 - G(\mathbf{x}^T\boldsymbol{\beta}_0|\mathbf{x})f_0(\mathbf{x}^T\boldsymbol{\beta}_0|\mathbf{x})\right\}\right].$$

Note that Γ_1 is continuous at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, and has a full rank under assumption A3. Therefore, there exists a constant K such that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \le K\|M(\widetilde{\boldsymbol{\beta}}, F_0, \boldsymbol{\gamma}_0)\|$ with probability tending to one. It then suffices to show that $\|M(\widetilde{\boldsymbol{\beta}}, F_0, \boldsymbol{\gamma}_0)\| = O_p(n^{-1/2})$.

Using similar arguments as in the proof of Lemma 2 in Wang and Wang (2009), we have

$$\|M_n(\boldsymbol{\beta}, F, \boldsymbol{\gamma}) - M_n(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0) - M(\boldsymbol{\beta}, F, \boldsymbol{\gamma})\| = o_p(n^{-1/2}), \quad (6.3)$$

uniformly over $(\boldsymbol{\beta}, F, \boldsymbol{\gamma})$ such that $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le a_n$, $\|F-F_0\|_{\mathcal{H}} \le a_n$ and $\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| \le a_n$, where $a_n \to 0$ as $n \to \infty$. Therefore, by the consistency of $\widetilde{\beta}$, Lemma 1 and assumption A4,

$$\|M(\widetilde{\boldsymbol{\beta}}, \widehat{F}, \widehat{\boldsymbol{\gamma}}) - M_n(\widetilde{\boldsymbol{\beta}}, \widehat{F}, \widehat{\boldsymbol{\gamma}}) + M_n(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0)\| = o_p(n^{-1/2}). \quad (6.4)$$

Under assumptions A1, A2 and A4, $\|M_n(\widetilde{\boldsymbol{\beta}}, \widehat{F}, \widehat{\boldsymbol{\gamma}}) - M_n(\widetilde{\boldsymbol{\beta}}, \widehat{F}, \boldsymbol{\gamma}_0)\| = o_p(n^{-1/2})$. In addition, combining the subgradient condition (Koenker, 2005) and assumption A1 gives $\|M_n(\widetilde{\boldsymbol{\beta}}, \widehat{F}, \widehat{\boldsymbol{\gamma}})\| = O_p(n^{-1})$. Therefore,

$$\|M_n(\widetilde{\boldsymbol{\beta}}, \widehat{F}, \boldsymbol{\gamma}_0)\| \le \|M_n(\widetilde{\boldsymbol{\beta}}, \widehat{F}, \widehat{\boldsymbol{\gamma}}) - M_n(\widetilde{\boldsymbol{\beta}}, \widehat{F}, \boldsymbol{\gamma}_0)\| + \|M_n(\widetilde{\boldsymbol{\beta}}, \widehat{F}, \widehat{\boldsymbol{\gamma}})\| = O_p(n^{-1/2}). \quad (6.5)$$

Let $\epsilon > 0$ and $F_\epsilon(t|z) = F_0(t|z) + \epsilon\{F(t|z) - F_0(t|z)\}$. Following some routine algebra, we can derive the functional derivative of $M(\boldsymbol{\beta}_0, F, \boldsymbol{\gamma}_0)$ at F_0 in the direction $[F - F_0]$ as

$$\begin{aligned} & \Gamma_2(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0)[F - F_0] \\ = \; & \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[M\{\boldsymbol{\beta}_0, F_0 + \epsilon(F - F_0)\}, \boldsymbol{\gamma}_0) - M(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0)\right] \\ = \; & (1-\tau)E\left[\mathbf{x}\int_{-\infty}^{\mathbf{x}^T\boldsymbol{\beta}_0} \frac{F(t|z_0) - F_0(t|z_0)}{1 - F_0(t|z_0)} g(t|\mathbf{x})dt\right]. \end{aligned}$$

Therefore,

$$\Gamma_2(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0)[\widehat{F} - F_0] = (1-\tau)E\left[\mathbf{x}\int_{-\infty}^{\mathbf{x}^T\boldsymbol{\beta}_0} \frac{\widehat{F}(t|z_0) - F_0(t|z_0)}{1 - F_0(t|z_0)} g(t|\mathbf{x})dt\right].$$

By plugging in the linear representation of $\widehat{F}(t|z) - F_0(t|z)$ in Lemma 1(ii) and by applying the Taylor expansion, we obtain

$$\Gamma_2(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0)[\widehat{F} - F_0] = n^{-1}\sum_{i=1}^n (1-\tau)\phi_i + o_p(n^{-1/2}), \quad (6.6)$$

where $\phi_i = \mathbf{x}_i \int_{-\infty}^{\mathbf{x}_i^T\boldsymbol{\beta}_0} g(t|z_{0i})\xi(Y_i, \delta_i, t, z_{0i})\{1 - F_0(t|\mathbf{x}_i)\}^{-1}dt$, and $\xi(Y, \delta, t, z)$ is defined in Lemma 1(ii). Here ϕ_i are independent random variables with mean zero, and $\Gamma_2(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0)[\widehat{F} - F_0] = O_p(n^{-1/2})$.

Let $\delta_n = o(1)$ be a positive sequence such that $P(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0| \ge \delta_n, \|\widehat{F} - F_0\|_{\mathcal{H}} \ge \delta_n) \to 0$. Following the routine Taylor expansion, we can show that under assumptions A1-A2, $\|\Gamma_2(\widetilde{\boldsymbol{\beta}}, F_0, \boldsymbol{\gamma}_0)[\widehat{F} - F_0] - \Gamma_2(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0)[\widehat{F} - F_0]\| = \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| o_p(1)$, and

$$\|M(\widetilde{\boldsymbol{\beta}},\widehat{F},\boldsymbol{\gamma}_0)-M(\widetilde{\boldsymbol{\beta}},F_0,\boldsymbol{\gamma}_0)-\Gamma_2(\widetilde{\boldsymbol{\beta}},F_0,\boldsymbol{\gamma}_0)[\widehat{F}-F_0]\|\le K\|\widehat{F}-F_0\|_{\mathcal{H}}^2 \quad (6.7)$$

for a constant $K \geq 0$. By Lemma 1(i), for any $\alpha > 0$ such that $1/(4\nu) < \alpha < 1/(2q)$, $\|\widehat{F}-F_0\|_{\mathcal{H}}^2 = o_p(n^{-1/2})$. This together with (6.6) gives

$$\begin{aligned}
& \|M(\widetilde{\boldsymbol{\beta}},F_0,\boldsymbol{\gamma}_0)\| \\
\le\ & \|M(\widetilde{\boldsymbol{\beta}},F_0,\boldsymbol{\gamma}_0)-M(\widetilde{\boldsymbol{\beta}},\widehat{F},\boldsymbol{\gamma}_0)\|+\|M(\widetilde{\boldsymbol{\beta}},\widehat{F},\boldsymbol{\gamma}_0) \\
\le\ & \|M(\widetilde{\boldsymbol{\beta}},\widehat{F},\boldsymbol{\gamma}_0)-M(\widetilde{\boldsymbol{\beta}},F_0,\boldsymbol{\gamma}_0)-\Gamma_2(\widetilde{\boldsymbol{\beta}},F_0,\boldsymbol{\gamma}_0)[\widehat{F}-F_0]\|+\|M(\widetilde{\boldsymbol{\beta}},\widehat{F},\boldsymbol{\gamma}_0) \\
& +\|\Gamma_2(\widetilde{\boldsymbol{\beta}},F_0,\boldsymbol{\gamma}_0)[\widehat{F}-F_0]-\Gamma_2(\boldsymbol{\beta}_0,F_0,\boldsymbol{\gamma}_0)[\widehat{F}-F_0]\|+\|\Gamma_2(\boldsymbol{\beta}_0,F_0,\boldsymbol{\gamma}_0)[\widehat{F}-F_0]\| \\
\le\ & \|\widehat{F}-F_0\|_{\mathcal{H}}^2+\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|o_p(1)+O_p(n^{-1/2}) \\
\le\ & \|M(\widetilde{\boldsymbol{\beta}},F_0,\boldsymbol{\gamma}_0)\|o_p(1)+O_p(n^{-1/2}). \qquad (6.8)
\end{aligned}$$

Therefore, combining (6.4)-(6.8) gives $\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 \le K\|M(\widetilde{\boldsymbol{\beta}},F_0,\boldsymbol{\gamma}_0)\| = O_p(n^{-1/2})$.

Recall from (6.5) that $M_n(\widetilde{\boldsymbol{\beta}},\widehat{F},\gamma_0) = O_p(n^{-1/2})$, The rest of the proof of normality is similar to those of Theorem 3.3 in Pakes and Pollard (1989) and Theorem 2 of Chen et al. (2003). Therefore, we just sketch the main steps here. Let $\Gamma_3 = \partial M(\boldsymbol{\beta}_0,F_0,\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}$. By Assumption 4(i) and Taylor expansion, we get $M(\widetilde{\boldsymbol{\beta}},\widehat{F},\widehat{\gamma})-M(\widetilde{\boldsymbol{\beta}},\widehat{F},\boldsymbol{\beta}_0) = \Gamma_3(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)+o_p(n^{-1/2})$. Define $\mathcal{L}_n(\boldsymbol{\beta}) = M_n(\boldsymbol{\beta}_0,F_0,\gamma_0)+\Gamma_1(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\Gamma_2(\boldsymbol{\beta}_0,F_0,\gamma_0)[\widehat{F}-F_0]+\Gamma_3(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)$. By the root-n consistency result shown above, Lemma 1, (6.3) and (6.7), we have

$$\begin{aligned}
& \|M_n(\widetilde{\boldsymbol{\beta}},\widehat{F},\widehat{\gamma})-\mathcal{L}_n(\widetilde{\boldsymbol{\beta}})\| \\
=\ & \|M_n(\boldsymbol{\beta}_0,F_0,\gamma_0)+M(\widetilde{\boldsymbol{\beta}},\widehat{F},\widehat{\gamma})-\mathcal{L}_n(\widetilde{\boldsymbol{\beta}}) \\
& +M_n(\widetilde{\boldsymbol{\beta}},\widehat{F},\widehat{\gamma})-M(\widetilde{\boldsymbol{\beta}},\widehat{F},\widehat{\gamma})-M_n(\boldsymbol{\beta}_0,F_0,\gamma_0)\| \\
\le\ & \|M(\widetilde{\boldsymbol{\beta}},F_0,\gamma_0)-\Gamma_1(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\| \\
& +\|M(\widetilde{\boldsymbol{\beta}},\widehat{F},\widehat{\gamma})-M(\widetilde{\boldsymbol{\beta}},F_0,\gamma_0)-\Gamma_2(\boldsymbol{\beta}_0,F_0,\gamma_0)[\widehat{F}-F_0]\| \\
& +\|M(\widetilde{\boldsymbol{\beta}},\widehat{F},\widehat{\boldsymbol{\gamma}})-M(\widetilde{\boldsymbol{\beta}},\widehat{F},\boldsymbol{\gamma}_0)-\Gamma_3(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\| \\
& +\|M_n(\widetilde{\boldsymbol{\beta}},\widehat{F},\widehat{\gamma})-M(\widetilde{\boldsymbol{\beta}},\widehat{F},\widehat{\gamma})-M_n(\boldsymbol{\beta}_0,F_0,\gamma_0)\| \\
=\ & o_p(n^{-1/2}).
\end{aligned}$$

Similarly, $\|M_n(\widetilde{\boldsymbol{\beta}},\widehat{F},\widehat{\gamma})-\mathcal{L}_n(\bar{\boldsymbol{\beta}})\| = o_p(n^{-1/2})$, where $\bar{\boldsymbol{\beta}}$ is the minimizer of $\mathcal{L}_n(\boldsymbol{\beta})$ satisfying $n^{1/2}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) = -\Gamma_1^{-1}\{M_0(\boldsymbol{\beta}_0,F_0,\gamma_0)+\Gamma_2(\boldsymbol{\beta}_0,F_0,\gamma_0)[\widehat{F}-F_0]+\Gamma_3(\widehat{\boldsymbol{\gamma}}-$

$\boldsymbol{\gamma}_0)\}$. Since Γ_1 is of full rank under assumption A3, with a bit more work, we get $n^{1/2}(\widetilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) = o_p(1)$. Note that $M_n(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0)$, $\Gamma_2(\boldsymbol{\beta}, F_0, \boldsymbol{\gamma}_0)[\widehat{F} - F_0]$ and $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$ are the averages of independent random vectors of means zero. Therefore, applying the central limit theorem gives

$$n^{1/2}\left\{M_n(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0) + \Gamma_2(\boldsymbol{\beta}_0, F_0, \boldsymbol{\gamma}_0)[\widehat{F} - F_0] + \Gamma_3(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\right\} \xrightarrow{D} N(0, V), \quad (6.9)$$

where $V = Cov(\mathbf{v}_i)$ with $\mathbf{v}_i = \mathbf{m}_i(\boldsymbol{\beta}_0, F_0, \gamma_0) + \boldsymbol{\phi}_i + \mathbf{d}_i$. The asymptotic normality of $\widetilde{\boldsymbol{\beta}}$ is thus proven.

Proof of Theorem 2.

Let $\widehat{\mathcal{A}_n} = \{j : \widehat{\boldsymbol{\beta}}_j \neq 0\}$. We first show that for any $j \notin \mathcal{A}$, $P(j \in \widehat{\mathcal{A}_n}) \to 0$ as $n \to \infty$. Suppose there exists a $k \in \mathcal{A}^c$ such that $|\widehat{\beta}_k| \neq 0$. Let $\boldsymbol{\beta}^*$ be a vector constructed by replacing $\widehat{\beta}_k$ with 0 in $\widehat{\boldsymbol{\beta}}$. For notational simplicity, we write $\widehat{w}_i = w_i(\widehat{F}, \widehat{\boldsymbol{\gamma}})$. Note that $|\rho_\tau(a) - \rho_\tau(b)| \leq |a - b| \max\{\tau, 1 - \tau\} < |a - b|$. Therefore, for large enough n,

$$\begin{aligned}
& L_{AL}(\widehat{\boldsymbol{\beta}}, \widehat{w}) - L_{AL}(\boldsymbol{\beta}^*, \widehat{w}) \\
= \; & \sum_{i=1}^n \widehat{w}_i \left\{\rho_\tau(y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}) - \rho_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^*)\right\} \\
& + \sum_{i=1}^n \widehat{w}_i \left\{\rho_\tau(y^{+\infty} - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}) - \rho_\tau(y^{+\infty} - \mathbf{x}_i^T\boldsymbol{\beta}^*)\right\} + \lambda_n v_k|\widehat{\beta}_k| \\
\geq \; & -\sum_{i=1}^n |\rho_\tau(y^{+\infty} - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}) - \rho_\tau(y^{+\infty} - \mathbf{x}_i^T\boldsymbol{\beta}^*)| - \sum_{i=1}^n \tau|\mathbf{x}_i^T\widehat{\boldsymbol{\beta}} - \mathbf{x}_i^T\boldsymbol{\beta}^*| + \lambda_n v_k|\widehat{\beta}_k| \\
\geq \; & -2\sum_{i=1}^n \|\mathbf{x}_i\| \cdot |\widehat{\beta}_k| + \lambda_n|\widetilde{\beta}_k|^{-r}|\widehat{\beta}_k| > 0, \qquad (6.10)
\end{aligned}$$

where the last inequality holds as $\sum_{i=1}^n \|\mathbf{x}_i\| = O_p(n)$ by Assumption A1, and $n^{-1}\lambda_n|\widetilde{\beta}_k|^{-r} \geq n^{r/2-1}\lambda_n \to \infty$. The (6.10) thus contradicts the fact that $L_{AL}(\widehat{\boldsymbol{\beta}}, \widehat{w}) \leq L_{AL}(\boldsymbol{\beta}^*, \widehat{w})$.

We next show that for any $j \in \mathcal{A}$, $P(j \notin \widehat{\mathcal{A}_n}) \to 0$. We write $b_{\mathcal{A}} = (b_j, j \in \mathcal{A})$ for any vector $b \in R^p$, and $B_{\mathcal{A}\mathcal{A}}$ as the sub-matrix of a $p \times p$ matrix B with both row and column indices in $\mathcal{A}$. Recall from the proof of Theorem 1 that

$$\begin{aligned}
M_n(\boldsymbol{\beta}_{\mathcal{A}}, F, \boldsymbol{\gamma}) \;=\; & M_n(\boldsymbol{\beta}_{0\mathcal{A}}, F_0, \boldsymbol{\gamma}_0) + \Gamma_{1\mathcal{A}\mathcal{A}}(\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) \\
& + \Gamma_{2\mathcal{A}\mathcal{A}}(\boldsymbol{\beta}_{\mathcal{A}0}, F_0, \boldsymbol{\gamma}_0)[F - F_0] + o_p(n^{-1/2}) \qquad (6.11)
\end{aligned}$$

uniformly over $\boldsymbol{\beta}_{\mathcal{A}}$, F and $\boldsymbol{\gamma}$ such that $\|\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}\| = O(n^{-1/2})$, $\|F - F_0\|_{\mathcal{H}} = o(n^{-1/4})$ and $\boldsymbol{\gamma} - \boldsymbol{\gamma}_0 = O(n^{-1/2})$. Denote $\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}} = n^{-1/2}\mathbf{u}$ and let K be some positive constant. By (6.11), for $\|\mathbf{u}\| = K$, we have

$$n\mathbf{u}^T M_n(\boldsymbol{\beta}_{\mathcal{A}}, \widehat{F}, \widehat{\boldsymbol{\gamma}}) = n\mathbf{u}^T\{M_n(\boldsymbol{\beta}_{0\mathcal{A}}, F_0, \boldsymbol{\gamma}_0) + \Gamma_{2\mathcal{A}\mathcal{A}}\} + n^{1/2}\mathbf{u}^T\Gamma_{1\mathcal{A}\mathcal{A}}\mathbf{u} + o_p(n^{1/2}) \quad (6.12)$$

where $\Gamma_{2\mathcal{A}\mathcal{A}} = \Gamma_{2\mathcal{A}\mathcal{A}}(\boldsymbol{\beta}_{\mathcal{A}0}, F_0, \boldsymbol{\gamma}_0)[\widehat{F} - F_0]$. Therefore, with probability tending to one,

$$\begin{aligned} & -n\mathbf{u}^T M_n(\boldsymbol{\beta}_{\mathcal{A}}, \widehat{F}, \widehat{\boldsymbol{\gamma}}) \\ \geq \quad & -n\mathbf{u}^T\{M_n(\boldsymbol{\beta}_{0\mathcal{A}}, F_0, \boldsymbol{\gamma}) + \Gamma_2\} - n^{1/2}\mathbf{u}^T\Gamma_1\mathbf{u} + o(n^{1/2}) \geq k_0 n^{1/2} \quad (6.13) \end{aligned}$$

for some positive k_0. However, the subgradient condition (see the proof of Theorem 1 in Wang and Wang, 2009) requires that

$$\|n\mathbf{u}^T M_n(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}, \widehat{F}, \widehat{\boldsymbol{\gamma}})\| + \lambda_n \sum_{j\in\mathcal{A}} v_j|\tau - I(\widehat{\boldsymbol{\beta}}_j < 0)| \leq O_p(\max_i \|\mathbf{x}_i\|). \quad (6.14)$$

When $\lambda_n = o(n^{1/2})$ and assumption A1 holds, (6.13) and (6.14) suggest that the subgradient condition cannot hold if $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}\| = Kn^{-1/2}$. Using the monotonicity argument in Jurečková (1977), we can show that the gradient condition also can not hold if $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}\| > Kn^{-1/2}$. Therefore, $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}\| \leq Kn^{-1/2}$ with probability tending to one. The proof of Theorem 2 is thus complete.

References

Adelstein, D. J., Li, Y., Adams, G. L., Wagner, H. J., Kish, J. A., Ensley, J. F., Schuller, D. E. & Forastiere, A. A. (2003). An intergroup phase iii comparison of standard radiation therapy and two schedules of concurrent chemoradiotherapy in patients with unresectable squamous cell head and neck cancer. *Journal of Clinical Oncology* **21**, 92–98.

Bang, H. & Tsiatis, A. (2002). Median regression with censored cost data. *Biometrics* **58**, 643–649.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Tech. rep., University of California, Berkeley.

Efron, B. (1967). The two-sample problem with censored data. *Porcessdings Fifth Berkeley Symposium in Mathematical Statistics, Vol. IV, eds. L. 88 Le Cam and J. Neyman, New York: Prentice-Hall* , 831–853.

Engler, D. & Li, Y. (2009). Survival analysis with high-dimensional covariates: an application in microarray studies. *Statistical Applications in Genetics and Molecular Biology* **8**, Issue 1, Article 14.

Hu, T. C. & Fan, J. (1992). Bias correction and higher order kernel functions. *Statistics and Probability Letters* **13**, 235-243.

Huang, J., Ma, S. & Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62**, 813–820.

Huang, Y. (2009). Quantile calculus and censored regression. *Annals of Statistics* **38**, 1607–1637.

Koenker, R. & Geling, O. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association* **96**, 458–468.

Koenker, R. & Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association* **94**, 1296–1310.

Leng, C. (2010). Variable selection and coefficient estimation via regularized rank regression. *Statistica Sinica* **20**, 167–181.

Leng, C. & Tong, X. (2011). A quantile regression estimator for censored data. Manuscript.

Li, K. C., Wang, J. L. & Chen, C. H. (1999). Dimension reduction for censored regression data. *Annals of Statistics* **27**, 1–23.

Lindgren, A. (1995). Quantile regression with censored data using generalized l1 minimization. *Computational Statistics and Data Analysis* **23**, 509–524.

McKeague, I. W., Subramanian, S. & Sun, Y. (2001). Median regression and the missing information principle. *Journal of Nonparametric Statistics* **13**, 709–727.

PENCINA, M. & D'AGOSTINO, R. (2004). Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine* **23**, 2109–2123.

PENG, L. & HUANG, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association* **103**, 637–649.

PORTNOY, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association* **98**, 1001–1012.

PORTNOY, S. & LIN, G. (2010). Asymptotics for censored regression quantiles. *Journal of Nonparametric Statistics* **22**, 115–130.

RAO, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* **44**, 50–57.

SCHWARZ, G. E. (1948). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

SERFLING, R. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons.

SHOWS, J. H., LU, W. & ZHANG, H. H. (2010). Sparse estimation and inference for censored median regression. *Journal of Statistical Planning and Inference* **140**, 1903–1917.

SUBRAMANIAN, S. (2002). Median regression using nonparametric kernel estimation. *Nonparametric Statistics* **14**, 583–605.

WANG, H. & WANG, L. (2009). Locally weighted censored quantile regression. *Journal of American Statistical Association* **104**, 1117–1128.

WANG, S., NAN, B., Z. J. & BEER, D. G. (2008). Doubly penalized buckley-james method for survival data with high-dimensional covariates. *Biometrics* **64**, 132–140.

WU, Y. & LIU, Y. (2009). Variable selection in quantile regression. *Statistica Sinica* **19**, 801–817.

Xia, Y., Zhang, D. & Xu, J. (2010). Dimension reduction and semiparametric estimation of survival models. *Journal of the American Statistical Association* **105**, 278–290.

Zhang, H. H. & Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94**, 691–703.

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

Department of Statistics, North Carolina State University

E-mail: hwang3@ncsu.edu

Department of Statistics, University of Virginia

E-mail: jz9p@virginia.edu

Department of Biostatistics, Harvard University

E-mail: yili@jimmy.harvard.edu

Table 1: Variable selection results for Example 1. TP denotes the average number of relevant variables that are correctly selected. FP denotes the average number of irrelevant variables that are incorrectly selected. OP denotes the oracle proportion, that is, the percentage of times that the true model is correctly selected.

	15% censoring						30% censoring					
	$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.25$			$\tau = 0.50$		
Method	TP	FP	OP	TP	FP	OP	TP	FP	OP	TP	FP	OP
						$n = 200$						
$PIPW$	3.96	1.72	0.31	4.92	1.33	0.36	3.94	2.55	0.19	4.89	2.30	0.19
$PROM_M$	3.95	0.19	0.80	4.64	0.12	0.58	3.92	0.16	0.78	4.45	0.12	0.41
$PROM_S$	3.95	0.15	0.82	4.66	0.15	0.58	3.91	0.15	0.78	4.47	0.10	0.41
						$n = 500$						
$PIPW$	4.00	0.71	0.58	5.00	0.52	0.68	4.00	1.27	0.40	5.00	1.07	0.46
$PROM_M$	4.00	0.08	0.92	5.00	0.08	0.91	4.00	0.09	0.92	4.91	0.05	0.87
$PROM_S$	4.00	0.09	0.92	5.00	0.09	0.90	4.00	0.09	0.92	4.90	0.07	0.83

Table 2: Mean squared errors (×100) of the estimates for nonzero quantile coefficients in Example 1.

	$\tau = 0.25$				$\tau = 0.50$				
	$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_2(\tau)$	$\beta_3(\tau)$	$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_2(\tau)$	$\beta_3(\tau)$	$\beta_4(\tau)$
$n = 200$, 15% censoring									
$Oracle_M$	0.67	1.95	2.05	1.79	1.14	1.75	1.69	1.49	2.20
$Oracle_S$	0.68	1.92	2.02	1.79	1.12	1.76	1.69	1.54	2.24
$PIPW$	1.46	2.38	3.33	2.56	1.50	2.01	2.61	2.19	4.79
$PROM_M$	0.77	2.01	4.27	1.85	1.81	1.95	3.25	1.72	9.12
$PROM_S$	0.78	2.03	4.20	1.89	1.76	1.86	2.80	1.65	8.80
$n = 200$, 30% censoring									
$Oracle_M$	0.78	2.29	2.36	2.03	1.54	2.06	1.94	1.67	2.83
$Oracle_S$	0.81	2.24	2.35	2.06	1.66	1.97	2.00	1.71	3.10
$PIPW$	1.75	3.08	3.54	3.03	1.80	2.89	3.51	2.89	6.00
$PROM_M$	1.02	2.42	5.95	2.24	2.70	2.32	5.22	2.00	12.86
$PROM_S$	1.08	2.42	6.32	2.12	2.80	2.14	4.91	2.26	13.36
$n = 500$, 15% censoring									
$Oracle_M$	0.23	0.75	0.71	0.79	0.44	0.63	0.64	0.71	1.08
$Oracle_S$	0.23	0.73	0.72	0.78	0.48	0.62	0.64	0.72	1.12
$PIPW$	0.48	0.84	1.01	0.96	0.51	0.68	0.87	0.83	1.60
$PROM_M$	0.26	0.75	0.70	0.81	0.49	0.65	0.65	0.71	1.55
$PROM_S$	0.26	0.73	0.71	0.79	0.49	0.64	0.64	0.73	1.63
$n = 500$, 30% censoring									
$Oracle_M$	0.27	0.80	0.77	0.88	0.84	0.76	0.67	0.81	1.59
$Oracle_S$	0.27	0.79	0.76	0.87	0.88	0.75	0.67	0.79	1.66
$PIPW$	0.56	0.98	1.10	1.10	0.65	0.91	1.10	1.00	2.03
$PROM_M$	0.30	0.82	0.81	0.89	0.92	0.77	0.69	0.83	3.23
$PROM_S$	0.31	0.78	0.77	0.90	1.03	0.75	0.68	0.84	3.59

Table 3: Results of PROM methods at different bandwidth values $h_n = cn^{-0.15}$ in Example 1 with 15% censoring, $\tau = 0.25$ and $n = 500$.

		Variable Selection			100×MSE			
Method	c	TP	FP	OP	β_0	β_1	β_2	β_3
$PROM_M$	0.2	4.00	0.10	0.91	0.29	0.72	0.73	0.80
	1.0	4.00	0.08	0.93	0.26	0.75	0.71	0.81
	2.0	4.00	0.08	0.93	0.25	0.73	0.70	0.80
$PROM_S$	0.2	4.00	0.09	0.92	0.29	0.73	0.72	0.79
	1.0	4.00	0.07	0.94	0.27	0.73	0.71	0.79
	2.0	4.00	0.08	0.93	0.26	0.73	0.71	0.79

Table 4: Variable selection results for Example 2. TP denotes the average number of relevant variables that are correctly selected. FP denotes the average number of irrelevant variables that are incorrectly selected. OP denotes the oracle proportion, that is, the percentage of times that the true model is correctly selected. The methods $PROM_S(q)$ and $PROM_S(\hat{q})$ are based on the true and the estimated number of indices, respectively.

	15% censoring						30% censoring					
	$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.25$			$\tau = 0.50$		
Method	TP	FP	OP	TP	FP	OP	TP	FP	OP	TP	FP	OP
						$n = 200$						
$PIPW$	4.99	39.64	0.00	5.00	25.16	0.01	4.98	68.70	0.00	4.99	62.65	0.00
$PROM_S(q)$	4.91	0.34	0.68	4.97	0.34	0.70	4.78	0.48	0.52	4.85	0.38	0.60
$PROM_S(\hat{q})$	4.91	0.40	0.64	4.97	0.27	0.76	4.77	0.53	0.50	4.85	0.43	0.58
						$n = 500$						
$PIPW$	5.00	5.04	0.18	5.00	3.19	0.24	5.00	11.37	0.08	5.00	9.48	0.06
$PROM_S(q)$	5.00	0.13	0.90	5.00	0.14	0.89	5.00	0.17	0.88	5.00	0.10	0.93
$PROM_S(\hat{q})$	5.00	0.16	0.87	5.00	0.13	0.90	5.00	0.19	0.86	5.00	0.09	0.92

Table 5: Mean squared errors ($\times 100$) of the estimates for nonzero quantile coefficients in Example 2. The methods $PROM_S(q)$ and $PROM_S(\hat{q})$ are based on the true and the estimated number of indices, respectively.

	$\tau = 0.25$					$\tau = 0.50$				
	$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_2(\tau)$	$\beta_3(\tau)$	$\beta_4(\tau)$	$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_2(\tau)$	$\beta_3(\tau)$	$\beta_4(\tau)$
					$n = 200$, 15% censoring					
$Oracle_S$	1.14	3.02	2.67	2.42	3.42	1.15	2.44	2.35	2.25	2.60
$PIPW$	2.86	4.25	4.48	4.44	5.69	1.33	3.43	3.94	3.18	3.33
$PROM_S(q)$	1.20	5.10	6.50	5.03	3.72	1.18	3.17	3.40	2.91	3.43
$PROM_S(\hat{q})$	1.27	5.24	6.03	5.36	4.11	1.17	3.20	3.45	3.10	3.50
					$n = 200$, 30% censoring					
$Oracle_S$	1.24	3.44	3.21	2.97	3.59	1.53	3.04	2.71	2.67	3.13
$PIPW$	6.42	11.13	12.24	9.89	11.58	3.03	8.93	8.77	7.73	7.97
$PROM_S(q)$	1.65	8.74	10.59	10.46	7.17	1.61	6.36	7.42	7.05	5.81
$PROM_S(\hat{q})$	1.61	7.18	12.14	9.60	8.67	1.59	5.90	9.32	6.83	4.91
					$n = 500$, 15% censoring					
$Oracle_S$	0.45	1.06	1.07	1.11	1.22	0.38	1.02	0.97	0.88	1.04
$PIPW$	0.58	1.18	1.35	1.36	1.36	0.37	1.30	1.14	1.07	1.28
$PROM_S(q)$	0.45	1.03	1.09	1.11	1.23	0.40	1.01	0.93	0.89	1.02
$PROM_S(\hat{q})$	0.46	1.02	1.10	1.11	1.25	0.37	1.04	0.99	0.89	1.04
					$n = 500$, 30% censoring					
$Oracle_S$	0.51	1.21	1.31	1.29	1.38	0.67	1.15	1.04	1.06	1.34
$PIPW$	0.67	1.47	1.56	1.38	1.69	0.49	1.59	1.46	1.29	1.55
$PROM_S(q)$	0.52	1.22	1.28	1.29	1.36	0.70	1.16	1.07	1.06	1.36
$PROM_S(\hat{q})$	0.50	1.25	1.28	1.30	1.36	0.51	1.17	1.09	1.08	1.33

Table 6: Spare coefficient estimates for the head and neck cancer data set.

	$\tau = 0.25$		$\tau = 0.5$	
Variable	$PIPW$	$PROM$	$PIPW$	$PROM$
(Intercept)	7.464	10.056	23.0092	35.296
Treatment B	0.000	2.371	0.000	4.806
Treatment C	0.000	0.000	0.000	5.556
Age	0.000	0.000	0.000	0.000
Tumor differentiation	0.000	0.000	0.000	0.000
Weight loss	0.000	0.000	0.206	0.000
White	0.000	0.000	0.000	2.789
Height	0.000	0.000	-4.222	-2.979
Weight	1.370	0.000	8.898	3.882
Gender	-0.088	0.000	-8.108	-5.972
Tumor size	-1.286	-0.641	0.000	0.000
Hypopharynx	0.000	1.281	0.000	6.756
Larynx	0.000	3.588	5.032	2.261
Oropharynx	0.000	0.000	0.000	0.000
Light smoker	0.000	0.000	0.000	-14.506
Moderate smoker	0.000	-0.542	0.000	-15.644
Heavy smoker	0.000	-0.268	-4.022	-22.128
Moderate drinker	0.000	-3.139	-4.701	-6.056
Heavy drinker	0.000	-3.531	-5.391	-5.561

Table 7: C_{t_0} statistics of the models selected by $PIPW$ and $PROM$ for the head and neck cancer data set.

	$\tau = 0.25$		$\tau = 0.5$	
t_0	$PIPW$	$PROM$	$PIPW$	$PROM$
20	0.580	0.613	0.612	0.647
40	0.566	0.602	0.600	0.638
60	0.565	0.599	0.599	0.637
80	0.565	0.598	0.598	0.637