Published in final edited form as:

Stat Comput. 2014 September 1; 24(5): 853–869. doi:10.1007/s11222-013-9406-4.

Shrinkage Estimation of Varying Covariate Effects Based On Quantile Regression

Limin Peng,

Department of Biostatistics and Bioinformatics, Emory University, 1518 Clifton Rd. NE, Atlanta, USA

Jinfeng Xu, and

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546

Nancy Kutner

Department of Rehabilitation Medicine, Emory University, Atlanta, USA 30322

Abstract

Varying covariate effects often manifest meaningful heterogeneity in covariate-response associations. In this paper, we adopt a quantile regression model that assumes linearity at a continuous range of quantile levels as a tool to explore such data dynamics. The consideration of potential non-constancy of covariate effects necessitates a new perspective for variable selection, which, under the assumed quantile regression model, is to retain variables that have effects on all quantiles of interest as well as those that influence only part of quantiles considered. Current work on l_1 -penalized quantile regression either does not concern varying covariate effects or may not produce consistent variable selection in the presence of covariates with partial effects, a practical scenario of interest. In this work, we propose a shrinkage approach by adopting a novel uniform adaptive LASSO penalty. The new approach enjoys easy implementation without requiring smoothing. Moreover, it can consistently identify the true model (uniformly across quantiles) and achieve the oracle estimation efficiency. We further extend the proposed shrinkage method to the case where responses are subject to random right censoring. Numerical studies confirm the theoretical results and support the utility of our proposals.

Keywords

Adaptive-LASSO; Censoring; Quantile regression; Shrinkage estimation; Variable selection; Varying covariate effects

1 Introduction

In regression analysis, there has been growing awareness that the association under study may be prone to considerable heterogeneity. Inhomogeneous associations, embodied by varying covariate effects, are often scientifically plausible and entail important practical

implications. There have been numerous reports in literature, for example, Kaslow et al (1987), Dickson et al (1989), Thorogood et al (1990), Verweij and Van Houwelingen (1995), Carey et al (1995), and Jensen et al (1997). As a more recent example, an analysis presented in Peng and Huang (2008) suggested that the severity of restless leg syndrome (RLS) symptoms may have prognostic power only in dialysis patients with high risk of mortality but not in long-term survivors. Such a varying association pattern sheds useful insight on dialysis prognosis but is precluded by traditional linear regression with the constraint of constant coefficients.

Quantile regression Koenker and Bassett (1978), by its modeling strategy, offers an excellent platform to accommodate and identify inhomogeneous associations. For example, a linear quantile regression model may take the form,

$$Q_Y(\tau|\mathbf{Z}) = \mathbf{Z}^T \boldsymbol{\beta}_0(\tau), \quad \tau \in (\tau_L, \tau_U), \quad (1)$$

where Y is a response variable, $\mathbf{Z} \equiv (1, \mathbf{Z}^T)^T$ is a $p \times 1$ vector of covariates, $\boldsymbol{\beta}_0(\tau) \equiv (\alpha_0(\tau), \boldsymbol{\beta}_0(\tau)^T)^T$ is a $p \times 1$ vector of unknown coefficients, $Q_Y(\tau|\mathbf{Z}) \equiv \inf \{y : \text{pr}(Y \leq y|\mathbf{Z}) = \tau\}$ denotes the τ -th conditional quantile of Y given $\mathbf{Z}$, and (τ_L, τ_U) specifies the quantile level range of interest with $0 < \tau_L < \tau_U < 1$. Here $\alpha_0(\tau)$ and $\boldsymbol{\beta}_0(\tau)$ represent intercept and covariate effects on the τ -th quantile of Y respectively, which are allowed to change over τ . As commented by Portnoy (2003), nonconstant coefficients in $\boldsymbol{\beta}_0(\tau)$ across τ reflect any sources of heterogeneity affecting Y but not captured by the covariates in $\mathbf{Z}$. Throughout this paper, we shall refer model (1) to as a “global” linear quantile regression model, reflecting the fact the linear quantile assumption is imposed over a τ -continuum, in contrast to a “local” model that only asserts quantile linearity for a single or a finite number of specified τ 's. It is worth emphasizing that model (1) is a strict extension of the traditional linear model, $Y = \mathbf{Z}^T \boldsymbol{\beta} + \varepsilon$, with or without distributional assumptions on the error term ε .

In practice, a large number of covariates are often collected at initial study stage. An appropriate variable selection procedure is thus crucial for building a model that excels in both interpretation and prediction accuracy. The consideration of varying covariate effects poses some conceptual subtlety for variable selection. Take model (1) for instance. Some covariates may exhibit a strong effect throughout the whole τ -range, namely a full effect, while some may impact part of but not all quantiles of interest and thus demonstrate a partial effect. In the dialysis example, RLS shows strong effects on lower quantiles of survival time but its effects seem to vanish for upper quantiles. It is scientifically important to retain such type of covariates in the final model because they can contribute a better understanding of the underlying response-covariate association and facilitate more profound predictions. This need calls for a new perspective for variable selection, which is to identify covariates with full effects as well as those with partial effects. Numerical studies presented in Section 3 indicate that existing variable selection methods that assume constant effects are incapable of fulfilling this need.

In this work, we adopt the varying coefficient quantile regression model (1) as the vehicle of tackling the variable selection problem that concerns potentially inhomogeneous associations. Penalization methods, as a special case of general shrinkage methods, have

been actively studied for “local” quantile regression models. For example, Koenker (2005) gave a summary of the use of LASSO penalty (Tibshirani, 1996) and other bridge penalties (Frank and Friedman, 1993) in the context of quantile regression. Li and Zhu (2005) developed the solution path of the L_1 penalty quantile regression. Zou and Yuan (2008) proposed a regression method, called composite quantile regression, which was shown to have some advantage in estimation efficiency over least squares methods. Wu and Liu (2009) presented a detailed study on the adaptation of the SCAD (Fan and Li, 2001) and adaptive-LASSO (Zou, 2006) penalties to quantile regression. Rocha et al (2009) extended the work of Knight and Fu (2000) to more general l_1 -penalized M -estimation settings and provided elegant theoretical results that can be applied to penalized quantile regression with fixed τ . However, all these approaches only concern covariate effects at a single or multiple specified quantile levels. Therefore, it is easy to perceive that they can miss important variables at a great chance when the quantile level targeted by the “local” quantile regression model is not in the τ -range where some covariates have influences. For example, a covariate that only influences the response from its 20th to 40th quantiles can be missed by the variable selection based on a median regression. An intuitive remedy for this problem may be to fit penalized “local” quantile regression at all possible τ 's of interest and remove variables that have shrunk coefficients at all τ 's considered. However, we shall show that this approach, which we call *pointwise penalization approach* hereafter, does not produce consistent variable selection.

Shrinkage estimation has been studied for other varying coefficient models. For example, Wang and Xia (2009) considered a linear model with coefficients formulated as functions of a univariate index variable. More recently, Kai et al (2011) extended the idea of composite quantile regression to estimate a similar type of varying coefficient model which contains both nonparametric coefficients and parametric coefficients. Both of these models permit extra data dynamics in contrast to a standard linear model by allowing coefficients to vary smoothly over the group stratified by the index variable. Of note, they require estimation based on kernel smoothing. As a result, the corresponding penalized estimators involve two levels of regularization; one for kernel bandwidth and one for penalization parameter. This may incur complexity in computation and theory.

The quantile regression model (1) considered in this work offers an alternative for accommodating varying associations. It offers a straightforward interpretation of the coefficient functions in $\beta_0(\tau)$, and has gained increased popularity recently, for example, in survival analysis (Portnoy, 2003; Neocleous et al, 2006; Peng and Huang, 2008; Portnoy and Lin, 2010; Huang, 2010). The fact that the estimation of model (1) does not require smoothing is expected to alleviate the computational complexity of the shrinkage estimation of $\beta_0(\cdot)$. The technical challenges now largely come from the need to deal with nonparametric coefficient functions and the non-differentiability of the loss functions tied to the estimation of $\beta_0(\tau)$.

Belloni and Chernozhukov (2011) studied a l_1 -penalized estimator for the “global” quantile regression model in high-dimensional sparse case, allowing for τ -varying coefficients. They established the elegant near-oracle convergence rate (uniformly in τ) and characterized conditions under which the selected model contains the true model as the submodel with

finite sample data. One critical condition required for their model selection properties is that the absolute value of any non-zero coefficient function must be bounded below by a positive separation bound uniformly in τ . While the separation bound can go to zero as n increases, it is not clear how their finite sample result on model selection can be extended to imply consistent model selection in the large sample sense without precluding the presence of covariates with partial effects (i.e. their coefficients are zero for some but not all τ 's), an important practical scenario like in the dialysis example discussed earlier.

In this paper, we develop a shrinkage estimation procedure based on the quantile regression model (1), which provides a useful variable selection tool that can flexibly accommodate inhomogeneous associations. Specifically, we follow the idea of adaptive LASSO (Zou, 2006) and propose a new penalization strategy to capture covariate effects' "global" departure from null. We show that such a strategy can lead to a shrinkage estimator of $\beta_0(\cdot)$ that enjoys the following oracle properties (Fan and Li, 2001): (i) the probability of correctly identifying all zero and nonzero coefficient functions approaches 1, which means consistent variable selection; (ii) the proposed estimator converges weakly to a Gaussian process in the same manner as if the set of irrelevant covariates was known, which entails efficient estimation. Note that, this is a stronger result than the oracle properties attained under a "local" quantile regression model as it asserts the uniform behavior of the coefficient functions. We do not require the full separation of non-zero coefficient functions from zero across τ and thus the new method is capable of detecting covariates that have effects on a partial range of quantile indices. Furthermore, we present an adaptation of the new shrinkage approach for randomly right censored data. This extension greatly enhances the applicability of this work in biomedical study settings, where censoring often arises. Findings from our numerical studies provide evidences for the utility of the proposed methods as well as the importance of accounting for the presence of varying covariate effects in variable selection. In this paper, we only concern the situation where p is fixed. The investigation of the case with $p \gg n$ will be reported in separate work.

2 The Proposed Methods

2.1 Uniform adaptive-LASSO estimation

Suppose the observed data include n independent and identically distributed replicates of $(Y, \mathbf{Z})$, denoted by $\{(Y_i, \mathbf{Z}_i)\}_{i=1}^n$. We use $u^{(j)}$ to denote the j th component of the vector $\mathbf{u}$, and adopt a shorthand notation $\tau \in \Delta$ for the quantile level range of interest, (τ_L, τ_U) . Without loss of generality, we assume that $\sup_{\tau \in \Delta} |\beta_0^{(j)}(\tau)| \neq 0$ for $2 \leq j \leq s$ and $\sup_{\tau \in \Delta} |\beta_0^{(j)}(\tau)| = 0$ for $s+1 \leq j \leq p$. Based on model (1), this means the first $s-1$ covariates in $\mathbf{Z}$ have some effects on Y while the rest do not.

It is worth noting that, here we adopt a global sparsity assumption,

$|\{2 \leq j \leq p: \sup_{\tau \in \Delta} |\beta_0^{(j)}(\tau)| > 0\}| = s-1 < \infty$. This represents a distinction of this work from penalized local quantile regression, which imposes sparsity only on $\beta_0(\tau)$ with a given τ (i.e. local sparsity). It is clear that the global sparsity would imply the local sparsity but not vice versa.

We begin with an unpenalized estimator of $\beta_0(\tau)$ (Koenker and Bassett, 1978), defined as $\tilde{\beta}_n(\tau) = \arg \min_{\beta} W_n(\beta; \tau)$, where

$$W_n(\beta; \tau) \equiv \sum_{i=1}^n \rho(Y_i - \mathbf{Z}_i^T \beta; \tau),$$

and $\rho(u; \tau) = u\{\tau - I(u < 0)\}$ is the so-called check function. Despite lack of a close form, $\tilde{\beta}_n(\tau)$ can be efficiently computed, even in the presence of a large number of covariates, by various algorithms, for example, Koenker and d'Orey (1987), Lustig et al (1994), and Madsen and Nielsen (1993).

To proceed the shrinkage estimation of $\beta_0(\tau)$, we adopt the adaptive LASSO idea from Zou (2006). In a general form, we present the penalized estimator as

$$\hat{\beta}_{n, \lambda_n}(\tau) = \arg \min_{\beta} W_{n, \lambda_n}(\beta; \tau),$$

where

$$W_{n, \lambda}(\beta; \tau) = W_n(\beta; \tau) + \lambda \sum_{j=2}^p w_{n, j}(\tau) \cdot |\beta^{(j)}|,$$

and $w_{n, j}(\tau)$ stands for the weight assigned to $|\beta^{(j)}|$. Note that when $w_{n, j}(\tau)$ takes the form of $|\tilde{\beta}_n^{(j)}(\tau)|^{-\gamma}$ with $\gamma > 0$, the above estimator becomes the adaptive-LASSO estimator studied by Wu and Liu (2009) for a “local” linear quantile regression model assumed for a fixed τ . The naive pointwise penalization approach discussed in Section 1 is to compute this type of estimator with $\gamma = 1$ separately for each $\tau \in \Delta$ to form a function estimator of $\beta_0(\cdot)$. As shown by our simulation studies, this approach may yield problematic variable selection results.

The proposed shrinkage estimator is $\hat{\beta}_{n, \lambda_n}^{US}(\tau)$ with $w_{n, j}(\tau) = w_{n, j}^{US}$, where

$$w_{n, j}^{US} = 1 / (\sup_{\tau \in \Delta} |\tilde{\beta}_n^{(j)}(\tau)|).$$

The key idea is about the new penalization strategy, by which we assign a high penalty to covariates, effects of which consistently exhibit small departure from zero for all τ , but not to those that have small effects at some τ 's but have large effects on the other τ 's. Using the uniform adaptive LASSO penalty $w_{n, j}^{US}$ in $W_{n, \lambda}(\beta; \tau)$ is expected to effectively differentiate zero coefficient function of τ versus those which are nonzero for some $\tau \in \Delta$. As elucidated in Section 2.2, uniform shrinkage behaviors across τ can be achieved by the proposed uniform adaptive LASSO estimator, $\hat{\beta}_{n, \lambda_n}^{US}(\tau)$.

To compute the proposed penalized estimator, $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$, one can adopt the existing routines for quantile regression, which have been well framed under linear programming (Koenker, 2005; Li and Zhu, 2005; Wu and Liu, 2009). More specifically, one may use the `rq()` function in R package *quantreg* (Koenker, 2011) to obtain $\tilde{\beta}_n(\cdot)$, the unpenalized estimator of $\beta_0(\tau)$ for $\tau \in (0, 1)$. Given that $\tilde{\beta}_n(\cdot)$ is piecewise constant, applying the `rq()` function would yield a finite number of function values of $\tilde{\beta}_n(\tau)$ along with the corresponding τ -jumping points, say $0 \equiv \tau_0 < \tau_1 < \dots < \tau_R < \tau_{R+1} = 1$. Based on these results, one can readily evaluate $\sup_{\tau \in \Delta} |\tilde{\beta}_n^{(j)}(\tau)|$ and thus $w_{n,j}^{US}$. Next, for any given $\tau \in (0, 1)$, one may compute $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$ by using the `rq.fit.lasso()` function in R package *quantreg* with the penalty parameter specified as $\lambda_n \cdot w_{n,j}^{US}$.

It is worth pointing out that in theory, $\hat{\beta}_{n,\lambda_n}^{US}(\cdot)$ obtained from a finite size dataset is also piecewise constant. However, it is not straightforward to locate the τ -points where $\hat{\beta}_{n,\lambda_n}^{US}(\cdot)$ jumps. To circumvent this complication, one can approximate $\hat{\beta}_{n,\lambda_n}^{US}(\cdot)$ by a cadlag step function which jumps at a sufficiently fine pre-specified grid in τ . For example, in our numerical studies, we chose the grid points as $(\tau_j + \tau_{j+1})/2, j = 0, \dots, R$. Our theoretical studies suggest that adopting a grid of size $O(n^{-1/2})$ would lead to only asymptotically negligible deviations from the exact $\hat{\beta}_{n,\lambda_n}^{US}(\cdot)$.

2.2 Asymptotic results

To study the asymptotic properties of $\hat{\beta}_{n,\lambda_n}^{US}(\cdot)$, we require the following regularity conditions:

- C1** (i) $\mathbf{Z}$ is bounded; that is, $\sup_i \|\mathbf{Z}_i\| < \infty$; (ii) $E(\mathbf{Z}\mathbf{Z}^T) > 0$; (iii) $\beta_0(\tau)$ is Lipschitz in τ for $\tau \in \tau_0$.
- C2** (i) $\inf_{\tau \in \tau_0} \int_{\mathcal{Z}} f_{\tau}(0|\mathbf{z}) > 0$; (ii) For some $A_0 > 0$ and $\delta_1 > 0$, $\sup_{\tau \in \tau_0} \int_{\mathcal{Z}} |f_{\tau}(x|\mathbf{z}) - f_{\tau}(0|\mathbf{z})| A_0 |x|$ for $|x| < \delta_1$. Here $f_{\tau}(x|\mathbf{z})$ denotes the density of $Y - \mathbf{Z}^T \beta_0(\tau)$ given $\mathbf{Z} = \mathbf{z}$.

Both C1 and C2 impose rather mild data assumptions. Here, we assume bounded $\mathbf{Z}$ to simplify asymptotic arguments. This condition can be relaxed to $E(\mathbf{Z}^2) < \infty$, similar to the theoretical studies on accelerated failure time (AFT) model (Ying, 1993). The positive definite covariate matrix assumption stated in C1(ii) is critical for the identifiability of $\beta_0(\tau)$, and is common in regression analysis. Condition C2 requires the smoothness and positiveness of $f_{\tau}(x|\mathbf{z})$ in a neighborhood of $x = 0$, which is often met in practice. Here and hereafter, we let $\ell_2 n$ be shorthand for $\log \log n$, and $\|\cdot\|$ denote the Euclidean norm.

Theorem 1—Assume that $(n^{1/2} \ell_2 n)^{-1} \lambda_n = O(1)$. Under regularity conditions C1-C2, we have $\sup_{\tau \in \Delta} \|\hat{\beta}_{n,\lambda_n}^{US}(\tau) - \beta_0(\tau)\| = O_p(n^{-1/2} \ell_2 n)$.

The uniform upper bound for, $\|\hat{\beta}_{n,\lambda_n}^{US}(\tau) - \beta_0(\tau)\|$, obtained in Theorem 1, implies the uniform consistency of $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$. This result also helps establish the oracle property of $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$ stated in the following theorem.

Theorem 2—Assume that $\lim_{n \rightarrow \infty} (\ell_2 n)^{-2} \lambda_n = \infty$ and $\lim_{n \rightarrow \infty} n^{-1/2} \lambda_n = 0$. Under regularity conditions C1-C2, we have

i. (selection consistency)

$$\lim_{n \rightarrow \infty} pr(\{2 \leq j \leq p: \sup_{\tau \in \Delta} |\hat{\beta}_{n,\lambda_n}^{US(j)}(\tau)| = 0\} = \{s+1, \dots, p\}) = 1;$$

ii. (weak convergence) $n^{1/2} \{\hat{\beta}_{n,\lambda_n}^{US(1:s)}(\tau) - \beta_0^{(1:s)}(\tau)\}$ converges weakly to a mean zero Gaussian process with a covariance matrix $\Sigma_{\text{oracle}}(\tau, \tau')$. Here, the superscript $(1:s)$ denotes the first s components of a vector, $\Sigma_{\text{oracle}}(\tau, \tau')$ is the same as the limit covariance matrix of $n^{1/2} \{\hat{\beta}_{\text{oracle}}(\tau) - \beta_0(\tau)\}$, where $\hat{\beta}_{\text{oracle}}(\tau)$ is the unpenalized estimator of $\beta_0^{(1:s)}(\tau)$ obtained as the minimizer of $W_n(\beta; \tau)$ with $\beta_0^{(j)}(\tau)$ ($s+1 \leq j \leq p$) known to be zero functions.

The oracle property established in Theorem 2 ensures the uniform shrinkage behaviors of $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$ over $\tau \in \Delta$, which is needed in the context that allows for varying covariate effects. To prove both theorems, we utilize empirical process theory to examine the asymptotic behavior of $\hat{\beta}_{n,\lambda_n}^{US}$ as a process of τ . Delicate probabilistic arguments are also used to handle the non-differentiability of $W_{n,\lambda_n}(\beta; \tau)$. Proofs are sketched in Appendices A and B, with a more detailed version presented in Supplementary Materials (available upon request).

Theorem 2 indicates the asymptotic order of λ_n which can yield consistent selection and efficient estimation. With a real dataset with finite sample size, we propose a BIC-like criterion to select a good penalization parameter, λ_n . Specifically, we define the BIC criterion as a sum of two terms:

$$BIC(\lambda) = \int_{\Delta} \log \hat{\sigma}_{\lambda}(\tau) d\tau + \frac{\log n}{n} |S_{\lambda}|,$$

where $S_{\lambda} = \{2 \leq j \leq p: \sup_{\tau \in \Delta} |\hat{\beta}_{n,\lambda}^{US(j)}(\tau)| \neq 0\}$, $|S_{\lambda}|$ denotes the size of S_{λ} , and

$$\hat{\sigma}_{\lambda}(\tau) = n^{-1} \sum_{i=1}^n \rho(Y_i - \mathbf{Z}_i^T \hat{\beta}_{n,\lambda}^{US}(\tau); \tau).$$

Note that the first term in $BIC(\lambda)$ quantifies the goodness of fit and the second term measures the model complexity. The use of $\log \hat{\sigma}_{\lambda}$ in BIC was motivated Koenker (2005)'s model selection criteria for quantile regression (see page 135). A similar form of BIC was also used by other authors, for example, Wang and Leng (2007). We select the penalization parameter as $\hat{\lambda}_n = \arg \min_{\lambda} BIC(\lambda)$. Simulations reported in Section 3.1 demonstrate satisfactory performance of $\hat{\lambda}_n$. We have also tested other methods such as AIC and

generalized cross validation (GCV) for selecting λ_n . The simulation results presented in Supplementary Materials (available upon request) show that, compared to BIC based procedure, the AIC and GCV based procedures have similar performance in terms of the model error but appear to select the true model at substantially lower rates. It would be important to further investigate the justification for our proposal of selecting the penalty parameter based on BIC.

2.3 Inference

The oracle property stated in Theorem 2 facilitates the statistical inference for $\beta_0(\tau)$ based on $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$. In this subsection, we outline the procedures for variance estimation and construction of confidence intervals.

Let $C_n = \{1\} \cup S_{\lambda_n}$ and $q_n = |C_n|$. For a vector $\mathbf{u}$ and an index set Φ , we let $\mathbf{u}^{(\Phi)}$ denote a subvector of $\mathbf{u}$ with Φ indicating the components of $\mathbf{u}$ included in $\mathbf{u}^{(\Phi)}$. In the rest of this

subsection, we use $\hat{\beta}_n(\tau)$ as shorthand notation for $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$. Define

$$\begin{aligned} \mathbf{r}_n(\tau) &= \left(w_{n,1}(\tau) \text{sgn}\{\hat{\beta}^{(1)}(\tau)\}, \dots, w_{n,p}(\tau) \text{sgn}\{\hat{\beta}^{(p)}(\tau)\} \right)^T, \\ \hat{\ell}_j(\tau) &= -\mathbf{Z}_j^{(C_n)} [\tau - I\{Y_j - \mathbf{Z}_j^T \hat{\beta}(\tau) < 0\} - \lambda_n \mathbf{r}_n^{(C_n)}(\tau)], \text{ and} \\ \Omega_n(\tau, \tau) &= n^{-1} \sum_{j=1}^n \hat{\ell}_j(\tau)^{\otimes 2}. \end{aligned}$$

We propose the following steps to obtain variance estimates for regression quantiles that do not shrinkage for all $\tau \in \cdot$:

1. Find a symmetric and nonsingular $q_n \times q_n$ matrix $\mathbf{E}_n(\tau) \equiv \{\mathbf{e}_{n,1}(\tau), \dots, \mathbf{e}_{n,q_n}(\tau)\}$ such that $\Omega_n(\tau, \tau) = \{\mathbf{E}_n(\tau)\}^2$.
2. Calculate the $q_n \times q_n$ matrix,

$$\mathbf{D}_n(\tau) = \left(\mathbf{b}_{e,1}(\tau) - \hat{\beta}^{(C_n)}(\tau), \dots, \mathbf{b}_{e,q_n}(\tau) - \hat{\beta}^{(C_n)}(\tau) \right),$$

where

$$\begin{aligned} \mathbf{b}_{e,k}(\tau) &= \arg \min_{\mathbf{b} \in R^{q_n}} \left\{ \sum_{i=1}^n \rho \left(Y_i - \mathbf{Z}_i^{(C_n)T} \mathbf{b}; \tau \right) + \right. \\ &\quad \left. \lambda_n \sum_{j=2}^{q_n} w_n, C_n(j) \cdot |b^{(j)}| - n^{1/2} \mathbf{b}^T \mathbf{e}_{n,k}(\tau) \right\} \end{aligned}$$

with $C_n(j)$ standing for the j th element in set C_n .

3. Estimate the variance matrix of $n^{1/2}\{\hat{\beta}_n^{(C_n)}(\tau) - \beta_0^{(C_n)}(\tau)\}$ by $n\{\mathbf{D}_n(\tau)\}^{\otimes 2}$, where $\mathbf{u}^{\otimes 2} = \mathbf{u}\mathbf{u}^T$.

The technique used for variance estimation is an adaptation of Huang (2002) and Peng and Fine (2009)'s methods. Based on the proof of Theorem 2 (ii), specifically equation (17) in

the Appendix B, and the uniform consistency of $\hat{\beta}(\tau)$ from Theorem 1, we can readily apply the lines of Peng and Fine (2009) to show that $n\{\mathbf{D}_n(\tau)\}^{\otimes 2}$ is a consistent variance estimate for $\hat{\beta}^{(1:s)}(\tau)$ when $\mathbf{C}_n = \{1, \dots, s\}$. This, coupled with Theorem 2 (i), implies that for any $\varepsilon > 0$,

$$\begin{aligned} & \lim_{n \rightarrow \infty} pr(\mathbf{C}_n = \{1, \dots, s\} \text{ and } \\ & \|n\{\mathbf{D}_n(\tau)\}^{\otimes 2} - \sum_{oracle}(\tau, \tau)\| \leq \varepsilon) \\ & \geq \lim_{n \rightarrow \infty} \{1 - pr(\mathbf{C}_n \neq \{1, \dots, s\}) - pr(\mathbf{C}_n = \{1, \dots, s\} \\ & \text{and } \|n\{\mathbf{D}_n(\tau)\}^{\otimes 2} - \sum_{oracle}(\tau, \tau)\| > \varepsilon)\} = 1. \end{aligned}$$

The above arguments provide a justification for the proposed variance estimation. Likewise, we propose to construct a $100(1 - \alpha)\%$ confidence interval for $\beta_0^{(j)}(\tau)$ as $(\hat{\beta}^{(j)}(\tau) - z_{\alpha/2} \hat{\sigma}_j(\tau), \hat{\beta}^{(j)}(\tau) + z_{\alpha/2} \hat{\sigma}_j(\tau))$ if $j \in \mathbf{C}_n$ and $\{0\}$ otherwise, where $\hat{\sigma}_j^2$ is the j th diagonal element of $\{\mathbf{D}_n(\tau)\}^{\otimes 2}$ and z_α is the $100(1 - \alpha)$ th percentile of $N(0, 1)$. Denote such a defined confidence interval as I_n . For $1 \leq j \leq s$, given that $pr(j \notin \mathbf{C}_n) = 0$ and $n^{1/2}\{\hat{\beta}^{(j)}(\tau) - \beta_0^{(j)}(\tau)\}$ is asymptotically normal with $n\hat{\sigma}_j(\tau)^2$ being a consistent variance estimate, we have

$$\begin{aligned} & \lim_{n \rightarrow \infty} pr(\beta_0^{(j)}(\tau) \in I_n) \\ & = \lim_{n \rightarrow \infty} pr(|\beta_0^{(j)}(\tau) - \hat{\beta}^{(j)}(\tau)| < z_{\alpha/2} \hat{\sigma}_j(\tau) \text{ and } j \in \mathbf{C}_n) \\ & = \lim_{n \rightarrow \infty} pr(|\beta_0^{(j)}(\tau) - \hat{\beta}^{(j)}(\tau)| < z_{\alpha/2} \hat{\sigma}_j(\tau) | j \in \mathbf{C}_n) = 1 - \alpha. \end{aligned}$$

This implies that the asymptotic confidence level of I_n is $(1 - \alpha)$.

2.4 An extension to randomly censored data

Censoring often occurs in biomedical study settings. In this subsection, we propose an extension of the shrinkage method proposed in Section 2.1 to accommodate scenarios with random censoring. Specifically, let C denote the censoring variable to the response Y . We assume C is independent of $(Y, \mathbf{Z})$. Note, this independent assumption is stronger than the standard random censoring assumption (i.e. C is independent of Y given $\mathbf{Z}$). Define $X = Y \wedge C$ and $\delta = I(Y = C)$, where $\wedge$ is the minimum operator. Observables include n independent and identically distributed replicates of $(X, \delta, \mathbf{Z})$, denoted by $\{(X_i, \delta_i, \mathbf{Z}_i)\}_{i=1}^n$.

For unpenalized estimation of model (1) in the presence of random censoring by C , we consider an adaptation of Ma and Yin (2010)'s estimator, which can be obtained as

$$\tilde{\beta}_{C,n}(\tau) = \arg \min_{\beta} V_n(\beta; \tau),$$

where

$$V_n(\beta; \tau) = \sum_{i=1}^n \frac{\delta_i}{\hat{G}(X_i)} \rho(X_i - \mathbf{Z}_i^T \beta; \tau).$$

Here $\hat{G}(\cdot)$ is the Kaplan-Meier estimator of $G(x) = \text{pr}(C \leq x)$. One can use the available software, for example, the R package *quantreg*, to compute $\hat{\beta}_{C,n}(\tau)$ as weighted regression quantiles.

Using the same reasoning as that in Section 2.1, we propose the shrinkage estimator of $\beta_0(\tau)$ as $\hat{\beta}_{C,n,\lambda_{C,n}}(\tau) = \arg \min_{\beta} V_{n,\lambda_{C,n}}(\beta; \tau)$, where

$$V_{n,\lambda}(\beta; \tau) = V_n(\beta; \tau) + \lambda \sum_{j=2}^p w_{C,n,j}(\tau) |\beta^{(j)}|$$

with $w_{C,n,j}(\tau) = (\sup_{\tau \in \Delta} |\hat{\beta}_{C,n}^{(j)}(\tau)|)^{-1}$.

By the analogy between $V_n(\beta; \tau)$ and a standard weighted quantile regression problem, the implementation of $\hat{\beta}_{C,n,\lambda_{C,n}}(\tau)$ is expected to be very similar to our proposal for the uncensored shrinkage estimator $\beta_{n,\lambda_n}^{US}(\tau)$. So is the asymptotic studies for $\hat{\beta}_{C,n,\lambda_{C,n}}(\tau)$. With some additional regularity condition on censoring, for example,

C3 There exists $\nu > 0$ such that $\text{pr}(C = \nu) > 0$ and $\text{pr}(C > \nu) = 0$;

we would have the uniform consistency of $\hat{G}(t)$ for $t < \nu$, and thus can obtain the oracle property of $\hat{\beta}_{C,n,\lambda_{C,n}}(\tau)$ as in the uncensored case. Note that condition C3 is a rather mild assumption and is often satisfied in practice in the presence of administrative censoring. As pointed out by Peng and Fine (2009), artificial truncation may be imposed to ensure this condition always met. The detailed results are stated in the theorem below with the proof omitted.

Theorem 3—Assume that $\lim_{n \rightarrow \infty} (\ell_2 n)^{-2} \lambda_{C,n} = \infty$, and $\lim_{n \rightarrow \infty} n^{-1/2} \lambda_{C,n} = 0$. Under regularity conditions C1-C2 and some mild assumption on censoring as those in Ma and Yin (2010), we have

i. (selection consistency)

$$\lim_{n \rightarrow \infty} \text{pr}(\{2 \leq j \leq p: \sup_{\tau \in \Delta} |\hat{\beta}_{C,n,\lambda_{C,n}}^{(j)}(\tau)| = 0\} = \{s+1, \dots, p\}) = 1,$$

ii.

(weak convergence) $n^{1/2} \{\hat{\beta}_{C,n,\lambda_{C,n}}^{(1:s)}(\tau) - \beta_0^{(1:s)}(\tau)\}$ converges weakly to a mean zero Gaussian process with a covariance matrix $\Sigma_{C,\text{oracle}}(\tau, \tau')$. Here, $\Sigma_{C,\text{oracle}}(\tau, \tau')$ is the same as the limit covariance matrix of $n^{1/2} \{\hat{\beta}_{C,\text{oracle}}(\tau) - \beta_0^{(1:s)}(\tau)\}$ where $\hat{\beta}_{C,\text{oracle}}(\tau)$ is the unpenalized estimator of $\beta_0^{(1:s)}(\tau)$ obtained as the minimizer of $V_n(\beta; \tau)$ with $\beta_0^{(j)}(\tau)$ ($s+1 \leq j \leq p$) known to be zero functions.

Similarly, we propose to select λ_n based on a BIC-like criterion:

$$\widetilde{BIC}(\lambda) = \int_{\Delta} \log \hat{\sigma}_{C,\lambda}(\tau) d\tau + \frac{\log n}{n} |S_{\lambda}|,$$

where

$$\hat{\sigma}_{C,\lambda}(\tau) = \frac{\sum_{i=1}^n \delta_i \rho(X_i - \mathbf{Z}_i^T \hat{\beta}_{C,n,\lambda}(\tau); \tau) / \hat{G}(X_i)}{\sum_{i=1}^n \delta_i / \hat{G}(X_i)}$$

Note that $\hat{\sigma}_{C,\lambda}$, like $\hat{\sigma}_\lambda$, serves as an approximation to $n^{-1} \sum_{i=1}^n \rho(Y_i - \mathbf{Z}_i^T \beta_0(\tau); \tau)$ with censoring to Y_i properly handled. For a given dataset, we choose $\lambda_{C,n}$ as

$\hat{\lambda}_{C,n} = \arg \min_{\lambda} \widehat{BIC}(\lambda)$. Inference, such as variance estimation and confidence intervals, may mimic the procedures proposed for the uncensored case.

3 Numerical Studies

3.1 Monte-Carlo simulations

Extensive simulations were conducted to evaluate the finite-sample performance of the proposed shrinkage estimators in both uncensored and censored cases.

We first considered situations where there is no censoring to the response Y . We generated Y based on model (1) with eight covariates, $Z^{(1)}, \dots, Z^{(8)}$. Only $Z^{(1)}$, $Z^{(2)}$ and $Z^{(5)}$ have non-zero effects, and the coefficients for the other covariates are zero functions of τ . We considered three set-ups, denoted by (I), (II), and (III). For each of these set-ups, we plot the non-zero coefficients for $Z^{(1)}$, $Z^{(2)}$, and $Z^{(5)}$ in Figure 1 in bold lines, dashed lines and dotted lines respectively. The intercept coefficient function in set-up (I) is $\sqrt{2}$ times the quantile function of the standard normal distribution, and that in set-up (II) or (III) is set as a zero function in τ . It is easy to see from the specification of $\beta_0(\tau)$ that a standard linear model with normal random errors holds in set-up (I) but not in set-ups (II) and (III). In set-up (I), $Z^{(1)}, \dots, Z^{(8)}$ were generated as standard normal variates with the correlation between $Z^{(i)}$ and $Z^{(j)}$ equal to $0.5^{|i-j|}$ ($1 \leq i, j \leq 8$), and were truncated to be between -3 and 3 . In set-ups (II) and (III), $Z^{(i)} = \Phi(\xi_i)$ ($i = 1, \dots, 8$), where $\xi_1, \dots, \xi_8$ are standard normal variates with the correlation between ξ_i and ξ_j equal to $0.5^{|i-j|}$ ($1 \leq i, j \leq 8$). Additional simulations were also conducted for cases with a larger number of covariates (i.e. $p = 40$) and are reported in Supplementary Materials (available upon request).

We assessed the proposed shrinkage estimation, denoted by US, and compare it with several other methods: (i) the naive pointwise penalization method described in Section 1 and 2.1, denoted by PS; (ii) the adaptive-LASSO estimator for “local” quantile regression at a single predetermined quantile level τ , denoted by SS(τ) with $\tau = 0.25, 0.50$ or 0.75 ; (iii) adaptive-LASSO estimator for the least-square estimation of linear regression of Y on $\mathbf{Z}$, denoted by LS. In the LS and SS(τ) approaches, it is implicitly assumed that the importance of a covariate can be adequately captured by its effect on the conditional mean or a conditional quantile at a known τ . We set $\tau_L = 0.1$ and $\tau_U = 0.9$. For each approach under comparison, the tuning parameter λ_n is selected as the minimizer of the corresponding BIC criterion among the $25(n/100)$ equally-spaced grid points between 0 and $12(n/100)$.

In Table 1, we present results based on 400 simulation replications for the proposed method and all the other methods described above. Specifically, we report mean number of correctly identified zero effects (NC), mean number of incorrectly identified zero effects (NIC), percentage of under-fitted models (PUF), percentage of correctly fitted models (PCF), and percentage of over-fitted models (POF). As a further clarification, NC counts the number of unselected covariates, which have zero coefficient functions (i.e. zero for all τ 's). In set-ups (I)–(III), we hope NC to be close to 5 as much as possible. We also evaluated the estimation accuracy of different estimators by examining their relative estimation errors (REE) with respect to the unpenalized estimator $\tilde{\beta}(\tau)$ or the oracle estimator $\hat{\beta}_{\text{oracle}}(\tau)$. The REE of $\beta(\cdot)$ is defined as

$$REE(\beta) = 100 \times \frac{\int_{\Delta} \sum_{i=1}^n \sum_{j=1}^p |\beta^{(j)}(\tau) - \beta_0^{(j)}(\tau)| d\tau}{\int_{\Delta} \sum_{i=1}^n \sum_{j=1}^p |\bar{\beta}^{(j)}(\tau) - \beta_0^{(j)}(\tau)| d\tau},$$

where $\tilde{\beta}(\tau)$ is either $\hat{\beta}(\tau)$ or $\hat{\beta}_{\text{oracle}}(\tau)$. The estimate from the SS(τ) or LS approach was taken as the constant value of estimated $\hat{\beta}_0(\tau)$ over τ . The median of REE with respect to the unpenalized estimator ($MREE_F$) and the median of REE with respect to the oracle estimator ($MREE_O$) are presented in Table 1.

From Table 1, we see that in set-up (I) where a linear model with normal random error holds, US, SS(0.25), SS(0.50), SS(0.75) and LS have quite similar performance in terms of model selection. The PCF's are all above 90%. Among them, LS shows the best estimation accuracy. Nevertheless, this advantage over the other approaches is only moderate. Note that the oracle estimator adopted here is defined based on the quantile regression model (1). Hence it is reasonable to observe $MREE_O$ below 0.8 for LS. Our finding on the superior estimation accuracy of LS estimator is consistent with the theory in Zou and Yuan (2008). It is important to note that the PS approach, though demonstrating good estimation accuracy as the other methods, shows a rather high tendency to produce an overfitted model, with POF=53.5% and 40.2% for $n = 200$ and 400 respectively. This strongly suggests that the intuitive PS approach may fail to generate consistent model selection.

In set-ups (II) and (III) where a standard linear model does not hold, LS has poor performance in identifying covariates that have effects on the response with PCF less than 65% when $n = 400$. The inconsistent shrinkage behavior also leads to inflated estimation errors. For example, $MREE_F$ for the LS estimator even exceeds 1 when $n = 400$. As in set-up (I), PS remains to have the overfitting problem in these configurations. When covariates with partial effects are present, the penalized “local” quantile regression method SS(τ) can be troublesome. For example, in set-up (II), $Z^{(1)}$ impacts Y at all quantile levels except for $\tau = 0.5$. When one simply applies the penalized median regression, $Z^{(1)}$ is identified as an unimportant covariate and its effects on other quantiles of Y are overlooked. This, reflected by NIC=1 and a low PCF of 2%, indicates a need to conduct a more thorough assessment of covariates' impact across various quantiles. This message is more clear in set-up (III) where $Z^{(1)}$ and $Z^{(2)}$ have partial effects on two disjoint quantile ranges. Finally, we note that in either set-up (II) or (III), US demonstrates satisfactory performance; PCF is above 99% when $n = 400$ and $MREE_O$ decreases to approach 1 as n increases.

We also considered situations where Y is subject to random censoring. In the new set-ups labeled as (IV), (V), (VI), Y and Z were generated in the same way as in set-ups (I), (II), (III) respectively, except that the intercept coefficient function in set-up (IV) is set as the quantile function of the extreme value distribution. As a result, the relationship between Y and Z in set-up (IV) satisfies a linear model with extreme value error distribution and also a proportional hazards model. The censoring variable, C , was generated as $Unif(-3, 14)$, $Unif(-2, 19)$, and $Unif(-2, 16)$ respectively in set-ups (IV), (V), and (VI) to produce 20% censoring. We set $\tau_L = 0.1$ and $\tau_U = 0.8$.

The simulation results for the censored cases are presented in Table 2 in the same manner as that for Table 1. The proposed method for censored data is now denoted by US_C , while the methods that we compared US_C with include PS_C , $SS_C(0.25)$, $SS_C(0.50)$, and $SS_C(0.75)$, which are modified PS , $SS(0.25)$, $SS(0.5)$, and $SS(0.75)$ that adjust for censoring. As a counter-part of LS in the censoring case, we considered adaptive LASSO for accelerated failure time (AFT) model, which was implemented by the method of least squares approximation (Wang and Leng, 2007) based on the log-rank estimator (Tsiatis, 1990; Wei and Gail, 1983). This approach is denoted by AAFT. In addition, we considered adaptive-LASSO estimator for the Cox proportional hazards model (Zhang and Lu, 2007), denoted by PH. From Table 2, we see that the empirical results support the oracle properties of $\hat{\beta}_{C,n,\lambda_{C,n}}(\cdot)$. Though PCF's may be lower than those observed in the uncensored case, which is not surprising given the information loss due to censoring, there is still a clear trend of PCF approaching 100 as n increases. The pointwise penalization approach remains to demonstrate a tendency of producing an overfitted model. Findings on the penalized local quantile regression approach are the same as those from Table 1. AAFT and PH perform well in set-up (IV) where data satisfy a linear model and a proportional hazards model. However, when the constant effect assumption is violated as in set-ups (V) and (VI), these methods can either overfit or underfit the model. By comparing PCF, $MREE_F$ and $MREE_O$ between US_C and AAFT, we see that assuming covariate effects constant when they are indeed varying can lead to not only misleading variable selection results but also substantially inflated estimation errors which entail very biased prediction of quantile profile of Y under model (1).

3.2 Applications to real datasets

Example 1—We considered the Air Pollution to Mortality Data (McDonald and Schwing, 1973), which are publicly available in the R package *SMPracticals*. This dataset contains information about a measure of total mortality (y) and 15 independent variables on 60 US metropolitan areas in 1959-1961. The 15 variables are precipitation (x_1), January temperature (x_2), July temperature (x_3), the percentage of age 60 or older (x_4), population per household (x_5), education (x_6), the percentage of soundhouse (x_7), population per square mile (x_8), the percentage of non-white (x_9), the percentage of white collar (x_{10}), the percentage of family income under 3000 dollars (x_{11}), HC potential (x_{12}), NOx potential (x_{13}), SO2 potential (x_{14}), and relative humidity (x_{15}). Of scientific interest is to predict y given $x_1, \dots, x_{15}$.

We applied the methods, including US, PS, SS(0.25), SS(0.5), SS(0.75) and LS, to the air pollution dataset. To evaluate different methods, we computed prediction errors (PE) as follows. First, we splitted the data randomly into a training set comprising 30 observations and a test set containing the rest 30 observations. We applied the method under evaluation to the training dataset and obtained a shrinkage estimator of $\beta_0(\tau)$, denoted by $\hat{\beta}^{train}(\tau)$. Then the prediction error was calculated as

$$PE = \frac{\sum_{i=1}^n I(i \text{ in Test Set}) \int \Delta \rho(Y_i - \mathbf{Z}_i^T \hat{\beta}^{train}(\tau); \tau) d\tau}{\sum_{i=1}^n I(i \text{ in Test Set})}.$$

In our analysis, all variables were standardized to have mean zero and unit variance. We set $\tau = [0.1, 0.9]$. In Table 3, we present the selected variables and average PE based on 10,000 replications of random splitting of training and test sets (APE).

From Table 3, we first note that PS may retain too much variables; a total of 13 variables are included in the final model. This result is consistent with the PS's overfitting behavior observed in the simulation studies. The proposed method, US, may yield the most parsimonious model with only 4 variables, which is associated with the smallest APE. To better understand the discrepancy in selected variables among different approaches, we plot in Figure 2 the coefficient functions from fitting model (1) with all covariates along with the corresponding pointwise 95% confidence intervals. We also depict the least square coefficient estimates and the corresponding 95% confidence intervals by horizontal lines. It is observed from the regression quantile curves that the effects of x_{10} and x_{14} demonstrate apparent non-constant patterns with different signs at lower quantiles and upper quantiles, while x_1 and x_9 exhibit fairly consistent significant effects in the whole τ -range of interest. This finding may help explain why LS selected x_1 and x_9 but not x_{10} and x_{14} . The variable selection results from penalized local quantile regression are closer to that from US compared to PS and LS. The APE's from SS(0.25), S(0.50), S(0.75) are all slightly larger than the APE from US.

Example 2—We also illustrated the proposed method for randomly censored data to the dialysis example discussed in §1. A more detailed study background can be found in Kutner et al (2002) and Peng and Huang (2008). The dataset contains the information on the vital status of 191 incident dialysis patients up to December of 2005, and 14 covariates including severity of restless leg syndrome (x_1), patient's age at study admission (x_2), patient's BMI at study admission (x_3), baseline dialysis modality (x_4), baseline hematocrit (x_5), baseline mental health scale score (x_6), baseline ferritin concentration (x_7), baseline serum albumin (x_8), primary diagnosis of diabetes (x_9), indicator for cardiovascular comorbidity (x_{10}), indicator for high education (x_{11}), gender (x_{12}), indicator for fish eater (x_{13}), and race (x_{14}). One goal of this study is to investigate the risk factors for dialysis mortality, assessed by time from study admission to death, subject to censoring due to renal transplantation or end of the study. The censoring rate in this dataset is 35%. In our analysis, the response variable y is taken as the log transformed survival time and τ is set as $[0.1, 0.9]$. We excluded ten subjects with missing covariates. The calculation of prediction error is slightly modified for

the censored case. Specifically, the training set consists of 91 randomly selected observations, and

$$PE = \frac{\sum_{i=1}^n I(i \text{ in Test Set}) \int_{\Delta} \delta_i \rho(X_i - \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}^{train}(\tau); \tau) d\tau / \hat{G}(X_i)}{\sum_{i=1}^n \delta_i I(i \text{ in Test Set}) / \hat{G}(X_i)}$$

Here the Kaplan-Meier estimator, $\hat{G}(\cdot)$, is computed based on the test dataset.

In Table 4, we summarize the results from the methods, US, PS, SS(0.25), SS(0.50), SS(0.75), AAFT, and PH. In Figure 3, we plot the regression quantiles and the estimates of AFT model coefficients when all covariates are included. It is seen from Table 4 that, as in example 1, PS tends to select more variables than the other methods, while US renders the smallest APE. With a particular interest in the impact of severity of restless leg syndrome on dialysis survival, we note that x_1 was selected by US and SS(0.25) but not SS(0.50), SS(0.75), AAFT and PH. This result may be explained by the regression quantile curve for x_1 in Figure 3. That is, the effect of x_1 on the τ -th quantile of y may have a changing trend with τ ; the coefficient magnitude decreases and approaches 0 around $\tau = 0.6$. As a result, the penalization method based on local quantile regression at a large τ would fail to identify restless leg syndrome as an important risk factor for predicting the mortality of dialysis patients. Fitting models that only allow for constant effects, such as the Cox proportional hazards model and the AFT model, also appears to attenuate the influence of restless leg syndrome on dialysis survival, thereby excluding x_1 in the final model. This example, again, demonstrates the importance of appropriately accounting for varying covariate effects in identifying important variables and building parsimonious predictive models.

4 Remarks

Meaningful heterogeneous associations between covariates and response are naturally embedded in many real datasets but are precluded by common regression models with constant coefficients. Some of empirical work presented in the paper shows that making inference based on a constant coefficient model in the presence of varying covariate effects can lead to misleading variable selection results and also deteriorated prediction accuracy.

In this work, we adopt a global quantile regression model as the platform to address potential inhomogeneity in the association of interest. This choice of model facilitates the development of the proposed variable selection method by avoiding smoothing and enabling stable and efficient computation.

Our empirical work also strongly suggests that fitting penalized “local” quantile regression pointwisely would often lead to an overfitted model. The intuition for this phenomenon shares similar flavor with the multiple comparison problem. To be more clear, consider the problem of testing the null hypothesis, $H_0: \beta_0^{(j)}(\tau) = 0$ for all $\tau \in \cdot$. The estimation of $\beta_0^{(j)}(\cdot)$ based on the pointwise penalization approach is analogous to the testing procedure that tests $H_0(\tau): \beta_0^{(j)}(\tau) = 0$ for each $\tau \in \cdot$ and rejects H_0 when $H_0(\tau)$ is rejected at any $\tau \in \cdot$. In this view, the chance that the pointwise estimator of a true zero coefficient function fails to

shrink uniformly in τ would be analogous to the type-I error in the above naive test procedure, which can be significantly inflated particularly when τ contains an infinite number of τ .

For randomly censored data, the proposed shrinkage method that assumes the independence between the censoring variable C and $\mathbf{Z}$ can be generalized to allow for covariate-dependent censoring. For example, one may replace the $\hat{G}(X_i)$ in $V_n(\boldsymbol{\beta}; \tau)$ by some adhoc estimator of $G(t|\mathbf{Z}) \equiv \text{pr}(C > t|\mathbf{Z})$ obtained via reasonable parametric or semiparametric modeling of $G(\cdot|\mathbf{Z})$.

Acknowledgments

The authors are grateful to the editor, associate editor, and the two referees for many helpful comments. This research has been supported by the National Heart, Lung, And Blood Institute of the National Institute of Health under Award Number R01HL 113548 (the first author). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Health.

Appendix A: Proof of Theorem 1

Define $B_{n,\tau}(C) = \{\boldsymbol{\beta}: \boldsymbol{\beta} = \boldsymbol{\beta}_0(\tau) + (n^{-1/2} \ell_2 n) \mathbf{u}, \|\mathbf{u}\| \leq C\}$, and let $\partial B_{n,\tau}(C)$ denote the boundary set of $B_{n,\tau}(C)$. Since $W_{n,\lambda_n}(\boldsymbol{\beta}; \tau)$ is convex in $\boldsymbol{\beta}$ for all $\tau \in \mathcal{T}$, it is sufficient to show that for any $\varepsilon > 0$, there exists $C_0 > 0$ and $N_0 > 0$ such that for $n \geq N_0$,

$$\text{pr}(\inf_{\tau \in \mathcal{T}} \inf_{\boldsymbol{\beta} \in \partial B_{n,\tau}(C_0)} W_{n,\lambda_n}(\boldsymbol{\beta}; \tau) - W_{n,\lambda_n}(\boldsymbol{\beta}_0(\tau); \tau) > 0) \geq 1 - \varepsilon. \quad (2)$$

To show equation (2), first note that

$$\begin{aligned} & n^{-1} W_{n,\lambda_n}(\boldsymbol{\beta}; \tau) - n^{-1} W_{n,\lambda_n}(\boldsymbol{\beta}_0(\tau); \tau) \\ &= n^{-1} \sum_{i=1}^n [\rho(Y_i - \mathbf{Z}_i^T \boldsymbol{\beta}; \tau) - \rho(Y_i - \mathbf{Z}_i^T \boldsymbol{\beta}_0(\tau); \tau)] \\ & \quad + n^{-1} \lambda_n \sum_{j=2}^p \{|\beta^{(j)}| w_{n,j}(\tau) - |\beta_0^{(j)}| w_{n,j}(\tau)\} \\ & \equiv I + II \end{aligned}$$

Write $e_i(\tau) = Y_i - \mathbf{Z}_i^T \boldsymbol{\beta}_0(\tau)$ and $\mathbf{u}(\tau; \boldsymbol{\beta}) = \boldsymbol{\beta} - \boldsymbol{\beta}_0(\tau)$, and let D and D_i respectively denote operators such that $D(U) = U - E(U)$ and $D_i(U) = U - E(U|\mathbf{Z}_i)$ for a random variable U . We can further decomposed the term I as $I = III + IV$, where

$$\begin{aligned}
III &= E[I\{0 < e_i(\tau) < \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \{-e_i(\tau) + \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\}] \\
&\quad + E[I\{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) < e_i(\tau) < 0\} \{e_i(\tau) - \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\}] \text{ and} \\
&\quad n \cdot (IV) \\
&= \sum_{i=1}^n D \left(I\{0 < e_i(\tau) < \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \frac{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) - e_i(\tau)}{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta)} \mathbf{Z}_i^T \right) \mathbf{u}(\tau; \beta) \\
&\quad + \sum_{i=1}^n D \left(I\{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) < e_i(\tau) < 0\} \frac{e_i(\tau) - \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)}{-\mathbf{Z}_i^T \mathbf{u}(\tau; \beta)} (-\mathbf{Z}_i^T) \right) \mathbf{u}(\tau; \beta) \\
&\quad + \sum_{i=1}^n I\{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) > 0\} D_i(I\{e_i(\tau) > \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\}) \{-\tau \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \\
&\quad + \sum_{i=1}^n I\{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) > 0\} D_i(I\{e_i(\tau) < 0\}) \{(1 - \tau) \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \\
&\quad + \sum_{i=1}^n I\{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) > 0\} D_i(I\{0 < e_i(\tau) < \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\}) \{-\tau \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \\
&\quad + \sum_{i=1}^n I\{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) < 0\} D_i(I\{e_i(\tau) < \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\}) \{(1 - \tau) \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \\
&\quad + \sum_{i=1}^n I\{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) < 0\} D_i(I\{e_i(\tau) > 0\}) \{-\tau \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \\
&\quad + \sum_{i=1}^n I\{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) < 0\} D_i(I\{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) < e_i(\tau) < 0\}) \{(1 - \tau) \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \\
&\quad \equiv \sum_{j=1}^8 n \cdot (IV_j)
\end{aligned}$$

For the term IV_1 , we can show that the function class,

$F = \left\{ I\{0 < e_i(\tau) < \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \frac{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) - e_i(\tau)}{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta)} \mathbf{Z}_i, \tau \in \Delta, \beta \in A(C_0) \right\}$, is a Donsker class (page 81 in Van der Vaart and Wellner (2000)), where $A(C_0) = \{\mathbf{b} \in R^p : \inf_{\tau \in \Delta} \|\mathbf{b} - \beta_0(\tau)\| < C_0\}$. Given the uniform boundedness of the functional class F and since $A(C_0)$ covers $B_{n, \tau}(C_0)$ for all $\tau \in \Delta$, we can apply the functional law of the iterated logarithm (LIL) (Goodman et al, 1981) to

$$\begin{aligned}
&n^{-1} \sum_{i=1}^n D \left(I\{0 < e_i(\tau) < \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \frac{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta) - e_i(\tau)}{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta)} \mathbf{Z}_i^T \right), \text{ and get} \\
&\sup_{\tau \in \Delta, \beta \in \partial B_{n, \tau}(C_0)} |IV_1| \leq O_p(n^{-1/2} \ell_2 n) (C_0 n^{-1/2} \ell_2 n).
\end{aligned}$$

Similarly, we can show that, for $j = 2, \dots, 8$,

$$\sup_{\tau \in \Delta, \beta \in \partial B_{n, \tau}(C_0)} \|IV_j\| \leq O_p(n^{-1/2} \ell_2 n) (C_0 n^{-1/2} \ell_2 n).$$

Therefore, we have

$$\sup_{\tau \in \Delta, \beta \in \partial B_{n, \tau}(C_0)} |IV| \leq O_p(n^{-1/2} \ell_2 n) (C_0 n^{-1/2} \ell_2 n). \quad (3)$$

For the term III , note that, under condition C2 (i), $\inf_{\tau \in \Delta} f_\tau(0|z) > 0$. By the definition of $B_{n, \tau}(C_0)$, there exists $N_1 > 0$ such that for $n \geq N_1$,

$$\sup_{\beta \in \partial B_{n,\tau}(C_0), \mathbf{z}} |\mathbf{z}^T \mathbf{u}(\tau; \beta)| \leq \delta_1 \wedge \inf_{\tau \in \Delta, \mathbf{z}} \{f_\tau(0|\mathbf{z}) / (2A_0)\}.$$

This, coupled with condition C2 (ii), implies that for $n > N_1$, $f_\tau(x|\mathbf{z}) > \inf_{\tau \in \Delta} f_\tau(0|\mathbf{z})/2$ for any $x \in (0, \mathbf{z}^T \mathbf{u}(\tau; \beta))$ with $\beta \in B_{n,\tau}(C_0)$. Let $\delta_2 \equiv \inf_{\tau} f_\tau(0|\mathbf{z})/2$ and $\delta_3 \equiv \text{eig min } E(\mathbf{Z}\mathbf{Z}^T)$, where $\text{eig min}(\cdot)$ denotes the minimum eigenvalue of a matrix. Let $E_{\mathbf{Z}}(\cdot)$ denote the expectation with regard to $\mathbf{Z}$. We get, for any $\tau \in \Delta$ and $\beta \in B_{n,\tau}(C_0)$,

$$\begin{aligned} & E[I\{0 < e_i(\tau) < \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} \{-e_i(\tau) + \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\}] \\ &= E_{\mathbf{Z}} \left[\int_0^{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta)} \{-x + \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} f_\tau(x|\mathbf{Z}) dx \right], \\ &\geq \delta_2 E_{\mathbf{Z}} \left[\int_0^{\mathbf{Z}_i^T \mathbf{u}(\tau; \beta)} \{-x + \mathbf{Z}_i^T \mathbf{u}(\tau; \beta)\} dx \right] \\ &= \delta_2 E_{\mathbf{Z}} \left\{ \frac{1}{2} \mathbf{u}(\tau; \beta)^T (\mathbf{Z} \mathbf{Z}^T) \mathbf{u}(\tau; \beta) \right\} \\ &\geq \frac{1}{2} \delta_2 \delta_3 C_0^2 (n^{-1/2} \ell_2 n)^2 > 0. \end{aligned}$$

Since a similar result can be shown for the second term in III, it follows that

$$\inf_{\beta \in \partial B_{n,\tau}(C_0), \tau \in \Delta} III \geq C_0^2 O_p((\ell_2 n)^2/n). \quad (4)$$

For the term II, it is easy to see that

$$II \geq -\frac{\lambda_n}{n} \sum_{j=2}^s |\beta_n^{(j)} - \beta_0^{(j)}(\tau)| w_{n,j}(\tau).$$

By the uniform consistency of $\tilde{\beta}_n(\tau)$, for $2 \leq j \leq s$,

$\sup_{\tau \in \Delta} |w_{n,j}(\tau)| = (\sup_{\tau \in \Delta} |\tilde{\beta}_n^{(j)}(\tau)|)^{-1} = O_p(1)$. Therefore, with $(n^{1/2} \ell_2 n)^{-1} \lambda_n = O(1)$,

$$\inf_{\beta \in \partial B_{n,\tau}(C_0), \tau \in \Delta} II \geq -C_0 \cdot O_p((\ell_2 n)^2/n). \quad (5)$$

Based on equations (3), (4), and (5), it follows that

$$\inf_{\tau \in \Delta, \beta \in \partial B_{n,\tau}(C_0)} [n^{-1} W_{n,\lambda_n}(\beta; \tau) - n^{-1} W_{n,\lambda_n}(\beta_0(\tau); \tau)] \geq C_0^2 \cdot O_p((\ell_2 n)^2/n) - C_0 \cdot O_p((\ell_2 n)^2/n).$$

Therefore, (2) holds if we choose C_0 large enough. This completes the proof of Theorem 1.

Appendix B: Proof of Theorem 2

Define $\text{sgn}(x) = I(x > 0) - I(x < 0)$, $U_{n,j}(\beta; \tau) = \frac{\partial W_{n,\lambda_n}(\beta; \tau)}{\partial \beta^{(j)}}$, and let $U_n(\beta; \tau) = (U_{n,1}(\beta; \tau), \dots, U_{n,p}(\beta; \tau))^T$. Define $\mu_j(\beta) = E[-Z_i^{(j)}\{\tau - I(Y_i - \mathbf{Z}_i^T \beta < 0)\}]$, $\boldsymbol{\mu}(\beta) = (\mu_1(\beta), \dots, \mu_p(\beta))^T$, and $\mathbf{A}(\beta) = \boldsymbol{\mu}(\beta)/\beta \equiv (\mathbf{A}_1(\beta)^T, \dots, \mathbf{A}_p(\beta)^T)^T$.

First, by Theorem 1, for $j = 2, \dots, s$, we have

$$\lim_{n \rightarrow \infty} \Pr(\sup_{\tau \in \Delta} |\hat{\beta}_{n,\lambda_n}^{US(j)}(\tau)| = 0) \leq \lim_{n \rightarrow \infty} \Pr(\sup_{\tau \in \Delta} |\hat{\beta}_{n,\lambda_n}^{US(j)}(\tau) - \beta_0^{(j)}(\tau)| \geq \sup_{\tau \in \Delta} |\beta_0^{(j)}(\tau)|) = 0. \quad (6)$$

Next, we note that given the uniform consistency of $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$ implied by Theorem 1, it can be shown by following the lines of Lemma 1 in Peng and Huang (2008) that

$$\begin{aligned} & \sup_{\tau \in \Delta} \left| -n^{-1/2} \sum_{i=1}^n Z_i^{(j)} [\tau - I\{Y_i - \mathbf{Z}_i^T \hat{\beta}_{n,\lambda_n}^{US}(\tau) < 0\}] \right. \\ & \quad \left. + n^{-1/2} \sum_{i=1}^n Z_i^{(j)} [\tau - I\{Y_i - \mathbf{Z}_i^T \beta_0(\tau) < 0\}] \right| \end{aligned} \quad (7)$$

$$-n^{1/2} [\mu_j\{\hat{\beta}_{n,\lambda_n}^{US}(\tau)\} - \mu_j\{\beta_0(\tau)\}] = o_p(1), \quad j=1, \dots, s. \quad (8)$$

This implies, for $j = s+1, \dots, p$,

$$\begin{aligned} & \sup_{\tau \in \Delta} |E_{n,j}(\tau)| \\ & \equiv \sup_{\tau \in \Delta} |n^{-1/2} U_{n,j}\{\hat{\beta}_{n,\lambda_n}^{US}(\tau); \tau\} - n^{-1/2} U_{n,j}\{\beta_0(\tau); \tau\}| \\ & \quad - n^{1/2} [\mu_j\{\hat{\beta}_{n,\lambda_n}^{US}(\tau)\} - \mu_j\{\beta_0(\tau)\}] \\ & \quad - n^{1/2} \lambda_n w_{n,j}(\tau) \text{sgn}\{\hat{\beta}_{n,\lambda_n}^{US(j)}(\tau)\} = o_p(1) \end{aligned} \quad (9)$$

By the definition of $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$,

$$\sup_{\tau \in \Delta} |n^{-1/2} U_{n,j}\{\hat{\beta}_{n,\lambda_n}^{US}(\tau); \tau\}| = o_p(1). \quad (10)$$

Applying Functional LIL to $n^{-1} U_{n,j}\{\beta_0(\tau); \tau\}$ gives

$$\sup_{\tau \in \Delta} |n^{-1/2} U_{n,j}\{\beta_0(\tau); \tau\}| = O_p(\ell_2 n). \quad (11)$$

An application of Taylor expansion to $\mu_j\{\hat{\beta}_{n,\lambda_n}^{US}(\tau)\} - \mu_j\{\beta_0(\tau)\}$, coupled with Theorem 1 and the uniform boundedness of $\mathbf{A}_j(\beta)$, shows that

$$\sup_{\tau \in \Delta} |n^{1/2} [\mu_j \{\hat{\beta}_{n,\lambda_n}^{US}(\tau)\} - \mu_j \{\beta_0(\tau)\}]| = O_p(\ell_2 n). \quad (12)$$

In addition, the proof of Theorem 1 can be used to justify that

$$\sup_{\tau \in \Delta} |\hat{\beta}_n^{(j)}(\tau)| = O_p(n^{-1/2} \ell_2 n) \text{ and thus}$$

$$\lim_{n \rightarrow \infty} \left(\frac{1}{\ell_2 n} \right) \left(n^{-1/2} \lambda_n w_{n,j}(\tau) \right) = \infty. \quad (13)$$

By (10), (11), (12), and (13), with a fixed $M > 0$, $\lim_{n \rightarrow \infty} \text{pr}(\sup_{\tau \in \Delta} |E_{n,j}(\tau)| \leq M,$

$\sup_{\tau \in \Delta} |\hat{\beta}_{n,\lambda_n}^{US(j)}(\tau)| \neq 0) = 0$. This, coupled with (9), implies that, for $j = s+1, \dots, p$,

$$\begin{aligned} & \lim_{n \rightarrow \infty} \text{pr}(\sup_{\tau \in \Delta} |\hat{\beta}_{n,\lambda_n}^{US(j)}(\tau)| \neq 0) \\ & \leq \lim_{n \rightarrow \infty} \{ \text{pr}(\sup_{\tau \in \Delta} |E_{n,j}(\tau)| \leq M, \sup_{\tau \in \Delta} |\hat{\beta}_{n,\lambda_n}^{US(j)}(\tau)| \neq 0) \} \end{aligned} \quad (14)$$

$$+ \text{pr}(|E_{n,j}(\tau)| > M) \} = 0. \quad (15)$$

The proof of Theorem 2 (i) is completed based on (6) and (14).

Let $\stackrel{a}{\equiv}$ denote asymptotic equivalence in the sense that the difference converges to zero in

probability uniformly in $\tau \in \Delta$. Define $\bar{\beta}_n(\tau) = \begin{pmatrix} \hat{\beta}_{n,\lambda_n}^{US(1:s)}(\tau) \\ 0 \end{pmatrix}$. By the result in Theorem 2 (i), we have

$$n^{-1/2} \mathbf{U}_n(\bar{\beta}_n; \tau) \stackrel{a}{\equiv} n^{-1/2} \mathbf{U}_n(\hat{\beta}_{n,\lambda_n}^{US}; \tau)$$

and hence $n^{-1/2} \mathbf{U}_n(\bar{\beta}_n; \tau) \stackrel{a}{\equiv} 0$. Using the result in (7) and applying Taylor expansion to $\mu_j \{\bar{\beta}_n(\tau)\} - \mu_j \{\beta_0(\tau)\}$, we get

$$\begin{aligned} 0 & \stackrel{a}{\equiv} n^{-1/2} \mathbf{U}_n(\bar{\beta}_n; \tau) \\ & \stackrel{a}{\equiv} -n^{-1/2} \sum_{i=1}^n \mathbf{Z}_i [\tau - I\{Y_i - \mathbf{Z}_i^T \beta_0(\tau) < 0\}] \\ & + [\mathbf{A} \{\bar{\beta}_n(\tau)\}] n^{1/2} \{\bar{\beta}_n(\tau) - \beta_0(\tau)\} + n^{-1/2} \lambda_n \mathbf{b}_n(\tau), \end{aligned} \quad (16)$$

where $\bar{\beta}_n(\tau)$ is on the line segment between $\beta_0(\tau)$ and $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$, and

$$\mathbf{b}_n(\tau) = (w_{n,1}(\tau) \text{sgn}\{\hat{\beta}_{n,\lambda_n}^{US(1)}(\tau)\}, \dots, w_{n,s}(\tau) \text{sgn}\{\hat{\beta}_{n,\lambda_n}^{US(s)}(\tau)\}, 0, \dots, 0)^T.$$

Since $\lim_{n \rightarrow \infty} n^{-1/2} \lambda_n = 0$, $\sup_{\tau \in \Delta} |w_{n,j}(\tau)| = O_p(1)$ for $j = 1, \dots, s$, and

$\mathbf{A} \{\bar{\beta}_n(\tau)\} \stackrel{a}{\equiv} \mathbf{A} \{\beta_0(\tau)\}$ given the uniform convergence of $\hat{\beta}_{n,\lambda_n}^{US}(\tau)$ to $\beta_0(\tau)$, it follows from (16) that

$$n^{1/2}\{\hat{\beta}_{n,\lambda_n}^{US(1:s)}(\tau) - \beta_0^{(1:s)}(\tau)\} \\ \stackrel{a}{=} \mathbf{A}_{11}\{\beta_0(\tau)\}n^{-1/2}\sum_{i=1}^n \mathbf{Z}_i^{(1:s)} I\{Y_i - \mathbf{Z}_i^T \beta_0(\tau) < 0\}, \quad (17)$$

where $\mathbf{A}_{11}(\cdot)$ stands for the submatrix of $\mathbf{A}(\cdot)$ formed by the first s rows and columns. An application of the Donsker theorem based on (17) thus shows that

$n^{1/2}\{\hat{\beta}_{n,\lambda_n}^{US(1:s)}(\tau) - \beta_0^{(1:s)}(\tau)\}$ converges weakly to a mean zero Gaussian process with the covariance matrix,

$$\Sigma(\tau', \tau) = E[\mathbf{A}_{11}\{\beta_0(\tau')\} \mathbf{Z}_i^{(1:s)} \mathbf{Z}_i^{(1:s)T} \mathbf{A}_{11}^T\{\beta_0(\tau)\} \\ I\{Y_i - \mathbf{Z}_i^T \beta_0(\tau') < 0\} I\{Y_i - \mathbf{Z}_i^T \beta_0(\tau) < 0\}].$$

Using similar steps, we can show that (17) still holds when $\hat{\beta}_{\text{oracle}}(\tau)$ is in place of $\hat{\beta}_{n,\lambda_n}^{US(1:s)}(\tau)$. This completes the proof of Theorem 2.

References

- Belloni A, Chernozhukov V. l_1 penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*. 2011; 82:82–130.
- Carey JR, Liedo P, Orozco D, Tatar M, Vaupel JW. A male-female longevity paradox in medfly cohorts. *The Journal of Animal Ecology*. 1995; 64:107–116.
- Dickson E, Grambsch P, Fleming T, Fisher L, Langworthy A. Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*. 1989; 10:1–7. [PubMed: 2737595]
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Frank I, Friedman J. A statistical view of some chemometrics regression tools. *Technometrics*. 1993
- Goodman V, Kuelbs J, Zinn J. Some results on the l_1 in Banach space with applications to weighted empirical processes. *Annals of Probability*. 1981; 9:713–752.
- Huang Y. Calibration regression of censored lifetime medical cost. *Journal of the American Statistical Association*. 2002; 98
- Huang Y. Quantile Calculus and Censored Regression. *The Annals of Statistics*. 2010; 38(3):1607–1637.
- Jensen G, Torp-Pedersen C, Hildebrandt P, Kober L, Nielsen F, Melchior T, Joen T, Andersen P. Does in-hospital fibrillation affect prognosis after myocardial infarction? *European Heart Journal*. 1997; 18:919–924. [PubMed: 9183582]
- Kai B, Li R, Zou H. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of Statistics*. 2011; 39:305–332. [PubMed: 21666869]
- Kaslow R, Ostrow D, Detels R, Phair J, Polk B, Rinaldo C. The multicenter AIDS cohort study: rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology*. 1987; 126:310–318. [PubMed: 3300281]
- Knight K, Fu W. Asymptotics for lasso-type estimators. *The Annals of Statistics*. 2000:1356–1378.
- Koenker R. Quantile regression. Cambridge University Press; 2005.
- Koenker R. quantreg: Quantile regression. 2011. <http://www.r-project.org>
- Koenker R, Bassett G. Regression quantiles. *Econometrica*. 1978; 46:33–50.
- Koenker R, d'Orey V. Computing regression quantiles. *Applied Statistics*. 1987; 36:383–393.

- Kutner N, Clow P, Zhang R, Aviles X. Association of fish intake and survival in a cohort of incident dialysis patients. *American Journal of Kidney Diseases*. 2002; 39:1018–1024. [PubMed: 11979345]
- Li Y, Zhu J. Quantile regression in reproducing kernel hilbert spaces. *Journal of Computational Graphical Statistics*. 2005; 17:163–185.
- Lustig I, Marsden R, Shanno D. Interior point methods for linear programming: Computational state of the art with discussion. *ORSA Journal on Computing*. 1994; 6:1–14.
- Ma Y, Yin G. Semiparametric median residual life model and inference. *Canadian Journal of Statistics*. 2010; 38:665–679.
- Madsen K, Nielsen H. A finite smoothing algorithm for linear l_1 estimation. *SIAMJ Optimization*. 1993; 3:223–235.
- McDonald G, Schwing R. Instabilities of regression estimates relating air pollution to mortality. *Technometrics*. 1973; 15:463–481.
- Neocleous T, Vanden Branden K, Portnoy S. Correction to censored regression quantiles by portnoy, s. (2003), 1001–1012. *J Am Statist Assoc*. 2006; 101(474):860–861.
- Peng L, Fine J. Competing risks quantile regression. *Journal of the American Statistical Association*. 2009; 104
- Peng L, Huang Y. Survival analysis with quantile regression models. *J Am Statist Assoc*. 2008; 103(482):637–649.
- Portnoy S. Censored regression quantiles. *J Am Statist Assoc*. 2003; 98(464):1001–1012.
- Portnoy S, Lin G. Asymptotics for censored regression quantiles. 2010
- Rocha, G.; Wang, X.; Yu, B. Asymptotic distribution and sparsistency for l_1 -penalized parametric mestimators with applications to linear svm and logistic regression. 2009. <http://arxiv.org/abs/09081940>
- Thorogood J, Persijn G, Schreuder G, D'amaro J, Zantvoort F, Van Houwelingen J, Van Rood J. The effect of hla matching on kidney graft survival in separate posttransplantation intervals. *Transplantation*. 1990; 50:146–150. [PubMed: 2368135]
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society: Series B*. 1996; 58:267–288.
- Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. *Ann Statist*. 1990; 18:354–372.
- Van der Vaart, A.; Wellner, J. *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Verlag; 2000.
- Verweij P, Van Houwelingen H. Time-dependent effects of fixed covariates in cox regression. *Biometrics*. 1995; 51:1550–1556. [PubMed: 8589239]
- Wang H, Leng C. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*. 2007; 102:1039–1048.
- Wang H, Xia Y. Shrinkage estimation of the varying coefficient model. *Journal of the Americal Statistical Association*. 2009; 104:747–757.
- Wei LJ, Gail MH. Nonparametric estimation for a scale-change with censored observations. *J Am Statist Assoc*. 1983; 78:382–388.
- Wu Y, Liu Y. Variable selection in quantile regression. *Statistica Sinica*. 2009; 19:801–817.
- Ying Z. A large sample study of rank estimation for censored data. *Annals of Statistics*. 1993:76–99.
- Zhang H, Lu W. Adaptive lasso for cox's proportional hazards model. *Biometrika*. 2007; 94:691–703.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the Americal Statistical Association*. 2006; 101:1418–1429.
- Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. *Annals of Statistics*. 2008; 36:1108–1126.

Fig. 1.
True coefficient functions in simulation set-ups.

Fig. 2. Air pollution example: Regression quantiles (bold dotted lines) and the corresponding pointwise 95% confidence intervals (shaded areas), and least square estimates (horizontal solid lines) and the corresponding pointwise 95% confidence intervals (horizontal dashed lines).

Fig. 3.

Dialysis example: Regression quantiles (dotted lines) and the corresponding pointwise 95% confidence intervals (shaded areas), and penalized AFT estimates (horizontal solid lines) and the corresponding pointwise 95% confidence intervals (horizontal dashed lines).

Table 1

Simulation results for uncensored data

Set-up	Method	n	NC	NIC	PUF (%)	PCF (%)	POF (%)	MREE _F (%)	MREE _O (%)
(I)	US	200	4.95	.00	.0	95.0	5.0	36.5	110.7
		400	5.00	.00	.0	99.5	.5	35.4	104.8
	PS	200	4.19	.00	.0	46.5	53.5	36.2	106.7
		400	4.49	.00	.0	59.8	40.2	34.6	103.4
	SS(0.25)	200	4.92	.00	.0	93.2	6.8	34.3	107.0
		400	4.96	.00	.0	95.8	4.2	32.1	99.1
	SS(0.50)	200	4.95	.00	.0	95.5	4.5	31.2	94.4
		400	4.97	.00	.0	97.5	2.5	29.4	91.9
	SS(0.75)	200	4.91	.00	.0	91.8	8.2	33.2	104.3
		400	4.97	.00	.0	97.8	2.2	31.7	100.7
	LS	200	4.90	.00	.0	90.8	9.2	24.3	73.8
		400	4.95	.00	.0	95.2	4.8	24.3	74.4
(II)	US	200	4.93	.00	.2	92.5	7.2	50.8	160.3
		400	4.99	.00	.0	99.2	.8	42.6	122.8
	PS	200	4.32	.00	.0	52.5	47.5	48.2	147.7
		400	4.76	.00	.0	78.8	21.2	44.7	129.4
	SS(0.25)	200	4.93	.19	13.8	80.5	5.8	115.4	364.9
		400	4.98	.00	.0	98.2	1.8	166.5	500.0
	SS(0.50)	200	4.99	.98	98.0	2.0	.0	84.7	267.6
		400	5.00	1.00	99.5	.5	.0	127.6	371.1
	SS(0.75)	200	4.96	.03	3.0	93.5	3.5	111.5	356.4
		400	5.00	.00	.2	99.8	.0	166.6	490.4
	LS	200	4.82	0.96	95.2	2.5	2.2	101.9	318.5
		400	4.93	0.98	98.2	1.8	0.0	142.1	423.2
(III)	US	200	4.94	0.01	1.0	93.8	5.2	33.5	105.1
		400	4.99	.00	.0	99.0	1.0	32.3	96.2
	PS	200	4.42	.00	.0	58.2	41.8	28.1	89.5

Set-up	Method	<i>n</i>	NC	NIC	PUF (%)	PCF (%)	POF (%)	MREE _F (%)	MREE _O (%)
		400	4.81	.00	.0	83.2	16.8	27.1	82.5
	SS(0.25)	200	4.95	1.00	99.2	.8	.0	94.5	296.1
		400	4.98	.99	99.0	.8	.2	141.2	429.6
	SS(0.50)	200	4.98	1.94	100.0	.0	.0	78.5	242.8
		400	4.99	1.94	99.8	0.2	.0	113.9	341.5
	SS(0.75)	200	4.96	0.98	98.0	1.8	0.2	95.2	297.3
		400	4.98	1.00	99.8	.2	.0	143.9	428.3
	LS	200	4.78	1.05	63.2	24.5	12.2	90.1	277.0
		400	4.79	.36	24.0	60.5	15.5	133.1	394.6

Table 2

Simulation results for censored data

Set-up	Method	n	NC	NIC	PUF (%)	PCF (%)	POF (%)	MREE _F (%)	MREE _O (%)
(IV)	US _C	200	4.77	.00	.0	83.0	17.0	38.3	113.2
		400	4.82	.00	.0	93.0	7.0	35.7	107.6
	PS _C	200	2.62	.00	.0	9.5	90.5	42.6	121.9
		400	2.45	.00	.0	8.2	91.8	42.0	118.1
	SS _C (0.25)	200	4.63	.00	.0	73.8	26.2	46.8	132.9
		400	4.62	.00	.0	76.5	23.5	42.8	127.4
	SS _C (0.50)	200	4.76	.00	.0	83.2	16.8	31.7	97.0
		400	4.72	.00	.0	81.0	19.0	33.4	99.8
	SS _C (0.75)	200	4.66	.00	.0	76.0	24.0	32.1	94.6
		400	4.7	.00	.0	80.8	19.2	31.7	93.5
	AAFT	200	4.75	.00	.0	77.8	22.2	26.4	78.1
		400	4.84	.00	.0	85.0	15.0	24.5	72.4
	PH	200	4.85	.00	.0	86.2	13.8	37.9	112.2
		400	4.91	.00	.0	91.0	9.0	33.6	99.4
(V)	US _C	200	4.88	.00	.2	89.5	10.2	54.1	169.5
		400	4.97	.00	.0	97.5	2.5	44.1	128.5
	PS _C	200	3.96	.00	.0	38.2	61.8	48.9	151.1
		400	4.48	.00	.0	59.0	41.0	46.1	136.9
	SS _C (0.25)	200	4.93	.31	21.5	73.5	5.0	93.1	302.7
		400	4.98	.02	1.8	96.0	2.2	124.5	379.3
	SS _C (0.50)	200	4.97	.95	94.5	5.2	.2	70.6	226.3
		400	5.00	.97	96.8	3.2	.0	98.9	300.2
	SS _C (0.75)	200	4.78	.06	6.5	75.0	18.5	112.2	360.0
		400	4.94	.00	.2	94.0	5.8	158.1	477.5
	AAFT	200	4.78	0.56	55.5	33.8	10.8	100.9	320.4
		400	4.80	0.38	38.0	46.5	15.5	143.1	438.4
	PH	200	4.63	.26	25.8	51.5	22.8	—	—
		400	4.75	.02	2.5	75.8	21.8	—	—

Set-up	Method	n	NC	NIC	PUF (%)	PCF (%)	POF (%)	MRE _{E_T} (%)	MRE _{E_O} (%)
(VI)	US _C	200	4.86	.11	11.2	78.0	10.8	34.9	111.5
		400	4.98	.02	2.0	95.8	2.2	30.9	92.6
	PS _C	200	3.64	.00	.2	24.5	75.2	32.3	101.5
		400	4.05	.00	.0	35.2	64.8	30.7	93.8
	SS _C (0.25)	200	4.96	1.05	98.5	1.5	.0	68.6	221.1
		400	4.99	1.00	100.0	.0	.0	103.4	309.2
	SS _C (0.50)	200	4.91	1.8	98.5	1.2	.2	63.5	200.9
		400	4.96	1.83	99.0	1.0	.0	93.6	277.3
	SS _C (0.75)	200	4.69	.93	92.0	4.2	3.8	92.6	294.6
		400	4.86	.95	95.2	3.2	1.5	136.2	414.2
	AAFT	200	4.60	1.03	89.2	4.2	6.5	81.9	261.6
		400	4.75	0.92	87.0	7.8	5.2	120.0	349.5
	PH	200	4.82	.99	89.8	6.8	3.5	—	—
		400	4.87	.80	79.8	15.5	4.8	—	—

Table 3

Example 1: Analysis of the air pollution to mortality data

Method	Selected variables	APE
US	x_1, x_9, x_{10}, x_{14}	0.215
PS	$x_1, x_2, x_3, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{12}, x_{13}, x_{14}, x_{15}$	0.345
LS	$x_1, x_2, x_6, x_9, x_{13}$	0.277
SS(0.25)	$x_1, x_3, x_9, x_{10}, x_{13}, x_{14}$	0.243
SS(0.5)	$x_1, x_8, x_9, x_{10}, x_{12}, x_{14}$	0.239
SS(0.75)	$x_6, x_8, x_9, x_{12}, x_{14}$	0.255

Table 4

Example 2: Analysis of the dialysis data

Method	Selected variables	APE
US	x_1, x_2, x_5, x_{14}	0.236
PS	$x_1, x_2, x_3, x_4, x_5, x_{11}, x_{14}$	0.244
SS(0.25)	x_1, x_{14}	0.244
SS(0.50)	x_2, x_{14}	0.242
SS(0.75)	x_2, x_{11}, x_{14}	0.243
AAFT	$x_2, x_4, x_{10}, x_{13}, x_{14}$	0.252
PH	$x_2, x_4, x_8, x_{10}, x_{11}, x_{13}, x_{14}$	0.258