Statistica Sinica Preprint No: SS-2014-0071R2

Title	Censored Quantile Regression in Reproducing Kernel Hilbert Space
Manuscript ID	SS-2014-0071.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202014.0071
Complete List of Authors	Seung Jun Shin, Hao Helen Zhang and Yichao Wu
Corresponding Author	Seung Jun Shin
E-mail	sjshin@korea.ac.kr

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

A Nonparametric Survival Function Estimator via Censored Kernel Quantile Regressions

Seung Jun Shin, Hao Helen Zhang, and Yichao Wu

Korea University, University of Arizona and North Carolina State University

Abstract: In survival data analysis, a central interest is to identify the relationship between a possibly censored survival time and explanatory covariates. In this article, a new censored quantile regression method is proposed and studied in the framework of reproducing kernel Hilbert spaces (RKHS). We first establish the joint piecewise linearity of the regression parameters as a function of regularization parameter λ and quantile level τ. An efficient algorithm is then developed to compute the entire two-dimensional solution surface over the $(\lambda \times \tau)$-plane. Finally, a piecewise linear conditional survival function estimator is constructed based on the solution surface. The method provides a new and flexible survival function estimator without requiring such rigid model assumptions as linearity of the survival time or proportionality of the hazards. One important advantage of the estimator is that it can handle moderately high-dimensional covariates. We carry out an asymptotic analysis to justify the proposed method theoretically, and numerical results are shown to illustrate its competitive finite-sample performance under various simulated scenarios and real applications.

Key words and phrases: Censored kernel quantile regression; Conditional survival function; Solution surface.

1 Introduction

In censored survival data analysis, the survival function can be regarded as a counterpart of the distribution function. But its estimation is difficult due to the presence of censoring. The

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

Kaplan-Meier (KM) estimator (Kaplan and Meier (1958)) is a milestone in survival function estimation and has been widely used in lifetime data analysis.

In a medical study, let T denote the survival time of a subject. Often a d-dimensional baseline covariate $\mathbf{X}$ is collected for each subject, containing such as treatment assignment, age, gender, and genetic information. One main purpose of survival analysis is to characterize the relationship between survival time and explanatory covariates. For example, the conditional survival function (CSF) given certain covariates, defined as $S(t|\mathbf{x}) = P(T > t|\mathbf{X} = \mathbf{x})$, is of primary interest for survival prediction. A variety of regression methods have been proposed over the last few decades, including the Cox proportional hazards (PH) model (Cox (1972)) and the accelerated failure time (AFT) model (Kalbfleisch and Prentice (1980)). The Cox PH model assumes that the hazard ratio for any two different configurations of covariates is constant, and its CSF estimator can be derived using the relationship between the hazard function and the survival function. Though the Cox PH model provides a flexible class of semi-parametric estimators, its estimation consistency relies heavily on the proportional hazards assumption. Also, the CSF estimator obtained from the Cox PH regression is piecewise constant, not continuous. Moreover, when the number of covariates is large, the standard PH procedure is computationally intensive and may fail to produce a reasonable estimate. A variety of penalized Cox regression methods have been studied (Tibshriani (1997); Fan and Li (2002); Zhang and Lu (2007); Zou (2008); Wang, Nan, Zhou and Zhu (2009), and many others), but they all rely on the PH assumption to assure valid estimation.

The AFT model assumes

$$T = \beta_0 + \mathbf{X}^T\boldsymbol{\beta} + \epsilon, \tag{1}$$

where T denotes the true survival time or its known monotone transformation; $\mathbf{X}$ is a d-dimensional covariate vector; ϵ is the random error with an unspecified distribution function F with mean zero and finite variance. The AFT model provides a useful alternative to the Cox PH regression (Kalbfleisch and Prentice (1980); Wei (1992)) thanks to its simplicity and easy interpretation. However, the classical AFT model is predominately fully parametric and (1) only provides estimates of the conditional mean of the survival time T rather than its entire CSF. Although semi-parametric extensions of the AFT model (Buckley and James (1979); Jin, Lin, Wei and Ying (2003); Zeng and Lin (2007); Zhang and Peng (2007)) have been studied,

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

they are not widely used in applications. In addition, when the covariate dimension d is large, the computational cost of fitting the classical AFT model with an unspecified error distribution can be very high.

Motivated by the AFT model (1), a variety of linear quantile regression methods have been proposed (Portnoy (2003); Peng and Huang (2008); Wang and Wang (2009); Leng and Tong (2013)). In this paper, we consider the nonparametric censored quantile regression model:

$$P(T \leq f_\tau(\mathbf{x})|\mathbf{X} = \mathbf{x}) = \tau, \quad \forall \tau \in [0, 1], \tag{2}$$

where the unknown function $f_\tau(\mathbf{x})$ denotes the τth conditional quantile of $T|\mathbf{X} = \mathbf{x}$. Conceptually, one can consider any conditional quantile estimates that solve (2) for all $\tau \in [0, 1]$. However, it is well known that some regression quantiles for censored response may not be estimable when τ is closed to 1 due to lack of information available from the data (Powell (1986); Peng and Huang (2008); Wang and Wang (2009)). We assume that the model (2) is estimable for any $\tau \in [0, \tau_0]$ where $\tau_0 < 1$ denotes the upper limit of estimable quantile levels. In Section 5.3, we discuss how to determine τ_0 from the data.

Based on (2), it is possible to handle more complicated data with heteroscedastic or heavy-tailed error distributions. However the estimation is not straightforward due to the presence of censoring. A natural way of dealing with censored data is to impose a weight on each observed data point, since the censoring essentially provides us biased information about observations that should be taken into account. The inverse censoring probability weighting (ICPW) technique provides a proper weight to adjust the biased information due to censoring. In the literature, the ICPW-based approaches have been used in various applications under the context of classical linear censored regression (Ying, Jung and Wei (1995); Bang and Tsiatis (2002); Zhou (2006), and many others).

We propose a censored kernel quantile regression (CKQR) to fit the model (2). The kernel trick is a widely-used nonparametric technique in machine learning, support vector machines (SVMs) for example. One of its main attractions is that it allows us to deal with a large number of covariates by offering great computational advantages (Zhang (2002); Mallick, Ghosh and Ghosh (2005)). The proposed CKQR is closely related to kernel quantile regression (KQR) that has an L1-type loss subject to a quadratic penalty, and a piecewise linear solution path (Hastie, Rosset, Tibshirani and Zhu (2004); Rosset and Zhu (2007)). There are two quantities associated with the KQR problem: the quantile level τ and the regularization parameter λ that controls the balance between the data fit and the model complexity. Correspondingly, the optimization problem has two types of marginal solution paths: the λ-path as a function of λ with τ being fixed, and the τ-path as a function of τ with λ being fixed. The computation of marginal KQR paths has been studied by Li, Liu and Zhu (2007) and Takeuchi, Nomura and Kanamori (2009). Here, we are not restricted to the marginal paths; instead we study the property of the CKQR solution as a bivariate function of (λ, τ) and establish joint piecewise linearity. This joint piecewise linearity enables us to develop an efficient algorithm to compute the entire two-dimensional CKQR solution surface over the $(\lambda \times \tau)$ plane. This two-dimensional solution surface contains the complete information of the CKQR and therefore greatly facilitates the process of selecting the optimal regularization parameter λ.

After computing $\hat{f}_\tau(\cdot)$ for all $\tau \in [0, \tau_0]$ with an appropriate λ, we can treat $\hat{f}_\tau(\mathbf{x})$ as a continuous function of τ for any given $\mathbf{x}$, and we propose to aggregate the information contained in $\hat{f}_\tau(\mathbf{x})$ to construct an estimator of $S(t|\mathbf{x})$. We show that the proposed CSF estimator is piecewise linear in time t since it is obtained from the piecewise linear solution surface of the CKQR. The new estimator is a flexible nonparametric estimator, and its prediction performance does not depend heavily on the covariate dimension d thanks to the kernel trick.

The rest of the article is organized as follows. In Section 2 we develop the CKQR by employing the ICPW scheme. In Section 3, a piecewise linear CSF estimator is proposed based on the joint piecewise linearity of the CKQR solution. An efficient algorithm for computing the entire CKQR solution surface is described in Section 4. Additional issues regarding the proposed CSF estimator are addressed in Section 5. Simulation studies and data analysis results are shown in Section 6 and 7, respectively. Final discussion follows in Section 8. Technical proofs and details of the solution surface algorithm are in the *supplementary materials*.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

2 Censored Kernel Quantile Regression

Suppose that we have a set of survival data $(Y_i, \delta_i, \mathbf{x}_i), i = 1, \cdots, n$, where $Y_i = \min(T_i, C_i)$, $\delta_i = \mathbb{1}(T_i \leq C_i)$, and $\mathbf{x}_i$ is a d-dimensional covariate vector for the ith subject. Here T_i and C_i denote the survival and censored time or their known monotone transformations, respectively. For identifiability, it is commonly assumed that the censoring time C is conditionally independent of the survival time T given the covariate $\mathbf{X}$, $T \perp C|\mathbf{X}$.

Standard quantile regression is characterized by minimizing the *check loss* function $\rho_\tau(u) = u(\tau - \mathbb{1}\{u \geq 0\})$. Without censoring, a nonparametric quantile regression model can be fitted by solving the optimization problem:

$$\operatorname*{argmin}_{f_\tau \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau \left(T_i - f_\tau(\mathbf{x}_i)\right) + \lambda J(f_\tau), \tag{3}$$

where $\mathcal{F}$ is a function space, J is a functional defined on $\mathcal{F}$ controlling the estimator's complexity to avoid over-fitting, and $\lambda > 0$ is the regularization parameter which balances the data fitting and the model complexity.

In survival analysis, the survival time T is often censored and not completely observed for all individuals. The ICPW scheme is a widely used approach for adjusting possible biases induced by censoring, which utilizes

$$E[\rho_\tau(T - f(\mathbf{X})|\mathbf{X}] = E\left[\frac{\delta}{G(Y|\mathbf{X})}\rho_\tau(Y - f(\mathbf{X}))|\mathbf{X}\right], \tag{4}$$

where $G(t|\mathbf{X}) = P(C \geq t|\mathbf{X})$ denotes the CSF of the censoring time C given covariate $\mathbf{X}$. Equation (4) follows from the assumption $T \perp C|\mathbf{X}$. Let $\hat{G}_n(\cdot|\mathbf{X})$ denote a reasonable estimator of $G(\cdot|\mathbf{X})$ and then, by (4), it is natural to solve

$$\operatorname*{argmin}_{f_\tau \in \mathcal{H}_K} \sum_{i=1}^{n} \frac{\delta_i}{\hat{G}_n(Y_i|\mathbf{x}_i)} \rho_\tau \left(Y_i - f_\tau(\mathbf{x}_i)\right) + \frac{\lambda}{2}\|f_\tau\|^2_{\mathcal{H}_K}, \tag{5}$$

where $\mathcal{H}_K$ is the reproducing kernel Hilbert space (RKHS, Wahba (1990)) generated by a non-negative kernel function $K(\mathbf{x}, \mathbf{x}')$. Here the term $\|\cdot\|^2_{\mathcal{H}_K}$ denotes the squared-norm in RKHS.

Censoring can be regarded as a special case of missing and the ICPW is one type of the inverse

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

probability weighting (IPW) commonly used in missing data analysis. The IPW provides an effective way of correcting or reducing the bias in the complete-case-only analysis. Compared to other methods of handling missing data, the IPW method is generally simpler and does not require rigid model assumptions (Tsiatis (2007) and references therein). The implementation of IPW requires a model for the probability that data are missing, for which a variety of choices are available.

The proposed method based on the ICPW scheme does not require the global linearity assumption, as needed by Portnoy (2003) and Peng and Huang (2008). Ying, Jung and Wei (1995) and Leng and Tong (2013) consider a slightly different weight $\hat{G}_n(f_\tau(\mathbf{x}_i)|\mathbf{x}_i)$ under the linear quantile model $f_\tau(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}_\tau$ to develop unbiased estimating equations. Under the linear quantile regression models, Wang and Wang (2009) introduce a different weighting scheme based on the redistribution-mass idea (Efron (1967)) and propose another way of formulating the censored quantile regressions. By contrast, our method does not depend on the possibly nonlinear quantile function $f_\tau(\mathbf{x})$ of the survival time and thus the weights are constant when estimated properly. Thus the ICPW idea can be naturally embedded in the loss-based quantile regression framework and we can derive the complete solution surfaces due to the joint piecewise linearity. This enables us to recover the entire quantile functionals and to construct a new CSF estimator by aggregating the complete quantile information.

For modeling the missing probability in the ICPW scheme, it is crucial to choose a proper estimator $\hat{G}_n(\cdot|\mathbf{X})$ for $G(\cdot|\mathbf{X})$. One simple choice is the KM estimator of the censored time (Zhou (2006); Shows, Lu and Zhang (2010)), but it ignores the information of $\mathbf{X}$ for modeling $G(\cdot|\mathbf{X})$ and requires the additional assumption $C \perp \mathbf{X}$. The local Kaplan-Meier estimator (Dabrowska, 1989) is another choice (Wang and Wang (2009); Leng and Tong (2013)), but it cannot handle a large d due to the curse of dimensionality. In principle, the proposed CKQR does not rely on the specific choice of $\hat{G}_n(\cdot|\mathbf{X})$ as long as it consistently estimates $G(\cdot|\mathbf{X})$. We refer to Lu and Li (2011) for more discussion of the choice of $\hat{G}_n(\cdot|\mathbf{X})$. We use the Cox PH regression to model the censoring time. The Cox model is more flexible than the parametric AFT model, but it still requires the PH assumption for valid estimation. Based on our limited empirical experiences, the proposed piecewise linear CSF estimator from the Cox-model-based weight is not overly sensitive to model misspecification.

Note that (5) has the form of weighted nonparametric quantile regression with pre-specified weights. Using the *Representer theorem* (Kimeldorf and Wahba (1971)), it is easy to show that the optimizer of (5) has a finite-parameter representation given by

$$f_\tau(\mathbf{x}) = b_\tau + \frac{1}{\lambda}\sum_{j=1}^{n}\theta_{j,\tau}K(\mathbf{x},\mathbf{x}_j). \tag{6}$$

Let $\omega_i = \delta_i/\hat{G}_n(Y_i|\mathbf{x}_i)$ and $\theta_{\tau,0} = \lambda b_\tau$. By plugging (6) into (5), we obtain the optimization problem

$$\underset{\theta_{0,\tau},\theta_{1,\tau},\cdots,\theta_{n,\tau}}{\text{argmin}} \sum_{i=1}^{n}\omega_i\rho_\tau\left(Y_i - \frac{1}{\lambda}\left\{\theta_{0,\tau} + \sum_{j=1}^{n}\theta_{j,\tau}K(\mathbf{x},\mathbf{x}_j)\right\}\right) + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\theta_{i,\tau}\theta_{j,\tau}K(\mathbf{x}_i,\mathbf{x}_j), \tag{7}$$

which we call the *censored kernel quantile regression*. Its minimizer, denoted by $\hat{\boldsymbol{\theta}}_\tau = (\hat{\theta}_{0,\tau},\hat{\theta}_{1,\tau},\cdots,\hat{\theta}_{n,\tau})^T$, is referred to as the *CKQR solution*. The CKQR solution depends on the value of λ and τ, so it can be viewed as a function of (λ,τ). For this reason, we use $\hat{f}(\mathbf{x};\tau,\lambda) = \lambda^{-1}\{\hat{\theta}_{0,\tau} + \sum_{j=1}^{n}\hat{\theta}_{j,\tau}K(\mathbf{x},\mathbf{x}_j)\}$ to denote the estimated conditional quantile function for a given pair of (λ,τ). We next establish that the CKQR solution is jointly piecewise linear over $(\lambda\times\tau)$-plane.

From now on, we may omit the subscript τ in f_τ, b_τ and $\theta_{\tau,j}$, as long as the τ value is fixed and clearly defined from the context.

3 Survival Function Estimation

We show here that the CKQR solution $\hat{\boldsymbol{\theta}}_\tau$ enjoys joint piecewise linearity as a function of (λ,τ), which implies the piecewise linearity of marginal paths. We propose a novel nonparametric estimator of the CSF for a given λ from the estimated conditional quantile function, and show that it is also piecewise linear.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

3.1 Joint Piecewise Linearity of the CKQR solutions

By introducing nonnegative slack variables, we can rewrite (7) as

$$\begin{aligned} \min_{\theta_0,\theta_1,\cdots,\theta_n} \quad & \tau\sum_{i=1}^{n} w_i\xi_i + (1-\tau)\sum_{i=1}^{n} w_i\zeta_i + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\theta_i\theta_j K(\mathbf{x}_i,\mathbf{x}_j) \\ \text{subject to} \quad & -\zeta_i \le Y_i - f(\mathbf{x}_i) \le \xi_i, i = 1,2,\cdots,n, \\ & \xi_i \ge 0, \zeta_i \ge 0, i = 1,2,\cdots,n, \end{aligned}$$

where ξ_i and ζ_i are the nonnegative slack variables. Similar to Li, Liu and Zhu (2007) for the standard KQR, we consider the three sets for any given pair of (λ,τ):

$$\mathcal{E} = \{i : y_i = \hat{f}(\mathbf{x}_i),\ -(1-\tau)w_i \le \hat{\theta}_i \le \tau w_i\} \text{ (elbow)},$$

$$\mathcal{L} = \{i : y_i < \hat{f}(\mathbf{x}_i),\ \hat{\theta}_i = -(1-\tau)w_i\} \text{ (left)},$$

$$\mathcal{R} = \{i : y_i > \hat{f}(\mathbf{x}_i),\ \hat{\theta}_i = \tau w_i\} \text{ (right)}.$$

The three sets and the solution $\hat{\boldsymbol{\theta}}$ change as λ and τ vary; we call it an *event* whenever the configuration of three sets changes.

Theorem 1 states that the CKQR solution moves linearly as long as no event happens, and therefore each component of $\hat{\boldsymbol{\theta}}$ forms a piecewise linear surface over $(\lambda \times \tau)$-plane. Proof of Theorem 1 is relegated to the *supplementary materials*.

Theorem 1. *Let $(\lambda^\ell, \tau^\ell)$ be a point on the $(\lambda \times \tau)$-plane and $\hat{\boldsymbol{\theta}}^\ell = (\hat{\theta}_0, \hat{\theta}_1, \cdots, \hat{\theta}_n)^T$ be the CKQR solution obtained at $(\lambda^\ell, \tau^\ell)$ with the associated sets $\mathcal{E}^\ell$, $\mathcal{L}^\ell$, and $\mathcal{R}^\ell$. If $\mathcal{S}^\ell$ is the largest region on the $(\lambda \times \tau)$-plane containing $(\lambda^\ell, \tau^\ell)$ such that no event happens within $\mathcal{S}^\ell$, then $\hat{\boldsymbol{\theta}}_{0,\mathcal{E}} = \{\hat{\theta}_i, i \in \{0\} \cup \mathcal{E}^\ell, (\lambda,\tau) \in \mathcal{S}^\ell\}^T$ moves as*

$$\hat{\boldsymbol{\theta}}_{0,\mathcal{E}} \equiv \hat{\boldsymbol{\theta}}_{0,\mathcal{E}}(\lambda,\tau) = \hat{\boldsymbol{\theta}}^\ell_{0,\mathcal{E}} + \mathbf{G}_\ell \boldsymbol{\Delta}, \qquad \forall(\lambda,\tau) \in \mathcal{S}^\ell, \tag{8}$$

where $\hat{\boldsymbol{\theta}}^\ell_{0,\mathcal{E}} = \{\hat{\theta}^\ell_i, i \in \{0\} \cup \mathcal{E}^\ell\}^T$ and $\boldsymbol{\Delta} = (\Delta_\lambda, \Delta_\tau)^T = (\lambda - \lambda^\ell, \tau - \tau^\ell)^T$. The gradient matrix $\mathbf{G}_\ell$ is given by

$$\mathbf{G}_\ell = \mathbf{A}_\ell^{-1}\mathbf{B}_\ell = \begin{pmatrix} 0 & \mathbf{1}_\ell^T \\ \mathbf{1}_\ell & \mathbf{K}_\ell \end{pmatrix}^{-1} \begin{pmatrix} 0 & -\sum_{j\notin\mathcal{E}^\ell} w_i \\ \mathbf{y}_\ell & -\mathbf{k}^*_\ell \end{pmatrix}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

Here $\mathbf{K}_\ell = \{K(\mathbf{x}_i, \mathbf{x}_j) : \textit{for } i, j \in \mathcal{E}^\ell\}$*;* $\mathbf{k}^*_\ell = \{\sum_{j \notin \mathcal{E}^\ell} w_j K(\mathbf{x}_i, \mathbf{x}_j) : i \in \mathcal{E}^\ell\}^T$*;* $\mathbf{y}_\ell = \{y_i : i \in \mathcal{E}^\ell\}^T$*;* $\mathbf{1}_\ell$ *is the one vector of length* $|\mathcal{E}^\ell|$*, where* $|S|$ *denotes the cardinality of a set* S*.*

It is easy to update the solution components corresponding to $\mathcal{L}$ or $\mathcal{R}$ from their definitions. The linear updating equation (8) provides us the complete information of the CKQR solutions for any pair $(\lambda, \tau) \in \mathcal{S}^\ell$. Theorem 1 can be regarded as a generalization of one-dimensional piecewise linearity of the KQR marginal solution path, as a function of either λ or τ (separately explored by Li, Liu and Zhu (2007) and Takeuchi, Nomura and Kanamori (2009)).

Corollary 1. *For any given* τ_0*, the solution* $\hat{\boldsymbol{\theta}}_{0,\mathcal{E}}$ *moves linearly in* $\lambda \in \{\lambda : (\lambda, \tau_0) \in \mathcal{S}^\ell\}$ *as*

$$\hat{\boldsymbol{\theta}}_{0,\mathcal{E}} = \hat{\boldsymbol{\theta}}^\ell_{0,\mathcal{E}} + \mathbf{g}^\ell_1 \Delta_\lambda. \tag{9}$$

Similarly, $\hat{\boldsymbol{\theta}}_{0,\mathcal{E}}$ *changes in* $\tau \in \{\tau : (\lambda_0, \tau) \in \mathcal{S}^\ell\}$ *for a given* λ_0 *as*

$$\hat{\boldsymbol{\theta}}_{0,\mathcal{E}} = \hat{\boldsymbol{\theta}}^\ell_{0,\mathcal{E}} + \mathbf{g}^\ell_2 \Delta_\tau, \tag{10}$$

where $\mathbf{g}^\ell_1 = \{g^\ell_{i1} : i \in \{0\} \cup \mathcal{E}^\ell\}^T$ *and* $\mathbf{g}^\ell_2 = \{g^\ell_{i2} : i \in \{0\} \cup \mathcal{E}^\ell\}^T$ *denote the first and second columns of* $\mathbf{G}_\ell$ *in (8), respectively.*

Using the joint piecewise linearity, we can further show that

$$\hat{f}(\mathbf{x}; \tau, \lambda) = \frac{\lambda^\ell}{\lambda}\{\hat{f}(\mathbf{x}; \tau^\ell, \lambda^\ell) - h^\ell_1(\mathbf{x})\} + h^\ell_1(\mathbf{x}) + \frac{\tau - \tau^\ell}{\lambda} h^\ell_2(\mathbf{x}), \tag{11}$$

where

$$\begin{aligned} h^\ell_1(\mathbf{x}) &= g^\ell_{01} + \sum_{i \in \mathcal{E}^\ell} g^\ell_{i1} K(\mathbf{x}, \mathbf{x}_i), \\ h^\ell_2(\mathbf{x}) &= g^\ell_{02} + \sum_{i \in \mathcal{E}^\ell} g^\ell_{i2} K(\mathbf{x}, \mathbf{x}_i) + \sum_{i \notin \mathcal{E}^\ell} w_i K(\mathbf{x}, \mathbf{x}_i). \end{aligned}$$

The quantile function estimate $\hat{f}(\mathbf{x}; \tau, \lambda)$ is not jointly piecewise linear as a function of (λ, τ), but it possesses the marginal piecewise linearity as a function of λ^{-1} or τ respectively, with the other value being fixed.

Rosset (2009) derived a similar result under the unweighted kernel quantile regression that $\hat{f}(\mathbf{x};\tau,\lambda)$ moves piecewise linearly as both λ and τ move together in a linear subspace $\{(\lambda,\tau):\tau=a\lambda+b \text{ with given } a,b\in\mathbb{R}\}$. Our theorem is more general in that it uncovers the complete behaviors of $\hat{\boldsymbol{\theta}}$ and $\hat{f}(\mathbf{x};\lambda,\tau)$ as a function of (λ,τ).

3.2 Piecewise Linear Survival Function Estimator

We propose a nonparametric CSF estimator based on the CKQR solutions described above. First, (11) implies that, given any fixed λ, the conditional quantile moves piecewise linearly in τ:

$$\hat{f}^{\lambda}(\mathbf{x};\tau)=\hat{f}(\mathbf{x};\tau,\lambda)=\hat{f}^{\lambda}(\mathbf{x};\tau^{\ell})+\frac{h_2^{\ell}(\mathbf{x})}{\lambda}(\tau-\tau^{\ell}). \tag{12}$$

Due to the piecewise linearity of $\hat{f}^{\lambda}(\mathbf{x};\tau)$ in (12), the solution path $\{(\tau^{\ell},\hat{f}^{\lambda}(\mathbf{x};\tau^{\ell})):\ell=1,\cdots,m_{\lambda}\}$ contains the complete conditional quantile information of $T|\mathbf{X}=\mathbf{x}$ for all $\tau\in[0,\tau_0]$, where m_{λ} denotes the number of knots in the piecewise linear paths. Any quantile other than those at the path knots can be easily obtained by interpolation. Using the fact that the quantile function is the inverse of the probability function, we propose the CSF estimator:

$$\hat{S}_{\lambda}(t|\mathbf{x})=\begin{cases}1 & \text{if } t\le\hat{f}^{\lambda}(\mathbf{x};0)\\ 1-\left\{\frac{\hat{f}^{\ell+1}-t}{\hat{f}^{\ell+1}-\hat{f}^{\ell}}\tau^{\ell}+\frac{t-\hat{f}^{\ell}}{\hat{f}^{\ell+1}-\hat{f}^{\ell}}\tau^{\ell+1}\right\} & \text{if } t\in\left(\hat{f}^{\ell},\hat{f}^{\ell+1}\right]\\ 1-\tau_0 & \text{if } t>\hat{f}^{\lambda}(\mathbf{x};\tau_0),\end{cases} \tag{13}$$

where $\hat{f}^{\ell}$ and $\hat{f}^{\ell+1}$ denote $\hat{f}^{\lambda}(\mathbf{x};\tau^{\ell})$ and $\hat{f}^{\lambda}(\mathbf{x};\tau^{\ell+1})$, respectively.

An advantage of the proposed estimator is that it can handle data with a moderately large d due to the employment of the kernel trick. The estimator can also be used for data from a heterogeneous or a heavy-tailed conditional survival time distribution of $T|\mathbf{X}=\mathbf{x}$.

3.3 Asymptotic Property

We carry out asymptotic analysis for the proposed estimator. In particular, the uniform convergence of the risk of the CKQR quantile estimator is established, providing theoretical justifications for the proposed CSF estimator. Under no censoring, we could consider the standard

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

quantile risk

$$R^*(f;\tau) = E\left[\rho_\tau(T - f_\tau(\mathbf{x}))|\mathbf{X} = \mathbf{x}\right] \tag{14}$$

that is minimized by the τth conditional quantile of $T|\mathbf{X} = \mathbf{x}$, denoted by $f_\tau^* = \operatorname{argmin}_f R^*(f;\tau)$. However, since the risk is not feasible due to presence of censoring, we consider the weighted quantile loss,

$$\varphi(\mathbf{Z}; f, \tau) = \frac{\delta}{G(Y|\mathbf{X})}\rho_\tau(Y - f_\tau(\mathbf{X})), \tag{15}$$

where $\mathbf{Z} = (Y, \delta, \mathbf{X})$. Now $R(f, \tau) = E\left[\varphi(\mathbf{Z}; f, \tau)|\mathbf{X} = \mathbf{x}\right]$ is identical to $R^*(f;\tau)$ in (14) by (4). The proposed CKQR minimizes

$$\hat{R}_{n,\text{reg}}(f;\tau) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{G}_n(Y_i|\mathbf{x}_i)}\rho_\tau(Y_i - f_\tau(\mathbf{x}_i)) + \frac{\alpha_n}{2}\|f_\tau\|^2_{\mathcal{H}_K}, \tag{16}$$

which is identical to (7) by letting $\alpha_n = \lambda/n$. Let $\hat{f}_\tau$ be the estimated quantile function from the CKQR solution, the minimizer of $\hat{R}_{n,\text{reg}}(f;\tau)$.

Theorem 2 states that, the entire trajectory of the estimated conditional quantile gets closer to the true one as n increases, in the sense that the associated risk (14) of $\hat{f}_\tau$ converges to that of f_τ uniformly over $\tau \in [0, \tau_0]$. This implies that the τ-path can be regarded as a reasonable quantile function estimate, which justifies the proposed survival function (13) due to their inverse relationship. The sketch of the proof is provided in the *supplementary materials*.

Theorem 2. *Assume that*

(A1) The quantile regression function $f_\tau(\mathbf{x})$ *is identifiable for all* $\tau \in [0, \tau_0]$

(A2) $\|\sup_{\mathbf{x}} K(\cdot, \mathbf{x})\|^2_{\mathcal{H}_K} < \infty$ *and* $\|f\|^2_{\mathcal{H}_K} < \infty$*;*

(A3) There exists a constant κ *such that* $P(C = \kappa) > 0$ *and* $P(C > \kappa) = 0$*;*

(A4) $\sup_{\mathbf{x}} \sup_{t\in[0,\kappa]} |\hat{G}_n(t|\mathbf{X}) - G(t|\mathbf{X})| \to 0$ *almost surely.*

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

Under (A1) – (A4), we have

$$\sup_{\tau\in[0,\tau_0]}\left|R^*(\hat{f}_\tau;\tau)-R^*(f^*_\tau;\tau)\right| \ \rightarrow\ 0 \textit{ almost surely.} \tag{17}$$

The first condition (A1) here states the underlying identifiability condition of the regression function. The regularity condition (A2) is quite standard in RKHS theory. (A3) is valid in many clinical studies with an administrative censoring and it simplifies theoretical arguments by ensuring Y is bounded. In practice, the maximum of follow-up can be used as κ. Condition (A4) states the strong uniform consistency of the censoring time survival function estimator. For example, the KM estimator enjoys the strong uniform consistency (Stute and Wang (1993)). If the PH model is correctly specified for the censoring time, then the strong uniform consistency of the survival function estimator from the Cox PH regression follows the strong consistency of the regression coefficient estimator (Tsiatis (1981); Andersen and Gill (1982)) and the uniform strong consistency of the cumulative baselines hazard estimator (Kosorok (2007)).

4 Computational Algorithm for Two-Dimensional Solution Surface

Joint piecewise linearity enables us to develop an efficient algorithm to compute the entire CKQR solution surface. Our algorithm iteratively identifies $\mathcal{S}^\ell$ on the $(\lambda\times\tau)$-plane and updates the solutions at the boundaries of $\mathcal{S}^\ell$ by applying Theorem 1. A key issue is how to accurately and efficiently identify $\mathcal{S}^\ell$. In fact, $\mathcal{S}_\ell$ is a convex polygon that satisfies the linear constraints

$$\begin{aligned}
\forall i\in\mathcal{E}^\ell: &\quad u_i^\ell-\omega_i\le g_{i1}^\ell\lambda+(g_{i2}^\ell-w_i)\tau\le u_i^\ell,\\
\forall i\in\mathcal{L}^\ell: &\quad \left\{y_i-h_1^\ell(\mathbf{x}_i)\right\}\lambda-h_2^\ell(\mathbf{x}_i)\tau\le v_i^\ell,\\
\forall i\in\mathcal{R}^\ell: &\quad \left\{y_i-h_1^\ell(\mathbf{x}_i)\right\}\lambda-h_2^\ell(\mathbf{x}_i)\tau\ge v_i^\ell,
\end{aligned}$$

where $u_i^\ell=g_{i1}^\ell\lambda^\ell+g_{i2}^\ell\tau^\ell-\theta_i^\ell$ and $v_i^\ell=\left\{\hat{f}^\ell(\mathbf{x}_i)-h_2^\ell(\mathbf{x}_i)\right\}\lambda^\ell-h_2^\ell(\mathbf{x}_i)\tau^\ell$. The basic idea of the algorithm for computing the CKQR solution surface is similar to that for the weighted support vector machine (WSVM) developed by Shin, Wu and Zhang (2014), where the computational

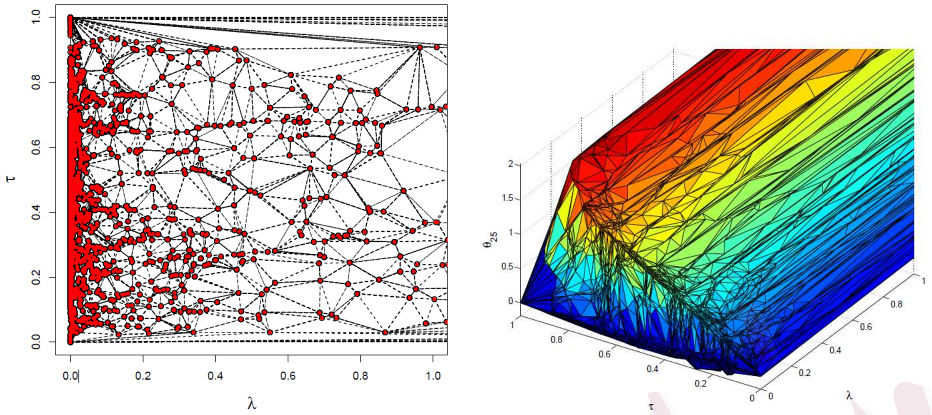
Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

complexity of the two-dimensional solution surface algorithm is rigorously explored. We relegate the details to the *supplementary materials*.

We use the *lung* data set in `survival` package in `R` (Loprinzi, Laurie, Wieand, Krook, Novotny, Kugler, Bartel, Law, Bateman and Klatt (1994)) to illustrate the proposed algorithm. The lung data set contains the survival times of 228 patients with advanced lung cancer from the North central cancer treatment group. We discarded 61 patients with at least one missing covariate. Of the 167 patients, 120 died and 47 survived during the study period. There were eight covariates measured at diagnosis, including the institution diagnosed at, age in years, gender, ECOG performance score, two versions of Karnofsky performance scores, calories consumed at meals, and weight loss in last six months. We used the radial kernel $K(\mathbf{x},\mathbf{x}') = \exp\{-\|\mathbf{x}-\mathbf{x}'\|^2/(2\sigma^2)\}$, where the bandwidth parameter σ was set to be the median pairwise distance of the predictors for uncensored data. It took 2.0525 minutes (on a PC equipped with i5-3590 CPU 3.40GHz and 4GB RAM) to compute the entire solution surface which consists of 38,777 vertices and 9,649 sets of $\mathcal{S}^\ell$, i.e., the pieces of the linear surfaces. In Figure 1, (a) depicts the $\mathcal{S}^\ell$s. The x-axis is λ (truncated at 1 for better visualization) and the y-axis is τ. The red dots and the (dashed) lines represent vertices and edges of all $\mathcal{S}^\ell$s produced during the algorithm. Here (b) is a three-dimensional plot of the piecewise linear solution surface of $\hat{\theta}_{25}$ over the $(\lambda \times \tau)$-plane. The x-, y-, and z-axis are λ, τ and θ_{25}, respectively. In fact, (a) can be regarded as a projection of the solution surface in (b) on the $(\lambda \times \tau)$-plane. Since there are 120 uncensored observations, we have solution surfaces of $\theta_1, \cdots, \theta_{120}$ in total. The other 119 solutions can be depicted as was θ_{25} in (b).

The proposed two-dimensional solution surface algorithm, as well as Theorem 1 and Corollary 1, are not restricted to the CKQR. These results are quite general and also hold for any WKQR problem with arbitrary non-negative weights. Censored observations have no effect on solving the CKQR problem once they are used for estimating $\hat{G}_n(\cdot|\mathbf{x})$, since $\omega_i = 0$ for all censored observation. It is thus enough to apply the algorithm to the reduced set of uncensored observations. Finally, it is straightforward to obtain a marginal path from the two-dimensional solution surface, making it convenient to tune λ as described in Section 5.1.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071



(a) Vertices of $\mathcal{S}^\ell$ (b) A CKQR Solution Surface of θ_{25}

Figure 1: Two-dimensional solution surfaces of CKQR for the *lung* data: (a) shows vertices and edges of the $\mathcal{S}^\ell$s produced from the proposed algorithm and (b) depicts the solution surface of $\hat{\theta}_{25}$.

5 Additional Issues

5.1 Tuning λ

The estimator $\hat{S}_\lambda(t|\mathbf{x})$ depends on the value of the regularization parameter λ, since the CKQR solution depends on both λ and τ. The choice of λ plays a crucial role in finite-sample performance. We propose a systematic way to select the optimal λ using cross validation.

We first define a proper tuning criterion. The conditional density estimator of $T|\mathbf{X} = \mathbf{x}$, denoted by $\hat{p}_\lambda(t|\mathbf{x})$ or simply $\hat{p}_\lambda$, can be derived from $\hat{S}_\lambda(t|\mathbf{x})$ using the relationship $S_\lambda(t|\mathbf{x}) = \int_t^\infty p_\lambda(u|\mathbf{x})du$. We use the Kullback-Leibler (KL) loss for $\hat{p}_\lambda$ as a tuning criterion. The KL loss is $L(p, \hat{p}_\lambda) = E\left[\log(p/\hat{p}_\lambda)\right]$, where p denotes the unknown true density of interest. Since minimizing the KL loss is equivalent to maximizing $E\left[\log(\hat{p}_\lambda)\right]$, one could maximize its empirical counter part, the conditional log-likelihood $\sum_{i\in\{i:\delta_i=1\}} \log \hat{p}_\lambda(y_i|\mathbf{x}_i)$ for uncensored observations, where the density $\hat{p}_\lambda(y|\mathbf{x})$ is obtained by differentiating the piecewise linear survival function estimate (13) and is therefore piecewise constant. As this fails to utilize information within censored observations, we use the complete likelihood of the censored data as a tuning criterion,

$$\sum_{i=1}^n \delta_i \log \hat{p}_\lambda(y_i|\mathbf{x}) + (1-\delta_i) \log \hat{S}_\lambda(y_i|\mathbf{x}), \tag{18}$$

and propose to select the optimal λ_{opt} by maximizing (18). A grid search is often employed to choose the optimal λ. To achieve this, we exploit the two-dimensional solution surface obtained by the algorithm given in Section 4. Similar to Shin, Wu and Zhang (2014) for the WSVM, we use all the distinctive λ values of the vertices of the $\mathcal{S}^\ell$s obtained during the algorithm as a grid. One advantage about the grid based on the two-dimensional surface, when compared to the regular lattice grid, is that the proposed grid is adaptive in the sense that the coarseness of the grid is automatically controlled by the complexity of the two-dimensional solutions. If the solution surface is complicated, the grid would be fine, and coarse if the solution surface is relatively simple.

Rather than (18), the (empirical) quantile risk could be considered as an alternative. In the standard KQR, Rosset (2009) developed an algorithm to track the path of $\hat{f}(\mathbf{x}; \lambda^*, \tau)$ as a function of τ, where λ^* minimizes the cross-validated quantile risk for a given τ.

5.2 Violation of Monotonicity

The monotone decreasing property is an essential feature of any survival function and should be satisfied by its estimator. However, we may have an estimate $\hat{S}_\lambda(\cdot|\mathbf{x})$ such that $\hat{S}_\lambda(t_1|\mathbf{x}) > \hat{S}_\lambda(t_2|\mathbf{x})$ for some $t_1 > t_2$. This happens due to the so-called *quantile crossing* (He (1997)). Quantile crossing is frequently encountered in nonparametric quantile regression. It occurs when $\hat{f}(\mathbf{x}; \tau_1) > \hat{f}(\mathbf{x}; \tau_2)$ for some $\mathbf{x}$ and quantile levels $\tau_1 < \tau_2$. Quantile crossing makes the estimated quantile function not invertible and thus the associated survival function cannot be properly defined. In the context of the standard quantile regression, Rosset (2009) proposed a simple remedy for quantile crossing by taking $\hat{f}^\lambda(\mathbf{x}; \tau_1)$ to be the same as $\hat{f}^\lambda(\mathbf{x}; \tau_2)$ (or vice versa). A similar idea can be applied here. If the conditional quantile path as a function of τ is decreasing in that region, we make the curve flat within the region.

Proposition 1. *For any given λ, if $\tau_1 < \tau_2$ but $\hat{f}^\lambda(\mathbf{x}; \tau_1) > \hat{f}^\lambda(\mathbf{x}; \tau_2)$ then either*

$$E\left[\rho_{\tau_1}\left(T - \hat{f}^\lambda(\mathbf{x}; \tau_1)\right)|\mathbf{X} = \mathbf{x}\right] \geq E\left[\rho_{\tau_1}\left(T - \hat{f}^\lambda(\mathbf{x}; \tau_2)\right)|\mathbf{X} = \mathbf{x}\right] \quad \textit{or}$$
$$E\left[\rho_{\tau_2}\left(T - \hat{f}^\lambda(\mathbf{x}; \tau_1)\right)|\mathbf{X} = \mathbf{x}\right] \leq E\left[\rho_{\tau_2}\left(T - \hat{f}^\lambda(\mathbf{x}; \tau_2)\right)|\mathbf{X} = \mathbf{x}\right].$$

Proposition 1 essentially states that the quantile risk is not escalated after such a correction.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

This is a direct extension of Proposition 7 in Rosset (2009), and the proof is omitted.

5.3 Non-identifiability Issue with τ close to 1

There is an issue that the upper regression quantile with τ close to 1 may not be estimable due to the loss of information caused by censorship. This can be true when all the censoring happens before $f_\tau(\mathbf{x})$ and there is not enough information available at or after $f_\tau(\mathbf{x})$. Wang and Wang (2009) provide a sufficient condition for $f_\tau(\mathbf{x})$ being estimable; $G(f_\tau(\mathbf{x})) > 0$. In order to check the condition at a given τ, consider $\hat{H}_n(\tau) = \{i : \hat{G}_n(\hat{f}_\tau(\mathbf{x}_i)|\mathbf{x}_i) = 0, i = 1, \cdots, n\}$. If there are not many such cases, $|\hat{H}_n(\tau)|/n$ is smaller than a pre-specified cutoff value $\gamma \in (0,1)$, then $f_\tau(\mathbf{x})$ is said to be estimable. For any given $\mathbf{x}$ we have a complete solution path of $\hat{f}_\tau(\mathbf{x})$ as a function of τ, and hence are able to restrict our attention to $[0, \hat{\tau}_0]$ where $\hat{\tau}_0 = \sup_\tau\{\tau : |\hat{H}_n(\tau)| \leq \gamma n\}$. Peng and Huang (2008) discussed how to choose τ_0 in practice.

6 Numerical Results

In this section, we report on numerical experiments to evaluate finite-sample performance of the proposed CSF estimator under various scenarios. We first generated the vector of covariates $\mathbf{x} \overset{iid}{\sim} N_d(\mathbf{0}, \mathbf{I})$, with $\mathbf{0}$ and $\mathbf{I}$, a d-dimensional zero vector and identity matrix, respectively. For the survival time T, we took

$$\log T = f(\mathbf{x}) + v(\mathbf{x})\epsilon, \tag{19}$$

which includes the AFT model as a special case. Different choices of the error distribution for ϵ with $v(\mathbf{x}) = 1$ leads to some common survival models: extreme value distribution for the PH model and logistic distribution for the proportional odds (PO) model.

In particular, our experiments include four different error distributions for ϵ: the standard extreme value, logistic, standard normal, and t distribution with degrees of freedom of 5; they are denoted by PH, PO, NR, and $t(5)$, respectively in the tables; two different shapes for $f(\mathbf{x})$: a linear shape with $f_1(\mathbf{x}) = \mathbf{x}^T\beta$, a non-linear shape $f_2(\mathbf{x}) = \cos(\mathbf{x}^T\beta)$, where the coefficient vector $\boldsymbol{\beta} = (1, \cdots, 1)^T/\sqrt{d}$; two different shapes for $v(\mathbf{x})$: the homoscedastic error $v_1(\mathbf{x}) = 1$, the heteroscedastic error, $v_2(\mathbf{x}) = \mathbf{x}^T\mathbf{x}/d$; two different input dimensions: $d = 3, d = 50$. The sample size and the censoring level were fixed at $n = 200$ and $P(\delta = 0) = 0.30$, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

We denote the conditional survival function estimator derived from CKQR by PLE. The methods under comparison included the parametric AFT model (1) with the extreme value distribution (AFT-PH), logistic distribution (AFT-PO), standard normal distribution (AFT-NR)), the standard Cox regression (Cox) method, and a smoothing-spline based nonparametric hazard regression (NP-Cox) of Leng and Zhang (2006). The AFT and Cox PH models were fitted with functions in the R-{survival} package. The method of Leng and Zhang (2006) was implemented by the R-{cosso} package. For the proposed CSF estimator, a radial kernel was employed; the associated bandwidth σ was set as in Section 4.

To evaluate performance of the various methods, we further generated an independent test set $\{Y_k, \delta_k, \mathbf{x}_k\}, k = 1, \cdots, \tilde{n}$, with $\tilde{n} = 1000$. The estimators were evaluated by $\bar{D}_{RISE} = \tilde{n}^{-1}\sum_{k=1}^{\tilde{n}} D_{k,RISE}$, where $D_{k,RISE}$ is the root integrated squared error (RISE) of the k subject in the test set,

$$D_{k,RISE} = \sqrt{\int_0^\infty \left[\hat{S}(t|\mathbf{x}_k) - S(t|\mathbf{x}_k)\right]^2 dt}. \tag{20}$$

The integration in $D_{k,RISE}$ was numerically computed over a fine grid on 99.9% of the support of the random variable $T|\mathbf{x}_k, k = 1, \cdots, \tilde{n}$.

The censoring time model used for ICPW is an important factor that affects performance of the proposed method. We took $C = \exp\left\{\gamma^T\mathbf{x} + \epsilon'\right\} - \Delta$, where $\gamma = (1, \cdots, 1)^T/\sqrt{d}$ and a non-random constant Δ controled the censoring level. We considered distribution of the random error for the censoring time ϵ' to be the extreme value (PH), logistic (PO), and standard normal distribution (NR). We used Cox PH regression to estimate $\hat{G}_n(\cdot|\mathbf{x})$; this is valid only when ϵ' has an extreme value distribution.

Table 1 summarizes the performance of different methods in terms of $\bar{D}_{RISE}$ when ϵ' follows the extreme value and $\hat{G}_n(\cdot|\mathbf{x})$ is estimated under the correctly specified model. With (f_1, v_1), the AFT model assumption is satisfied, so it is not surprising that the AFT model with the true error distribution performs the best. For example, the AFT model under the extreme value distribution error (AFT-PH) outperforms all of the others. However, the AFT model's performance is not good when $f(\mathbf{x})$ is highly nonlinear. The AFT assumes iid errors, so the model also suffers when $h(\mathbf{x})$ is not constant. This is echoed in Table 1.

The two Cox PH models do not perform well when the PH assumption is violated. The NP-Cox shows better performance than the standard Cox model when d is large due to the use

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

p	(f, v)	ϵ	AFT-PH	AFT-PO	AFT-NR	Cox	NP-Cox	PLE
3	(f_1, v_1)	PH	0.367	0.598	0.699	0.551	0.997	0.911
		PO	0.982	0.410	0.441	0.676	1.061	0.970
		NR	0.734	0.464	0.433	0.764	1.259	1.047
		t(5)	0.578	0.382	0.463	0.679	1.013	0.911
	(f_1, v_2)	PH	1.831	1.954	2.300	1.764	2.047	1.294
		PO	1.681	1.645	1.883	1.705	1.916	1.412
		NR	1.975	1.758	1.954	1.906	2.154	1.446
		t(5)	1.734	1.605	1.875	1.728	2.019	1.280
	(f_2, v_1)	PH	0.955	1.087	1.171	1.034	1.238	0.943
		PO	1.246	0.729	0.761	0.878	1.167	0.999
		NR	1.344	1.222	1.221	1.290	1.526	1.082
		t(5)	1.030	0.938	0.991	1.037	1.182	0.935
	(f_2, v_2)	PH	2.274	2.422	2.701	2.144	1.964	1.359
		PO	1.868	1.845	2.075	1.681	1.892	1.423
		NR	2.297	2.217	2.383	2.045	2.100	1.463
		t(5)	2.010	1.958	2.193	1.779	1.972	1.298
50	(f_1, v_1)	PH	1.717	1.856	1.898	1.899	2.425	1.769
		PO	1.902	1.668	1.650	1.921	2.139	1.490
		NR	1.963	1.901	1.819	2.155	2.898	2.079
		t(5)	1.685	1.613	1.643	1.867	2.411	1.674
	(f_1, v_2)	PH	1.839	1.945	2.026	1.975	2.511	1.849
		PO	1.925	1.740	1.747	1.986	2.114	1.563
		NR	2.070	1.971	1.919	2.229	2.993	2.135
		t(5)	1.773	1.687	1.738	1.936	2.463	1.722
	(f_2, v_1)	PH	2.002	1.906	1.920	2.131	1.803	1.245
		PO	2.031	2.123	2.147	2.150	1.722	1.255
		NR	2.109	1.822	1.799	2.081	2.008	1.219
		t(5)	2.350	2.285	2.208	2.454	1.779	1.556
	(f_2, v_2)	PH	2.143	2.230	2.289	2.229	2.189	1.371
		PO	2.130	1.898	1.904	2.142	1.886	1.303
		NR	2.420	2.356	2.306	2.525	1.840	1.633
		t(5)	2.054	1.968	2.002	2.171	1.863	1.305

Table 1: Averaged $\bar{D}_{RISE}$ over 100 independent repetitions for different scenarios when the censoring time error satisfies the PH assumption: The proposed PLE outperform or at least comparable to all other competing methods except the case with $p = 3$ and (f_1, v_1). Five number summary of the MC standard deviations of each cells is $(0.074, 0.134, 0.193, 0.242, 0.971)$.

p	(f, v)	ϵ	PO				NR			
			AFT	Cox	NPCox	PLE	AFT	Cox	NPCox	PLE
3	(f_1, v_1)	PH	0.365	0.547	1.008	0.908	0.362	0.553	0.999	0.994
		PO	0.410	0.661	1.080	0.930	0.401	0.691	1.164	1.101
		NR	0.435	0.777	1.308	1.034	0.432	0.764	1.241	1.202
		t(5)	0.385	0.700	1.050	0.906	0.394	0.694	1.226	1.065
	(f_1, v_2)	PH	1.861	1.802	1.973	1.300	1.831	1.762	1.849	1.404
		PO	1.661	1.714	1.998	1.396	1.638	1.706	1.843	1.548
		NR	1.976	1.975	2.147	1.448	1.944	1.884	2.133	1.578
		t(5)	1.622	1.789	1.982	1.295	1.611	1.726	1.968	1.383
	(f_2, v_1)	PH	0.956	1.033	1.282	0.929	0.955	1.033	1.271	1.023
		PO	0.729	0.820	1.098	0.884	0.728	0.881	1.170	1.052
		NR	1.218	1.275	1.595	1.066	1.216	1.299	1.490	1.207
		t(5)	0.937	1.005	1.227	0.891	0.941	1.067	1.296	1.040
	(f_2, v_2)	PH	2.293	2.147	2.024	1.352	2.271	2.143	1.997	1.410
		PO	1.854	1.672	1.903	1.368	1.846	1.698	1.872	1.509
		NR	2.390	2.047	2.208	1.452	2.384	2.048	2.150	1.568
		t(5)	1.963	1.774	1.854	1.261	1.968	1.789	1.959	1.380
50	(f_1, v_1)	PH	1.706	1.909	2.458	1.750	1.719	1.874	2.387	1.817
		PO	1.687	1.941	2.033	1.425	1.649	1.918	2.101	1.551
		NR	1.843	2.182	2.892	2.062	1.767	2.109	2.938	2.118
		t(5)	1.642	1.909	2.342	1.652	1.575	1.828	2.345	1.710
	(f_1, v_2)	PH	1.825	1.980	2.478	1.816	1.837	1.949	2.478	1.887
		PO	1.764	2.019	2.182	1.506	1.722	1.991	2.160	1.628
		NR	1.939	2.250	2.976	2.115	1.857	2.159	2.967	2.168
		t(5)	1.714	1.972	2.436	1.706	1.654	1.903	2.406	1.771
	(f_2, v_1)	PH	2.022	2.161	1.820	1.207	2.005	2.107	1.850	1.326
		PO	1.840	2.049	1.643	1.076	1.795	2.073	1.786	1.308
		NR	2.227	2.456	2.080	1.472	2.171	2.417	2.036	1.618
		t(5)	1.925	2.132	1.791	1.161	1.881	2.134	1.812	1.295
	(f_2, v_2)	PH	2.132	2.237	1.883	1.316	2.119	2.201	1.858	1.425
		PO	1.921	2.123	1.862	1.165	1.889	2.139	1.741	1.386
		NR	2.325	2.535	2.215	1.552	2.286	2.507	2.109	1.693
		t(5)	1.983	2.166	1.783	1.237	1.958	2.190	1.812	1.357

Table 2: Averaged $\bar{D}_{RISE}$ over 100 independent repetitions for different scenarios when the censoring time errors do not satisfy the PH assumption: The results are similar to the case when the censoring time is correctly specified, meaning that the proposed PLE is not overly sensitive against the model misspecification for C. Five number summary of the MC standard deviations of each cells is $(0.076, 0.130, 0.184, 0.233, 1.030)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

of penalization. Furthermore, it is observed that the Cox PH regression does not perform so well as expected under the PH assumption with homogeneous variance (v_1), regardless of the shape of the mean function. One reason for this is that some survival functions from the Cox regression never touch zero even for a very large survival time due to the censoring. In this regard, $\bar{D}_{RISE}$ (20) is not a favorable measure for the Cox regression estimator although the support for integration is truncated.

We observe that the proposed estimator (PLE) generally outperforms the competing methods in terms of $\bar{D}_{RISE}$, except in the simplest case with $p = 3$ and (f_1, v_1). In addition, its performance is quite stable over various scenarios, while other methods depend on the underlying data generating structure which is unknown *a priori*. This is not surprising because our estimator is constructed from the flexible CKQR solution, which does not heavily rely on the particular model assumption. Finally, the PLE performs very well for $d = 50$, since the CKQR exploits the kernel trick and is stable even for a large-dimensional covariate (Zhang (2002)).

Table 2 reports performance of the methods when ϵ' follows a logistic or normal distribution and the censoring time C does not satisfy the PH assumption. To avoid redundancy, we only report the best AFT model fit among the three. Throughout all the experiment settings, the proposed PLE performs very well compared to other methods, even when C does not satisfy the PH assumption, implying that the proposed estimator is not overly sensitive to the model misspecification of the censoring time.

In summary, the proposed piecewise linear survival function estimator is promising in practice when the covariate dimension d is moderately large, and/or when there is not enough information about underlying data structure.

7 Data Analysis

In this section, we revisit the lung cancer data set to demonstrate performance of the proposed piecewise linear CSF estimator. Based on the two-dimensional solution surface in Section 4, we obtained an adaptive tuning grid for λ and the corresponding survival function estimator $\hat{S}_\lambda(t; \mathbf{x})$ defined in (13). We then applied a leave-one-out cross validation (LOOCV) to tune λ. The cross-validated conditional log-likelihood (18) is plotted for different values of λ in Figure

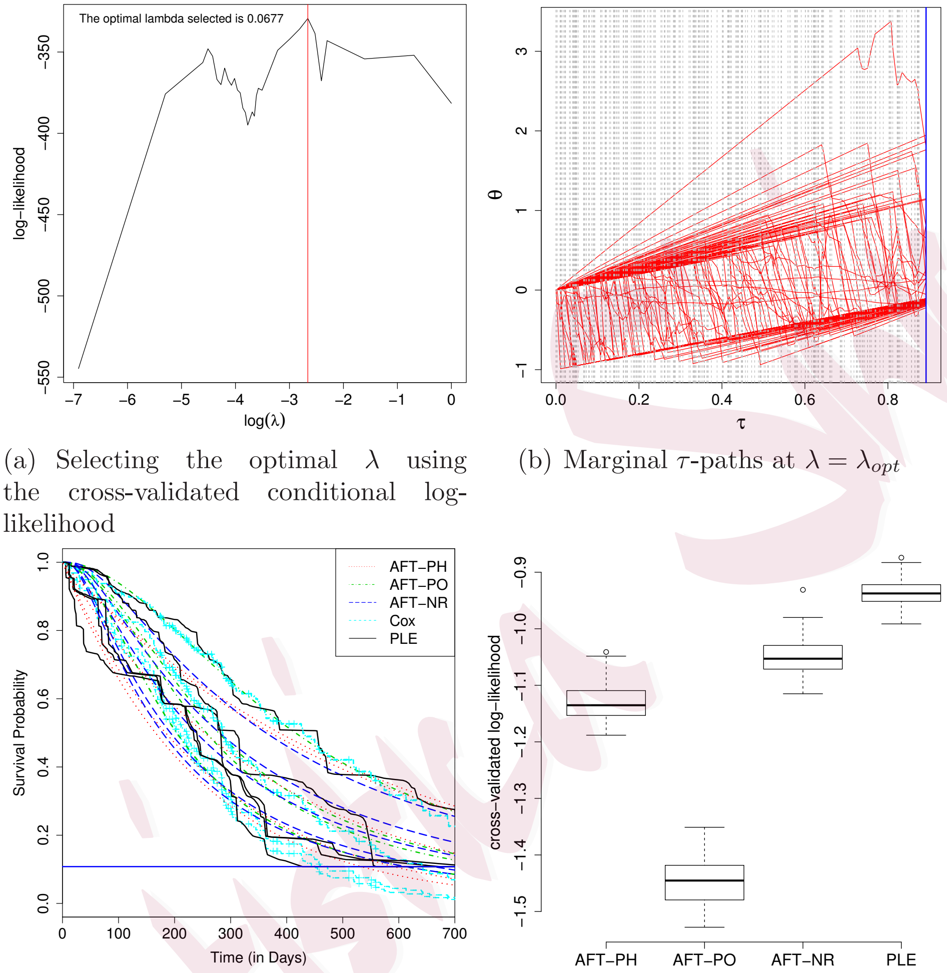
Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

2-(a), and the likelihood is maximized at $\lambda_{\text{opt}} = 0.0677$ (red vertical line). In Figure 2–(b), the τ-paths of $\hat{\theta}_1, \cdots, \hat{\theta}_{120}$ at the optimal λ_{opt} for $\tau \in [0, \tau_0]$ are depicted. The τ-paths are simply cross-sections at λ_{opt} of the two-dimensional solution surfaces, and are piecewise linear, as shown by Corollary 1. We estimate the upper limit of the estimable quantile level $\hat{\tau}_0 = 0.892$ with a cutoff value of $\gamma = 0.05$ (the blue solid vertical line at the right-end). Panel (c) illustrates the proposed piecewise linear estimates for the first five patients in the data set, along with the estimates given by the competing methods described in Section 6. The proposed PLE estimates are horizontally cut at $1 - \hat{\tau}_0 = 0.108$ (marked by the blue horizontal line) below which the function value is not estimable.

In order to evaluate the performance of the proposed method, we further carried out the LOOCV and report the cross-validated log-likelihood. The Cox-regression-based models cannot be evaluated in terms of the log-likelihood since they return piecewise constant survival function estimates, which often gives zero values of likelihood, and are therefore not considered here. For the same reason, it is not fair to directly compare the proposed piecewise estimator to the smooth ones from the AFT model. To overcome this, we consider an additional step to make the associated CSF estimates smooth: we first generate N random samples from the estimated survival function, then compute the kernel density estimate (KDE) from these N samples. The likelihood can be evaluated based on the smooth KDEs. Although the AFT models do not require such an additional step, we applied it for a fair comparison. This additional step cannot be applied to the estimate from the Cox model, which often provides incomplete survival curves in the sense that some of them may fail to converge to zero even for large T. The cross-validated log-likelihood evaluated from the KDE is a random quantity, and its standard error gets smaller as N increases; we set N sufficiently large, say $10,000$. Figure 2–(d) depicts the boxplot of the LOOCV log-likelihoods. It is clear that the PLE performs significantly better than the AFT models.

8 Discussion

We develop nonparametric quantile regression in RKHS and propose a survival function estimator for censored data analysis. The new nonparametric estimator works very well even in the



(a) Selecting the optimal λ using the cross-validated conditional log-likelihood

(b) Marginal τ-paths at $\lambda = \lambda_{opt}$

(c) The estimated CSFs by different methods

(d) Boxplots of LOOCV log-likelihoods of different methods

Figure 2: Conditional survival function estimator for *lung* Data: the panel (a) shows cross-validated conditional log-likelihood for different λ and the (red) vertical line represents the selected optimal λ_{opt}; (b) depicts solution paths of $\hat{\theta}_1, \cdots, \hat{\theta}_{48}$ as a function $\tau \in [0, \tau_0]$ at λ_{opt} where τ_0 is estimated by 0.892 at a cutoff value of $\gamma = 0.05$ (blue solid vertical line); (c) shows estimated CSFs for the first five patients and the results look similar; (d) depicts boxplots of LOOCV log-likelihoods computed from different CSF estimates and PLE outperforms all others.

case of large-dimensional covariates or heteroscedastic errors, showing favorable finite-sample performance. The asymptotic analysis provides a theoretical justification for the proposed approach.

It is not clear yet how to build a reasonable confidence band of the survival function estimator obtained from the CKQR solution surface. The asymptotic distribution of the quantile for any given τ can be possibly derived by extending the asymptotic results for SVM (Jiang, Zhang and Cai (2008); Li, Artemiou and Li (2011)). What we need, however, is the variability of survival probability for any given time, t. This is not straightforward and is worth further exploration. Another interesting problem for the proposed estimator is how to incorporate variable selection in the estimation. Penalized regression is an appealing approach, but there are some difficulties in this context. For each fixed τ and λ, one can conduct variable selection using some shrinkage methods, but how to assemble sparse estimators at different quantile levels to obtain an overall sparse survival function estimator is an open question. This is one of the research directions in our follow-up investigation.

In practice, the proposed algorithm for computing the two-dimensional solution surface may be slow when n is large. Instead we can use the marginal solution path algorithm developed by Takeuchi, Nomura and Kanamori (2009). In this case, the exhaustive grid search for λ based on cross-validation may be infeasible due to heavy computation. Our experiences suggest setting $\lambda = cn$ for a value of c smaller than .1.

Supplementary Materials

Proofs of Theorem 1 and 2, and the two-dimensional solution surface algorithm for the censored kernel quantile regression can be found in the supplementary materials.

Acknowledgements

We are grateful to Dr. Wenbin Lu for helpful discussions. We also thank the Co-Editor Dr. Zhiliang Ying, an associate editor, and two referees for their constructive comments and suggestions, which have greatly improved the paper. Zhang's research is partially supported by NSF grants DMS-1309507 and DMS-1418172, and National Natural Science Foundation of China grant

NSFC-11571009. Wu's research is partially supported by NSF grant DMS-1055210, NIH/NCI grants P01-CA142538 and R01-CA149569.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

Bibliography

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–1120.

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

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–436.

Cox, D. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187–220.

Dabrowska, D. M. (1989). Uniform of the kernel conditional Kaplan-Meier estimate. *Ann. Statist.* **17**, 1157–1167.

Efron, B. (1967). The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 4. 831–853. Prentice-Hall Engewood Cliffs, NJ.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74–99.

Hastie, T., Rosset, R., Tibshirani, R., and Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Reasearch* **5**, 1931–1415.

He, X. (1997). Quantile curves without crossing. *American Statistician* **51**, 186–192.

Jiang, B., Zhang, X., and Cai, T. (2008). Estimating the confidence interval for prediction erros of support vector machine classifiers. *Journal of Machine Learning Research* **9**, 521–540.

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

Kalbfleisch, J. and Prentice, R. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, Hoboken, NJ.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33**, 82–95.

Kosorok, M. R. (2007). *Introduction to Empirical Processes and Semiparametric Inference.* Springer.

Leng, C. and Tong, X. (2013). A quantile regression estimator for censored data. *Bernoulli* **19**, 344–361.

Leng, C. and Zhang, H. H. (2006). Model selection in nonparametric hazard regression. *Nonparametric Statistics* **18**, 417–429.

Li, B., Artemiou, A., and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann. Statist.* **39**, 3182–3210.

Li, Y., Liu, Y., and Zhu, J. (2007). Quantile regression in reproducing kernel hilbert spaces. *J. Amer. Statist. Assoc.* **102**, 255–268.

Loprinzi, C. L., Laurie, J. A., Wieand, H. S., Krook, J. E., Novotny, P. J., Kugler, J. W., Bartel, J., Law, M., Bateman, M., and Klatt, N. E. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group. *Journal of Clinical Oncology* **12**, 601–607.

Lu, W. and Li, L. (2011). Sufficient dimension reduction for censored regressions. *Biometrics* **67**, 513–523.

Mallick, B. K., Ghosh, D., and Ghosh, M. (2005). Bayesian classification of tumours by using gene expression data. *J. Roy. Statist. Soc. Ser. B* **67**(2), 219–234.

Statistica Sinica: Preprint
doi:10.5705/ss.202014.0071

Peng, L. and Huang, Y. (2008). Survival analysis with quantile regression models. *J. Amer. Statist. Assoc.*

Portnoy, S. (2003). Censored regression quantiles. *J. Amer. Statist. Assoc.* **98**, 1001–1012.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics* **32**, 143–155.

Rosset, S. (2009). Bi-level path following for cross validated solution of kernel quantile regression. *Journal of Machine Learning Research* **10**, 2473–2503.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35**, 1012–1030.

Shin, S. J., Wu, Y., and Zhang, H. H. (2014). Two-dimensional solution surface for weighted support vector machines. *Journal of Computational and Graphical Statistics* **23**, 383–402.

Shows, J., Lu, W., and Zhang, H. (2010). Sparse estimation and inference for censored median regression. *Journal of Statistical Planning and Inference* **140**, 1903–1917.

Stute, W. and Wang, J. (1993). The strong law under random censorship. *Ann. Statist.* **21**, 1691–1607.

Takeuchi, I., Nomura, K., and Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation* **21**, 533–559.

Tibshriani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385–395.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer.

Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9**, 93–108.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *J. Amer. Statist. Assoc.*

Wang, S., Nan, B., Zhou, N., and Zhu, J. (2009). Hierarchically penalized Cox regression for censored data with grouped variables and its oracle property. *Biometrika* **96**, 307–332.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine* **11**, 1871–1879.

Ying, Z., Jung, S., and Wei, L. (1995). Survival anaysis with median regression models. *J. Amer. Statist. Assoc.* **90**, 178–184.

Zeng, D. and Lin, D. (2007). Efficient estimation in the accelerated failure time model. *J. Amer. Statist. Assoc.* **102**, 1387–1396.

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

Zhang, J. and Peng, Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine* **26**, 3157–3171.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research* **2**, 527–550.

Zhou, L. (2006). A simple censored median regression estimator. *Statistica Sinica* **16**, 1043–1058.

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

Department of Statistics, Korea University, Seoul, South Korea.

E-mail: (sjshin@korea.ac.kr)

Department of Mathematics, University of Arizona, Tucson, Arizona, U.S.A.

E-mail: (hzhang@math.arizona.edu)

Department of Statistics, North Carolina State University, Raleigh, North Carolina, U.S.A.

E-mail: (wu@stat.ncsu.edu)