Taylor & Francis
Taylor & Francis Group

Check for updates

Inference for High-Dimensional Censored Quantile Regression

Zhe Fei[a], Qi Zheng[b], Hyokyoung G. Hong[c], and Yi Li[d]

[a]Department of Biostatistics, University of California, Los Angeles, CA; [b]Department of Bioinformatics and Biostatistics, University of Louisville, KY; [c]Department of Statistics and Probability, Michigan State University, East Lansing, MI; [d]Department of Biostatistics, University of Michigan, Ann Arbor, MI

ABSTRACT

With the availability of high-dimensional genetic biomarkers, it is of interest to identify heterogeneous effects of these predictors on patients' survival, along with proper statistical inference. Censored quantile regression has emerged as a powerful tool for detecting heterogeneous effects of covariates on survival outcomes. To our knowledge, there is little work available to draw inferences on the effects of high-dimensional predictors for censored quantile regression (CQR). This article proposes a novel procedure to draw inference on all predictors within the framework of global CQR, which investigates covariate-response associations over an interval of quantile levels, instead of a few discrete values. The proposed estimator combines a sequence of low-dimensional model estimates that are based on multi-sample splittings and variable selection. We show that, under some regularity conditions, the estimator is consistent and asymptotically follows a Gaussian process indexed by the quantile level. Simulation studies indicate that our procedure can properly quantify the uncertainty of the estimates in high-dimensional settings. We apply our method to analyze the heterogeneous effects of SNPs residing in lung cancer pathways on patients' survival, using the Boston Lung Cancer Survival Cohort, a cancer epidemiology study on the molecular mechanism of lung cancer.

1. Introduction

Lung cancer presents much heterogeneity in etiology (McKay et al. 2017; Dong et al. 2012; Huang et al. 2009), and some genetic variants may insert different impacts on different quantile levels of survival time. For example, in the Boston Lung Cancer Survival Cohort (BLCSC) (Christiani 2017), a cancer epidemiology cohort of over 11,000 lung cancer cases enrolled in the Boston area since 1992, it was found that SNP AX.37793583 (rs115952579), along with age, gender, cancer stage and smoking status, had heterogeneous effects on different quantiles of survival time. A total of 674 patients in the study were genotyped, with the goal of identifying lung cancer survival-predictive SNPs. Target gene approaches, which focus on SNPs residing in cancer-related gene pathways, are appealing for increased statistical power in detecting significant SNPs (Moon, Oh, and Roth 2003; Risch and Plass 2008; Ho et al. 2019), and the investigators have identified SNPs residing in 14 well-known lung cancer-related genes (Zhu et al. 2017; Korpanty et al. 2014; Yamamoto et al. 2008; Kelley, Li, and Harpole 2001). One goal was to investigate whether and how each SNP might play a different role among the high-risk (i.e., lower quantiles of overall survival) and low-risk (i.e., higher quantiles of overall survival) cancer survivors.

Quantile regression (QR) (Koenker and Bassett Jr 1978) is a significant extension of classic linear regression. By permitting the effects of active variables to vary across quantile levels, quantile regression can naturally accommodate and examine the heterogeneous impacts of biomarkers on different segments of the response variable's conditional distribution. As survival data are subject to censoring and may be incomplete, QR methods developed for complete data may be unsuitable. Efforts have been devoted to developing censored quantile regression (CQR) (Powell 1986; Portnoy 2003; Peng and Huang 2008, among others), which has become a useful alternative strategy to traditional survival models, such as the Cox model and the accelerated failure time model. QR has also been widely studied to accommodate high-dimensional predictors. For example, Wang, Wu, and Li (2012) dealt with variable selection using nonconvex penalization; Zheng, Gallagher, and Kulasekera (2013) proposed an adaptive penalized quantile regression estimator that can select the true sparse model with high probability; and Fan, Fan, and Barut (2014) studied the penalized quantile regression with a weighted L_1 penalty in an ultra-high-dimensional setting. As to high dimensional CQR (HDCQR), He, Wang, and Hong (2013) provided a model-free variable screening procedure for ultra-high-dimensional covariates, and Zheng, Peng, and He (2018) proposed a penalized HDCQR built upon a stochastic integral-based estimating equation. However, most of the existing works in HDCQR were designed to select a subset of predictors and estimate the effects of the selected variables, instead of drawing inference on all predictors.

Progress in high-dimensional inferences has been made for linear and nonlinear models (Zhang and Zhang 2014;

Bühlmann, Kalisch, and Meier 2014; Javanmard and Montanari 2014; Ning and Liu 2017; Fei et al. 2019; Fei and Li 2021). For example, Meinshausen, Meier, and Bühlmann (2009) proposed to aggregate p-values from multi-sample splittings for high-dimensional linear regression. Another line of works referred to as *post-selection inference* includes Berk et al. (2013), Lee et al. (2016), and Belloni, Chernozhukov, and Kato (2019), which provided post-selection inference at fixed quantiles for complete data. However, these methods may not handle censored outcomes. For censored median regression, Shows, Lu, and Zhang (2010) provided sparse estimation and inference, but it cannot handle high-dimensional data.

We propose to draw inferences on high-dimensional HDCQR based on a splitting and fusing scheme, termed Fused-HDCQR. Utilizing a variable selection procedure for HDCQR such as Zheng, Peng, and He (2018), our method operates partial regression followed by smoothing. Specifically, partial regression allows us to estimate the effect of each predictor, regardless of whether or not it is chosen by variable selection. The fused estimator aggregates the estimates based on multiple data-splittings and variable selection, with a variance estimator derived by the functional delta method (Efron 2014; Wager and Athey 2018). To comprehensively assess the covariate effects on the survival distribution, we adopt a "global" quantile model (Zheng, Peng, and He 2015) with the quantile level varying over an interval, instead of a local CQR that focuses only on a few pre-specified quantile levels. The global quantile model can address the molecular mechanism of lung cancer, our motivating disease, that hypothesizes that some genetic variants may cause heterogeneous impacts on different but unspecified segments of survival distribution (McKay et al. 2017; Dong et al. 2012; Huang et al. 2009).

Our work presents several advantages. First, compared to high-dimensional Cox models (Zhao and Li 2012; Fang, Ning, and Liu 2017; Kong et al. 2021), the employed HDCQR stems from the accelerated failure time model (Wei 1992) and offers straightforward interpretations (Hong, Christiani, and Li 2019). Second, utilizing the global conditional quantile regression, it uses various segments of the conditional survival distribution to improve the robustness of variable selection and capture global sparsity. Third, our splitting-and-averaging scheme avoids the technicalities of estimating the precision matrix by inverting a $p \times p$ Hessian matrix of the log-likelihood, which is a major challenge for debiased-LASSO type methods (Zhang and Zhang 2014; Van de Geer et al. 2014) and is even more so if we apply debiased-LASSO to the CQR setting. Finally, as opposed to post-selection inferences (Belloni, Chernozhukov, and Kato 2019, among others), Fused-HDCQR accounts for variations in model selection and draws inferences for all of the predictors.

The rest of the article is organized as follows. Section 2 introduces the method, and Section 3 details the asymptotic properties. Section 4 derives a nonparametric variance estimator, Section 5 conducts simulation studies, and Section 6 applies the proposed method to analyze the BLCSC data. The technical details, such as proofs and additional lemmas, are relegated to the online supplementary materials.

2. Model and Method

2.1. High-Dimensional Censored Quantile Regression

Let T and C denote the survival outcome and censoring time, respectively. We assume that C is independent of T given $\widetilde{\mathbf{Z}}$, a $(p-1)$-dimensional vector of covariates ($p > 1$). Let $X = \min\{T, C\}, \Delta = 1\{T \leq C\}$, and $\mathbf{Z} = (1, \widetilde{\mathbf{Z}}^{\mathrm{T}})^{\mathrm{T}}$, where $1\{\cdot\}$ is the binary indicator function. The observed data, $D^{(n)} = \{(X_i, \Delta_i, \mathbf{Z}_i), i = 1, \ldots, n\}$, are n identical and independently distributed (iid) copies of $(X, \Delta, \mathbf{Z})$. With $Y = \log T$, let $Q_Y(\tau|\mathbf{Z}) = \inf\{t : \mathrm{P}(Y \leq t|\mathbf{Z}) \geq \tau\}$ be the τth conditional quantile of Y given $\mathbf{Z}$. A global CQR model stipulates

$$Q_Y(\tau|\mathbf{Z}) = \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau), \ \tau \in (0, 1), \tag{1}$$

where $\boldsymbol{\beta}^*(\tau)$ is a p-dimensional vector of coefficients at τ. We aim to draw inferences on $\beta_j^*(\tau)$ for each $\tau \in (0, \tau_U]$ and for all $j \in \{1, \ldots, p\}$, where $0 < \tau_U < 1$ is an upper bound for estimable quantiles subject to identifiability constraint caused by censoring (Peng and Huang 2008).

Let $N(t) = 1\{\log X \leq t, \Delta = 1\}$, $\Lambda_T(t|\mathbf{Z}) = -\log(1 - \mathrm{P}(\log T \leq t|\mathbf{Z}))$, and $H(u) = -\log(1-u)$. Then, $M(t) = N(t) - \Lambda_T(t \wedge \log X|\mathbf{Z})$ is a martingale process under model (1) (Fleming and Harrington 2011) and hence $E(M(t)|\mathbf{Z}) = 0$. We use $N_i(t)$ and $M_i(t)$, $i = 1, \ldots, n$, to denote the sample analogs of $N(t)$ and $M(t)$. Let $\theta_i(\tau) = \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}(\tau)$ and

$$\mathbf{U}_n(\boldsymbol{\beta}, \tau) = n^{-1} \sum_{i=1}^{n} \mathbf{Z}_i \left\{ N_i(\theta_i(\tau)) - \int_0^\tau 1\{\log X_i \geq \theta_i(u)\} dH(u) \right\}.$$

We denote by $\mathbf{u}(\boldsymbol{\beta}, \tau)$ the expectation of $\mathbf{U}_n(\boldsymbol{\beta}, \tau)$.

The martingale property implies $\mathbf{u}(\boldsymbol{\beta}^*, \tau) = 0$ with $\tau \in [0, \tau_U]$, entailing an estimating equation with $\tau \in (0, \tau_U]$

$$n^{1/2}\mathbf{U}_n(\boldsymbol{\beta}, \tau) = n^{-1/2} \sum_{i=1}^{n} \mathbf{Z}_i \Bigg\{ N_i(\theta_i(\tau)) \\ - \int_0^\tau 1\{\log X_i \geq \theta_i(u)\} dH(u) \Bigg\} = 0. \tag{2}$$

The stochastic integral in Equation (2) naturally suggests sequential estimation with respect to τ. We define a grid of quantile values $\Gamma_m = \{\tau_0, \tau_1, \ldots, \tau_m\}$ to cover the interval $[\nu, \tau_U]$, where $\tau_0 = \nu$ and $\tau_m = \tau_U$. The assumption on the lower bound $\nu > 0$ is made to circumvent the singularity problem with CQR at $\tau = 0$, as detailed in Assumption (A1). In practice, ν is chosen such that only a small proportion of observations are censored below the νth quantile.

Then, $\hat{\boldsymbol{\beta}}(\tau_k)$'s, the estimates of $\boldsymbol{\beta}(\tau_k)$'s, $\tau_k \in \Gamma_m$, can be sequentially obtained by solving

$$n^{-1/2} \sum_{i=1}^{n} \mathbf{Z}_i \left(N_i(\theta_i(\tau_k)) - \sum_{r=0}^{k-1} \int_{\tau_r}^{\tau_{r+1}} 1\{\log X_i \geq \acute{\theta}_i(\tau_r)\} dH(u) \right) = 0,$$

where $\acute{\theta}_i(\tau_k) = \mathbf{Z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}(\tau_k)$. Due to the monotonicity of $\theta_i(\tau)$ in τ, $\hat{\boldsymbol{\beta}}(\tau_k)$ can be solved efficiently via L_1-minimization. And $\hat{\boldsymbol{\beta}}(\tau)$, $\tau \in [\nu, \tau_U]$, is defined as a right-continuous piece-wise constant function that only jumps at the grid points. It can be shown that $\hat{\boldsymbol{\beta}}(\tau)$ is uniformly consistent and converges weakly

to a mean zero Gaussian process for $\tau \in [\nu, \tau_U]$ when $p = o(n)$. More importantly, $\hat{\boldsymbol{\beta}}(\tau)$ provides a comprehensive understanding of the covariate effects on the conditional survival distribution over the quantile interval $[\nu, \tau_U]$. We incorporate this sequential estimating procedure for low-dimensional CQR estimation in our proposed method.

In addition, our method requires dimension reduction, which can be accomplished by existing methods, including the screening method proposed by He, Wang, and Hong (2013) and the penalized estimation and selection procedure developed by Zheng, Peng, and He (2018). Specifically, Zheng, Peng, and He (2018) incorporated an L_1 penalty into the stochastic integral-based estimating equation in (2) to obtain an L-HDCQR estimator, which achieves a uniform convergence rate of $\sqrt{q \log(p \vee n)/n}$, and results in "sure screening" variable selection with high probability, where q is defined in Condition (A4). Zheng, Peng, and He (2018) also proposed an AL-HDCQR estimator by employing the Adaptive Lasso penalties, which attains a uniform convergence rate of $\sqrt{q \log(n)/n}$ and selection consistency.

2.2. Fused-HDCQR Estimator

Our proposed Fused-HDCQR procedure consists of splitting data multiple times, selecting variables, fitting low-dimensional CQRs with partitioned data, applying *append-and-estimate* to all predictors, and aggregating those estimates.

1. With the full data $D^{(n)}$, determine via cross-validation the tuning parameter(s) λ_n of $\mathcal{S}$, an HDCQR variable selection method.
2. Let B be a large positive integer. For each $b = 1, \ldots, B$,

   (i) randomly split the data into equal halves, D_1^b and D_2^b;
   (ii) on D_1^b, apply $\mathcal{S}$ with λ_n on $[\nu, \tau_U]$, to select a subset of predictors, denoted by $\hat{S}_{\lambda_n}^b$, or $\hat{S}^b$ for short;
   (iii) on D_2^b, for each $j = 1, \ldots, p$, append j to $\hat{S}^b$ such that $\hat{S}_{+j}^b = \{j\} \cup \hat{S}^b$, fit a partial CQR on the covariates indexed by $\hat{S}_{+j}^b$, and denote their coefficient estimates by $\tilde{\boldsymbol{\beta}}_{\hat{S}_{+j}^b}(\tau), \tau \in [\nu, \tau_U]$. Here, $\tilde{\boldsymbol{\beta}}_{\hat{S}_{+j}^b}(\tau)$ is a right-continuous piecewise-constant function with jumps only at the grid points of $\tau_k \in \Gamma_m$;
   (iv) denote by $\tilde{\beta}_j^b(\tau) = \left(\tilde{\boldsymbol{\beta}}_{\hat{S}_{+j}^b}(\tau)\right)_j$ the entry of $\tilde{\boldsymbol{\beta}}_{\hat{S}_{+j}^b}(\tau)$ corresponding to Z_j.

3. Fusing: the final estimate of $\beta_j^*(\tau)$, $\tau \in [\nu, \tau_U]$, $j = 1, \ldots, p$ is

$$\hat{\beta}_j(\tau) = \frac{1}{B} \sum_{b=1}^{B} \tilde{\beta}_j^b(\tau). \tag{3}$$

Remark 1. We could select different tuning parameters for $\mathcal{S}$ in each data split, but with much added computation. Our numerical evidence suggests that a globally chosen λ_n work well.

Remark 2. Our procedure needs a variable selection procedure to reduce dimension. For example, L-HDCQR selects a subset: $\{j \in \{2, \ldots, p\} : \max_k |\hat{\gamma}_j(\tau_k)| > a_0, \tau_k \in \Gamma_m\}$, where $\hat{\gamma}_j(\tau_k)$'s are the L-HDCQR estimates, $a_0 > 0$ is a predetermined threshold,

and j starts with 2 as the intercept term (corresponding to $j = 1$) is always included in the model. For the choice of variable selection methods, our experience suggests that we adopt the screening method in He, Wang, and Hong (2013) for fast computation, use L-HDCQR for detecting any nonzero effects in the quantile interval $[\nu, \tau_U]$, and choose AL-HDCQR if we opt to select fewer predictors.

Remark 3. We select λ_n by minimizing a K-fold cross-validation error defined by deviance residuals in the presence of censored outcomes (Zheng, Peng, and He 2018). Specifically, we partition the data to K folds, and let $\hat{\boldsymbol{\beta}}_\lambda^{(-k)}(\tau)$ be the penalized estimate of $\boldsymbol{\beta}(\tau)$ using all of the data excluding the kth fold with a tuning parameter λ and $\tau \in [\nu, \tau_U]$, where $k = 1, \ldots, K$. Under the global CQR model (1), we define the cross-validation error as

$$\text{CV Error}(\lambda) = \sum_{k=1}^{K} \sum_{i \in \text{fold } k} \int_\nu^{\tau_U} |D_i[\hat{\boldsymbol{\beta}}_\lambda^{(-k)}(\tau)]| d\tau, \tag{4}$$

where

$$D_i[\boldsymbol{\beta}(\tau)] = \frac{\text{sign}\{M_i(\boldsymbol{\beta}(\tau))\}}{\sqrt{-2M_i(\boldsymbol{\beta}(\tau)) + \Delta_i \log\{\Delta_i - M_i(\boldsymbol{\beta}(\tau))\}}}$$

with $M_i(\boldsymbol{\beta}(\tau)) = N_i(\mathbf{Z}_i^T \boldsymbol{\beta}(\tau)) - \int_\nu^\tau 1\{\log X_i \geq N_i(\mathbf{Z}_i^T \boldsymbol{\beta}(u))\} dH(u) - \nu$. Here, $H(u) = -\log(1-u)$, $N_i(\cdot)$ is the counting process, and $M_i(\boldsymbol{\beta}(\tau))$ is the martingale residual under model (1) (Zheng, Peng, and He 2018).

Remark 4. When carrying out quantile regression at each grid point, we formulate it as a linear programming problem (Koenker 2005), which can be solved by a simplex algorithm with a computational complexity of $O(n^2 p)$ (Klee and Minty 1972). Since our grid size is $O(n)$ and the number of resampling, B, is $O(n)$, the computational complexity of our procedure is $O(n^4 p)$.

3. Theoretical Studies

3.1. Notation and Regularity Conditions

For any vector $\boldsymbol{\delta} \in \mathbf{R}^p$ and a subset $S \subset \{1, \ldots, p\}$, denote by S^C its complementary set, and define $\|\boldsymbol{\delta}\|_{r,S} = \|\boldsymbol{\delta}_S\|_r$, the l_r-norm of the sub-vector $\boldsymbol{\delta}_S$, in which $\delta_{jS} = \delta_j$ if $j \in S$ and $\delta_{jS} = 0$ if $j \in S^C$. We set the following conditions.

(A1) There exist a quantile level ν and a constant $c > 0$ such that

$$n^{-1} \sum_{i=1}^{n} 1\left\{\log C_i \leq \mathbf{Z}_i^T \boldsymbol{\beta}^*(\nu)\right\} (1 - \Delta_i) \leq cn^{-1/2}$$

holds for sufficiently large n.

(A2) (*Bounded observations*) $\|\mathbf{Z}\|_\infty \leq C_0$. Without loss of generality, we assume $C_0 = 1$. In addition, $E|\log X| < \infty$.

(A3) (*Bounded densities*) Let $F_T(t|\mathbf{Z}) = P(\log T \leq t|\mathbf{Z})$, $\Lambda_T(t|\mathbf{Z}) = -\log(1 - F_T(t|\mathbf{Z}))$, $F(t|\mathbf{Z}) = P(\log X \leq t|\mathbf{Z})$, and $G(t|\mathbf{Z}) = P(\log X \leq t, \Delta = 1|\mathbf{Z})$. Also, define $f(t|\mathbf{Z}) = dF(t|\mathbf{Z})/dt$, and $g(t|\mathbf{Z}) = dG(t|\mathbf{Z})/dt$.

(a) There exist constants $\underline{f}, \bar{f}, \underline{g}$ and $\bar{g}$ such that

$$\underline{f} \leq \inf_{\mathbf{z}, \tau \in [\nu, \tau_U]} f(\mathbf{z}^T \boldsymbol{\beta}^*(\tau)|\mathbf{z}) \leq \sup_{\mathbf{z}, \tau \in [\nu, \tau_U]} f(\mathbf{z}^T \boldsymbol{\beta}^*(\tau)|\mathbf{z}) \leq \bar{f},$$

$$\underline{g} \leq \inf_{\mathbf{z}, \tau \in [\nu, \tau_U]} g(\mathbf{z}^T \boldsymbol{\beta}^*(\tau)|\mathbf{z}) \leq \sup_{\mathbf{z}, \tau \in [\nu, \tau_U]} g(\mathbf{z}^T \boldsymbol{\beta}^*(\tau)|\mathbf{z}) \leq \bar{g}.$$

(b) There exist constants $\kappa > 0$ and A such that, when $|t| \leq \kappa$,

$$\sup_{\mathbf{z}, \tau \in [\nu, \tau_U]} \left| f(\mathbf{z}^T \boldsymbol{\beta}^*(\tau) + t|\mathbf{z}) - f(\mathbf{z}^T \boldsymbol{\beta}^*(\tau)|\mathbf{z}) \right| \leq A|t|,$$

$$\sup_{\mathbf{z}, \tau \in [\nu, \tau_U]} \left| g(\mathbf{z}^T \boldsymbol{\beta}^*(\tau) + t|\mathbf{z}) - g(\mathbf{z}^T \boldsymbol{\beta}^*(\tau)|\mathbf{z}) \right| \leq A|t|.$$

(A4) (*Sparsity*) Assume $\log p = o(n^{1/2})$, and let

$$S_\tau = \{j : \beta_j^*(\tau) \neq 0\}, \quad S^* = \bigcup_{\tau \in [\nu, \tau_U]}$$

$$S_\tau = \left\{ j : \sup_{\tau \in [\nu, \tau_U]} |\beta_j^*(\tau)| > 0 \right\}, \quad \text{and} \quad q = |S^*|.$$

Let $\widehat{S}$ be the index set of covariates selected by $\mathcal{S}$ with a tuning parameter λ_n. There exist constants $0 \leq c_1 < 1/3$, $c_2, K_1, K_2 > 0$ such that $q \leq K_1 n^{c_1}$, $|\widehat{S}| \leq K_1 n^{c_1}$, and

$$P\left(S^* \subseteq \widehat{S}\right) \geq 1 - K_2 (p \vee n)^{-1-c_2}.$$

(A5) Let $\tilde{\mu}(\tau) = E\left[1 \left\{ \log X > \mathbf{Z}^T \boldsymbol{\beta}^*(\tau) \right\} \right]$. There exists a constant $L > 0$ such that $|\beta_j^*(\tau_1) - \beta_j^*(\tau_2)| \leq L|\tau_1 - \tau_2|$ and $|\tilde{\mu}(\tau_1) - \tilde{\mu}(\tau_2)| \leq L|\tau_1 - \tau_2|$, for all $\tau_1, \tau_2 \in (\nu, \tau_U]$ and $1 \leq j \leq p$.

(A6) (*Bounded eigenvalues*) $\boldsymbol{\delta}^T E[\mathbf{Z}_i \mathbf{Z}_i^T] \boldsymbol{\delta} / \|\boldsymbol{\delta}\|^2$ is bounded below and above by $\lambda_{\min}$ and $\lambda_{\max}$, respectively, over $\|\boldsymbol{\delta}\|_0 \leq K_1 n^{c_1}, \boldsymbol{\delta} \neq \mathbf{0}$, where $0 < \lambda_{\min} < \lambda_{\max}$; (*Nonlinear impact*) $c_2 := \inf_{\|\boldsymbol{\delta}\|_0 \leq K_1 n^{c_1}, \boldsymbol{\delta} \neq \mathbf{0}} E[(\mathbf{Z}_i^T \boldsymbol{\delta})^2]^{3/2} / E[|\mathbf{Z}_i^T \boldsymbol{\delta}|^3] > 0$.

(A7) Γ_m is equally gridded with $\tau_k - \tau_{k-1} = \epsilon_n = c_0 n^{-1}$ for $\tau_k \in \Gamma_m$ $(k = 1, \ldots, m)$ and a constant $c_0 > 0$.

Assumption (A1) requires the number of censored observations below the ν-th quantile not to exceed $cn^{1/2}$, which is satisfied if the lower bound of C's support is greater than the lower bound of T's support, a reasonable scenario in real applications. As recommended in Zheng, Peng, and He (2018), ν is chosen such that only a small proportion of the observed survival times below the νth quantile are censored. (A2) assumes that the covariates are uniformly bounded. As pointed out by Zheng, Peng, and He (2015), the global linear quantile regression model is most meaningful when the covariates are confined to a compact set to avoid crossing of the quantile functions. (A3) ensures the positiveness of $f(t|\mathbf{Z})$ between $\mathbf{Z}^T \boldsymbol{\beta}^*(\nu)$ and $\mathbf{Z}^T \boldsymbol{\beta}^*(\tau_U)$, which is essential for the identifiability of $\boldsymbol{\beta}^*(\tau)$ for $\tau < \tau_U$. (A4) restricts the order of data dimensions, as well as the sparsity of $\boldsymbol{\beta}^*(\tau)$, which is necessary for the convergence of the low-dimensional estimator in Equation (2) (Condition C4 in Wang, Wu, and Li (2012)). (A4) also characterizes the "sure screening" property by $\mathcal{S}$. This asymptotic property does not assess the variability of selection with a finite sample; it is crucial to account for such variability for high-dimensional inference (Fei et al. 2019; Fei and Li 2021). Also, several variable selection methods for high-dimensional CQR satisfy the sure screening property in (A4) with additional mild conditions.

- L-HDCQR: by Corollary 4.1 of Zheng, Peng, and He (2018), a *beta-min* condition is required in addition to the set of conditions in this article. Explicitly, there exist constants $C_1, C_2 > 0$, such that

$$\inf_{j \in S^*} \sup_{\tau \in [\tau_L, \tau_U]} |\beta_j^*(\tau)| > C_1 \exp(C_2 q \tau_U) \sqrt{q \log(p \vee n)/n} + L\sqrt{q} \epsilon_n.$$

- AL-HDCQR: by Corollary 4.2 of Zheng, Peng, and He (2018), AL-HDCQR achieves the stronger *selection consistency* property, which implies the sure screening property.
- Quantile-adaptive Screening: by Theorem 3.3 of He, Wang, and Hong (2013), with a proper threshold value in their technical conditions, the screening procedure achieves the sure screening property.

(A5) characterizes the smoothness of $\boldsymbol{\beta}^*(\tau)$. The eigenvalue condition in (A6) is the sparse Riesz condition in Zhang and Huang (2008), satisfied by many commonly used covariance structures, including the compound symmetry structure and the first-order autoregressive structure (AR(1)) (Zhang and Huang 2008). Also, the nonlinear impact condition controls the minoration of the quantile regression objective function by a quadratic function, as adopted in Zheng, Peng, and He (2018), for establishing the consistency of L-HDCQR estimator. The condition is satisfied when the covariates $\mathbf{Z_i}$ have a log-concave density, which includes the commonly used normal distribution, Wishart distribution and Dirichlet distribution (Lovász and Vempala 2007). (A7) details the fineness of Γ_m, which renders an adequate approximation to the stochastic integration in Equation (2).

3.2. Theoretical Properties of Fused-HDCQR

We first extend the results in Peng and Huang (2008) from a fixed p to a *p-diverges-but-less-than-n* case. The results are novel and critical since we allow the true model size $q = |S^*|$ to increase with n, while the selected $\widehat{S}^b$'s in the fusing procedure vary around S^*. Specifically, we assume a subset $S \subset \{1, \ldots, p\}$ in Theorems 1 and 2, where $|S| \leq K_1 n^{c_1}$, $0 \leq c_1 < 1/3$ and $K_1 > 0$. Let $\grave{\boldsymbol{\beta}}_S(\tau), \tau \in [\nu, \tau_U]$ be the estimator from Peng and Huang (2008) of fitting the CQR with $\mathbf{Z}_S$, the covariates indexed by S, over the τ-grid Γ_m.

Theorem 1. (Consistency with a diverging number of covariates) Under Conditions (A1)–(A7) and given a subset $S \subset \{1, \cdots, p\}$ such that $S^* \subseteq S$ and $|S| \leq K_1 n^{c_1}$, there exist positive constants ζ_1 and ζ_2 such that

$$\sup_{\nu \leq \tau \leq \tau_U} \|\grave{\boldsymbol{\beta}}_S(\tau) - \boldsymbol{\beta}^*(\tau)\| \leq \zeta_1 \exp(\zeta_2)(K_1 n^{c_1-1} \log n)^{1/2}$$

with probability at least $1 - 20 c_0^{-2} K_1 n^{c_1-2}$.

Remark 5. From the proof of this theorem (in particular, the proofs of Propositions 1 and 2 in the supplementary materials that lead to this theorem), it can be seen that ζ_1 and ζ_2 do not depend on the choice of S or n. Thus, ζ_1 and ζ_2 are universal for all possible S satisfying $S^* \subseteq S$ and $|S| \leq K_1 n^{c_1}$.

Next, we derive the weak convergence of $\grave{\beta}_j$ for any $j \in S$.

Theorem 2. (Weak convergence with a diverging number of covariates) Under Conditions (A1)–(A7) and given a $S \subset \{1, \cdots, p\}$ such that $S^* \subseteq S$ and $|S| \le K_1 n^{c_1}$, it holds that

$$\sqrt{n}\left(\acute{\beta}_j(\tau) - \beta_j^*(\tau)\right)$$

converges weakly to a mean zero Gaussian process for $\tau \in [\nu, \tau_U]$ and any $j \in S$.

In high-dimensional settings, the next theorem shows that the fused estimator enjoys desirable theoretical properties.

Theorem 3. Consider the Fused-HDCQR estimator in Equation (3). Under Assumptions (A1)–(A7), for any $j \in \{1, \ldots, p\}$,

$$\sqrt{n}\left(\widehat{\beta}_j(\tau) - \beta_j^*(\tau)\right)$$

converges weakly to a mean zero Gaussian process for $\tau \in [\nu, \tau_U]$.

Our framework enables us to obtain the joint distribution of any K-dimensional estimated coefficients, where K is a finite number. Let $\mathcal{K}$ be the collection of the indices of K covariates of interest. We can show that the weak convergence result of $\acute{\boldsymbol{\beta}}_{\mathcal{K}}(\tau)$, a K-dimensional subvector of the oracle estimator, still holds for $\tau \in [\nu, \tau_U]$, that is, $\sqrt{n}(\acute{\boldsymbol{\beta}}_{\mathcal{K}}(\tau) - \boldsymbol{\beta}_{\mathcal{K}}^*(\tau)), \tau \in [\nu, \tau_U]$ converges to a K-dimensional Gaussian distribution at any $\tau \in [\nu, \tau_U]$. We only need to replace $\acute{\boldsymbol{\beta}}_j(\tau)$ by $\acute{\boldsymbol{\beta}}_{\mathcal{K}}(\tau)$ in the proof of Theorem 2 in the supplementary materials and slightly modify the arguments accordingly. Consequently, term I in the proof of Theorem 3 still converges weakly to a mean zero Gaussian distribution, while the norms of terms II and III are still $o_p(1)$. Therefore, Theorem 3 still holds for any K-dimensional subvector of $\hat{\boldsymbol{\beta}}_{\mathcal{K}}(\tau)$, that is, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{K}}(\tau) - \boldsymbol{\beta}_{\mathcal{K}}^*(\tau))$ converges to a mean zero K-dimensional Gaussian distribution at any $\tau \in [\nu, \tau_U]$.

As shown in the proof, the covariance function of $\widehat{\beta}_j(\tau)$ depends on the unknown active set S^*, the unknown conditional density functions $f(t|\mathbf{Z})$ and $g(t|\mathbf{Z})$, and other unknown quantities. Thus, it is not calculable. The next section proposes an alternative model-free variance estimator based on the functional delta method and the multi-sample splitting properties (Efron 2014; Fei and Li 2021).

4. A Variance Estimator via the Functional Delta Method

Let $J_{bi} \in \{0, 1\}$ indicate whether the ith observation is in the bth sub-sample D_2^b, and $J_{\cdot i} = B^{-1}\sum_{b=1}^B J_{bi}$. For each $i = 1, \ldots, n$, we define the resampling covariance between J_{bi} and $\widetilde{\beta}_j^b(\tau_k)$ at $\tau_k \in \Gamma_m$ as

$$\mathbf{s}_{ij}(\tau_k) = \frac{1}{B}\sum_{b=1}^B (J_{bi} - J_{\cdot i})\left(\widetilde{\beta}_j^b(\tau_k) - \widehat{\beta}_j(\tau_k)\right).$$

Define $\mathbf{S}_j(\tau_k) = \left(\mathbf{s}_{1j}(\tau_k), \mathbf{s}_{2j}(\tau_k), \ldots, \mathbf{s}_{nj}(\tau_k)\right)^{\mathrm{T}}$ and let $n_1 = |D_2^b|$. It follows that the covariance between $\widehat{\beta}_j(\tau_k)$ and $\widehat{\beta}_j(\tau_\ell)$

can be consistently estimated by

$$\widehat{\mathrm{Cov}}_j(\tau_k, \tau_\ell) = \frac{n-1}{n}\left(\frac{n}{n-n_1}\right)^2 \sum_{i=1}^n \mathbf{s}_{ij}(\tau_k)\mathbf{s}_{ij}(\tau_\ell)$$

$$= \frac{n(n-1)}{(n-n_1)^2}\mathbf{S}_j^{\mathrm{T}}(\tau_k)\mathbf{S}_j(\tau_\ell),$$

where the multiplier $n(n-1)/(n-n_1)^2$ is a finite-sample correction for sub-sampling (Wager and Athey 2018). In particular, by taking $\tau_\ell = \tau_k$, a variance estimator for $\widehat{\beta}_j(\tau_k)$ is

$$\widehat{V}_j(\tau_k) = \frac{n(n-1)}{(n-n_1)^2}\mathbf{S}_j^{\mathrm{T}}(\tau_k)\mathbf{S}_j(\tau_k). \quad (5)$$

As in Wager and Athey (2018), it follows that $\widehat{V}_j(\tau_k)/\mathrm{Var}\left(\widehat{\beta}_j(\tau_k)\right) \xrightarrow{p} 1$ with $n, B \to \infty$. Furthermore, for a finite B, we propose a bias corrected version of Equation (5)

$$\widehat{V}_j^B(\tau_k) = \widehat{V}_j(\tau_k) - \frac{nn_1}{B(n-n_1)}\left\{B^{-1}\sum_{b=1}^B\left(\widetilde{\beta}_j^b(\tau_k) - \widehat{\beta}_j(\tau_k)\right)^2\right\}. \quad (6)$$

The correction term in Equation (6) is a suitable multiplier of the re-sampling variance of $\widetilde{\beta}_j^b(\tau_k)$'s, and converges to zero with $n \to \infty$ and $n_1 = O(n)$. Thus, the two variance estimators in Equations (5) and (6) are asymptotically equal. However, $\widehat{V}_j(\tau_k)$ in (5) requires B to be of order $n^{3/2}$ to reduce the Monte Carlo noise below the sampling noise, while $\widehat{V}_j^B(\tau_k)$ in (6) only requires B to be of order n to achieve the same (Wager, Hastie, and Efron 2014).

Since $\widehat{\beta}_j(\tau)$ converges weakly to a Gaussian process by Theorem 3, and our variance estimators are consistent on the grid points, we define an asymptotic $100(1-\alpha)\%$ point-wise confidence interval for $\beta_j^*(\tau_k)$ at any $\tau_k \in \Gamma_m$ as

$$\left(\widehat{\beta}_j(\tau_k) - \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_j^B(\tau_k)}\,,\right.$$

$$\left.\widehat{\beta}_j(\tau_k) + \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_j^B(\tau_k)}\right),$$

where $\widehat{V}_j^B(\tau_k)$ is as defined in Equation (6), and Φ is the standard normal cumulative distribution function. The p-value of testing $H_0 : \beta_j^*(\tau_k) = 0$ for a $\tau_k \in \Gamma_m$ is

$$2 \times \left\{1 - \Phi\left(|\widehat{\beta}_j(\tau_k)|/\sqrt{\widehat{V}_j^B(\tau_k)}\right)\right\}.$$

5. Simulation Studies

In various settings, we compare the proposed method, Fused-HDCQR (referred to as "Fused" in the tables and figures hereafter), with some competing methods in quantile regression or high-dimensional inference. These methods include Wang, Wu, and Li (2012) ("W12") and Fan, Fan, and Barut (2014) ("F14") for quantile regression; Zheng, Peng, and He (2018) ("Z18") for CQR; and Meinshausen, Meier, and Bühlmann (2009) ("M09") for inference with aggregated p-values from multi-sample splittings.

In the simulations and the later data analysis, we choose L-HDCQR described in Section 3 as the variable selection tool for

Fused-HDCQR. We also explore the feasibility of using other alternatives for variable selection, such as Fan, Samworth, and Wu (2009) ("F09") and M09.

When implementing Fused-HDCQR, we specify the number of splits as $B = 300$, the quantile interval as $[\nu, \tau_U] = [0.1, 0.8]$, and the grid length as $m = n/\log p$. As regards to the selection of tuning parameters, Theorems 1 and 2 suggest that our procedure not be sensitive to tuning parameters as long as they can ensure sure screening. In practical settings, we recommend to select tuning parameters by minimizing the 5-fold cross-validation error as in Equation (4), which may help achieve sure screening and works well in our simulations. We study the following examples with sparse nonzero effects, some of which are heterogeneous.

Example 1. The event times are generated by

$$\log T_i = \widetilde{\mathbf{Z}}_i^{\mathrm{T}} \mathbf{b} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the coefficient vector $\mathbf{b}$ is sparse with $b_{20} = 0.5, b_{40} = 1, b_{60} = 1.5, b_j = 0$ for all other j's, and $\varepsilon_i \sim N(0, 1)$. Therefore, the true coefficients satisfy $\boldsymbol{\beta}^*(\tau) = (Q_\varepsilon(\tau), \mathbf{b}^{\mathrm{T}})^{\mathrm{T}}$ for all $\tau \in (0, 1)$, where $Q_\varepsilon(\tau)$, the τth quantile of the distribution of ε, is the intercept. The covariates $\widetilde{Z}_{j,i}$'s are iid from $\mathrm{Unif}(-1, 1)$ and are independent across $j \in \{1, \ldots, p\}$. The censoring times are generated independently as $\log C_i \sim N(3, 17.25)$, giving a censoring rate around 25%.

Example 2. The event times are generated as follows:

$$\log T_i = \widetilde{\mathbf{Z}}_i^{\mathrm{T}} \mathbf{b} + 1.5\widetilde{Z}_{3,i}\varepsilon_i, \tag{7}$$

where $b_{20} = 1, b_{40} = 1.5, b_{60} = 2, b_j = 0$ for all other j's, and $\varepsilon_i \sim N(0, 1)$. We first generate $\acute{\mathbf{Z}}_i \sim N_p(0, \Sigma)$ with an AR(1) $\Sigma = (\sigma_{k\ell})_{p \times p}$, where $\sigma_{k\ell} = 0.3^{|k-\ell|}$, and then let $\widetilde{\mathbf{Z}}_i = \acute{\mathbf{Z}}_i$, except that the third covariate $\widetilde{Z}_{3,i} = |\acute{Z}_{3,i}| + 0.5$. Thus, $\beta_1^*(\tau) = 0, \beta_4^*(\tau) = 1.5Q_\varepsilon(\tau)$, and $\beta_{j+1}^*(\tau) = b_j$, for all other j's. The censoring times are generated independently as $\log C_i \sim N(4, 17.25)$, giving a censoring rate around 23%.

Example 3. The event times are generated as follows:

$$\log T_i = \widetilde{\mathbf{Z}}_i^{\mathrm{T}} \mathbf{b} + \phi_1(\xi_i)\widetilde{Z}_{1,i} + \phi_4(\xi_i)\widetilde{Z}_{4,i},$$

where $b_{20} = 1, b_{40} = 1.5, b_{60} = 2, b_j = 0$ for all other j's, $\xi_i \sim N(0, 1)$, and ϕ_1, ϕ_4 are monotone functions as the dashed lines in Figure 1, both of which are continuous with zero and nonzero segments over τ. We first generate $\acute{\mathbf{Z}}_i \sim N_p(0, \Sigma)$ as in Example 2, and then let $\widetilde{\mathbf{Z}}_i = \acute{\mathbf{Z}}_i$, except that $\widetilde{Z}_{1,i} = |\acute{Z}_{1,i}| + 0.5$ and $\widetilde{Z}_{4,i} = |\acute{Z}_{4,i}| + 0.5$. Therefore, $\beta_1^*(\tau) = 0, \beta_2^*(\tau) = \phi_1(\tau), \beta_5^*(\tau) = \phi_4(\tau)$, and $\beta_{j+1}^*(\tau) = b_j$, for all other j's. The censoring times are generated independently as $\log C_i \sim N(6, 17.25)$, which gives a censoring rate around 20%.

For each of these examples, we set $(n, p) = (300, 1000)$ and $(700, 1000)$ to study the impacts of the sample size and the number of variables on the performance, and, in particular, how the methods fare when $p > n$. In Example 3, which mimics the real data example in Section 6 most closely, we have also explored $(n, p) = (700, 2000)$, which is roughly equal to the dimension of the real dataset. For every parameter configuration, a total of 100

independent datasets are generated, and we report the average results based on these replications. We choose 100 replications because the penalized methods for high-dimensional CQR in general take much computing time (Table 5).

We first evaluate the feasibility of using various variable selection tools for our proposed method. Comparisons of true positives and false negatives among F09, M09, and L-HDCQR under Examples 1–3 are reported in Table 1. F09 presents a subpar performance because, by taking intersections of variables selected from different partitions of data, it tends to miss out some true signals and thus have fewer true positives. In contrast, L-HDCQR retains more true positives than both F09 and M09, while having more false positives. Because our method requires the variable selection step to include the true signals with high probability, even at the cost of some false positives, we use L-HDCQR as the screening tool for our method.

We next compare Fused-HDCQR with the other high dimensional quantile regression methods under Example 1, with results reported at $\tau = .25, .5$ and $.75$. As a benchmark for comparisons, we also compute the oracle estimates based on the true model (with S^* known). Since W12, F14, and Z18 only provide coefficient estimates without standard errors (SEs), we only report the estimation biases for them, while reporting the average SEs, empirical standard deviations (SDs) and coverage probabilities of the confidence intervals for our method. Table 2 shows that Fused-HDCQR presents the smallest biases, which are comparable to those of the oracle estimates. In contrast, Z18 has smaller biases when the sample size is large, and larger biases otherwise, while W12 and F14 incur substantial biases since they are not designed for censored data. Moreover, the average SEs based on Fused-HDCQR agree with the empirical SDs of the estimates. The consistent estimates of coefficients and SEs obtained by Fused-HDCQR lead to proper coverage probabilities around the 0.95 nominal level. In addition, the coverage probabilities become closer to 0.95 as n increases.

Table 2 also concerns the power for detection of signals. Since W12, F14, and Z18 cannot draw inferences and, in general, there is a lack of literature that deals with inference for HDCQR, we compare our method with the aggregated p-value approach (M09) in the quantile setting, though M09 originated from linear regression. The results indicate that Fused-HDCQR outperforms M09, presenting more power when the effect size is moderate or large.

Table 3 summarizes the results from Example 2 with the heterogeneous effect β_4 varying with τ. We compare the estimation accuracy between Fused-HDCQR and Z18, as well as the statistical power between Fused-HDCQR and M09. Again, Fused-HDCQR presents smaller biases than Z18 and a higher power than M09. To assess whether the tuning parameters selected as in Remark 3 help the variable selection method (L-HDCQR), used by Fused-HDCQR, satisfy Assumption (A4) in Section 3, we report the selection frequency of each signal variable in Table 3 (and also in Table 4), and observe that the selection frequency increases as the sample size increases, hinting that Assumption (A4) may be satisfied with these selected tuning parameters.

Table 4 summarizes the results based on Example 3. For the two heterogeneous effects β_2 and β_5 that vary with τ, the estimation biases of Fused-HDCQR become smaller and the
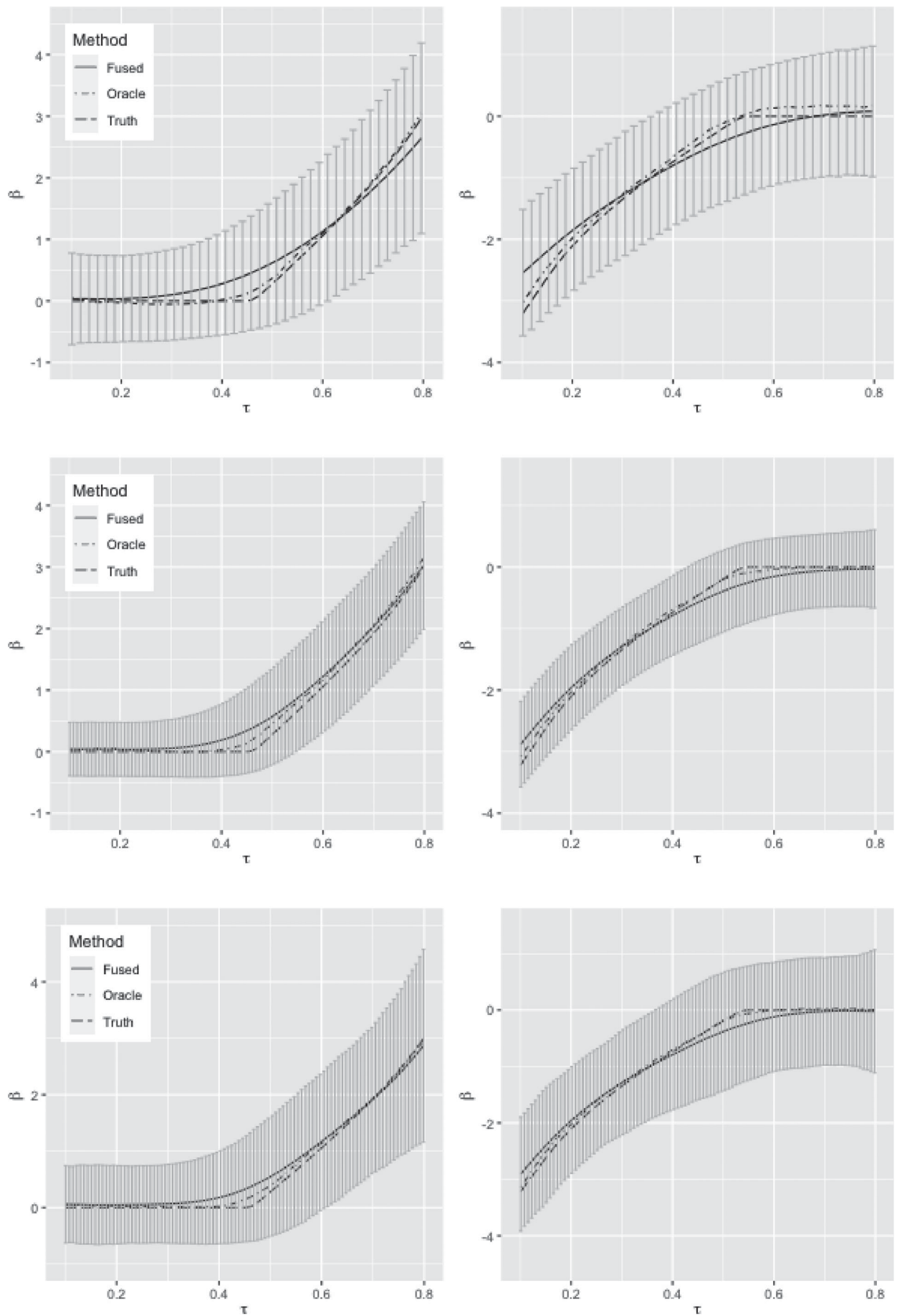
Figure 1. Estimated heterogeneous effects and confidence intervals of Fused-HDCQR using Example 3: $\beta_2^*(\cdot)$ (left panel) and $\beta_5^*(\cdot)$ (right panel). From the top to the bottom are the plots for $(n, p) = (300, 1000), (700, 1000),$ and $(700, 2000)$, respectively.

Table 1. Summary of variable selection results based on the simulated datasets.

				TP			FP		
	(n, p)	CR	q	L-HDCQR	M09	F09	L-HDCQR	M09	F09
Example 1	(300,1000)	0.25	3	2.67	2.12	1.64	7.95	0.00	0.19
	(700,1000)	0.25	3	2.98	2.78	2.27	13.08	0.01	0.34
Example 2	(300,1000)	0.22	4	3.60	3.58	2.22	12.45	0.00	0.22
	(700,1000)	0.23	4	3.99	3.99	3.54	11.29	0.00	0.64
Example 3	(300,1000)	0.20	5	3.82	3.63	1.91	10.00	0.00	0.17
	(700,1000)	0.20	5	4.81	4.77	4.35	11.73	0.01	0.54
	(700,2000)	0.19	5	4.78	4.76	4.17	16.34	0.00	0.47

NOTE: CR, average censoring rate; $q = |S^*|$; TP, average true positives; FP, average false positives; M09, Meinshausen, Meier, and Bühlmann (2009); F09, Fan, Samworth, and Wu (2009); L-HDCQR, Zheng, Peng, and He (2018).

Table 2. Results of Example 1 based on the simulated datasets.

		Bias				EmpSD	SE	Cov	Power	
	Oracle	Fused	Z18	F14	W12		Fused		Fused	M09
					$n = 300, p = 1000$					
$\beta_{21} = 0.5$	0.02	0.02	−0.38	−0.50	−0.48	0.14	0.13	0.93	0.97	0.06
	0.02	0.01	−0.24	−0.49	−0.48	0.12	0.13	0.95	0.98	0.04
	0.01	0.01	−0.13	−0.50	−0.48	0.12	0.13	0.96	1.00	0.02
$\beta_{41} = 1$	−0.01	−0.01	−0.02	−0.91	−0.33	0.14	0.13	0.92	1.00	0.99
	−0.00	−0.00	−0.03	−0.68	−0.32	0.14	0.12	0.92	1.00	0.98
	0.02	0.01	−0.01	−0.70	−0.30	0.17	0.14	0.93	1.00	0.94
$\beta_{61} = 1.5$	−0.00	0.01	0.00	−0.92	−0.24	0.12	0.13	0.92	1.00	1.00
	0.00	0.01	0.01	−0.64	−0.25	0.11	0.13	0.97	1.00	1.00
	0.02	0.01	0.02	−0.70	−0.25	0.13	0.14	0.95	1.00	1.00
					$n = 700, p = 1000$					
$\beta_{21} = 0.5$	−0.02	−0.01	−0.01	−0.47	−0.23	0.09	0.08	0.92	1.00	0.56
	−0.01	−0.01	−0.01	−0.39	−0.22	0.08	0.08	0.89	1.00	0.65
	−0.01	−0.01	−0.01	−0.40	−0.23	0.10	0.09	0.89	1.00	0.44
$\beta_{41} = 1$	0.00	0.00	0.04	−0.53	−0.17	0.09	0.08	0.91	1.00	1.00
	−0.00	0.00	0.03	−0.49	−0.19	0.09	0.08	0.90	1.00	1.00
	−0.01	−0.01	0.01	−0.53	−0.18	0.08	0.10	0.87	1.00	1.00
$\beta_{61} = 1.5$	0.01	0.01	0.06	−0.54	−0.21	0.10	0.09	0.93	1.00	1.00
	0.01	0.01	0.03	−0.62	−0.21	0.08	0.08	0.93	1.00	1.00
	−0.00	0.00	0.03	−0.71	−0.21	0.07	0.09	0.94	1.00	1.00

NOTE: Each β has three rows corresponding to $\tau = .25, .5, .75$ from the top to the bottom; EmpSD, empirical standard deviation; SE, average standard error; Cov, coverage probability; Oracle, Oracle estimator; Z18, Zheng, Peng, and He (2018).F14, Fan, Fan, and Barut (2014); W12, Wang, Wu, and Li (2012); M09, Meinshausen, Meier, and Bühlmann (2009).

Table 3. Results of Example 2 based on the simulated datasets.

		Bias		EmpSD	SE	Cov	Freq	Power	
	Oracle	Fused	Z18		Fused			Fused	M09
				$n = 300, p = 1000$					
$\beta_4 = 1.5Q_\varepsilon(\tau)$	0.01	0.13	0.29	0.32	0.31	0.88	0.73	0.82	0.16
	−0.05	−0.07	0.06	0.33	0.29	0.90		0.11	0.00
	0.01	−0.14	−0.05	0.31	0.34	0.82		0.62	0.10
$\beta_{21} = 1$	−0.01	−0.01	−0.01	0.14	0.13	0.90	0.69	1.00	0.88
	−0.03	−0.01	−0.05	0.12	0.12	0.91		1.00	0.92
	−0.01	−0.00	−0.02	0.14	0.13	0.92		1.00	0.84
$\beta_{41} = 1.5$	0.01	0.01	0.03	0.13	0.13	0.90	0.99	1.00	1.00
	−0.01	0.01	0.03	0.12	0.13	0.93		1.00	1.00
	−0.00	0.02	−0.02	0.13	0.14	0.93		1.00	1.00
$\beta_{61} = 2$	−0.03	−0.03	0.04	0.13	0.13	0.91	1.00	1.00	1.00
	−0.03	−0.02	0.03	0.11	0.13	0.92		1.00	1.00
	−0.01	−0.01	−0.00	0.12	0.15	0.95		1.00	1.00
				$n = 700, p = 1000$					
$\beta_4 = 1.5Q_\varepsilon(\tau)$	0.03	0.08	0.19	0.19	0.21	0.92	0.89	0.99	0.61
	0.02	0.03	0.14	0.18	0.19	0.89		0.11	0.00
	0.04	−0.03	−0.01	0.21	0.23	0.92		0.97	0.56
$\beta_{21} = 1$	0.01	0.01	0.05	0.09	0.08	0.94	0.99	1.00	1.00
	0.01	0.01	0.01	0.08	0.08	0.87		1.00	1.00
	0.01	0.01	0.05	0.10	0.09	0.89		1.00	1.00
$\beta_{41} = 1.5$	−0.01	0.00	0.08	0.08	0.08	0.94	1.00	1.00	1.00
	−0.00	0.00	0.05	0.09	0.08	0.92		1.00	1.00
	0.00	0.01	0.04	0.09	0.09	0.95		1.00	1.00
$\beta_{61} = 2$	−0.01	−0.01	0.10	0.08	0.09	0.93	1.00	1.00	1.00
	−0.01	−0.01	0.06	0.08	0.09	0.91		1.00	1.00
	−0.00	−0.00	0.07	0.09	0.10	0.90		1.00	1.00

NOTE: See the footnote of Table 2; Freq, average selection frequency in B splits.

Table 4. Results of Example 3 based on the simulated datasets.

	Bias			EmpSD	SE	Cov	Freq	Power	
	Oracle	Fused	Z18		Fused			Fused	M09
				$n = 300, p = 1000$					
$\beta_2 = \phi_1(\tau)$	0.08	0.06	0.59	0.34	0.36	0.94	0.71	0.06	0.00
	0.34	0.37	1.01	0.52	0.51	0.89		0.20	0.00
	0.08	−0.20	−0.05	0.80	0.72	0.89		0.87	0.06
$\beta_5 = \phi_4(\tau)$	0.10	0.14	0.27	0.65	0.50	0.90	0.67	0.77	0.36
	−0.16	−0.20	−0.36	0.62	0.51	0.91		0.19	0.00
	0.02	0.06	−0.03	0.56	0.52	0.90		0.10	0.00
$\beta_{21} = 1.5$	0.02	0.03	0.04	0.25	0.23	0.95	0.65	1.00	0.77
$\beta_{41} = 2$	0.01	−0.00	0.02	0.23	0.25	0.93	0.93	1.00	0.99
$\beta_{61} = 2.5$	−0.02	0.07	0.19	0.21	0.26	0.94	0.99	1.00	1.00
				$n = 700, p = 1000$					
$\beta_2 = \phi_1(\tau)$	0.01	0.04	0.27	0.21	0.23	0.94	0.96	0.06	0.00
	0.13	0.30	0.79	0.37	0.40	0.88		0.27	0.01
	0.00	0.08	0.35	0.51	0.51	0.90		1.00	0.77
$\beta_5 = \phi_4(\tau)$	0.06	0.09	0.18	0.33	0.33	0.91	0.92	0.99	0.92
	−0.09	−0.19	−0.23	0.35	0.34	0.85		0.21	0.00
	−0.01	−0.04	−0.08	0.37	0.31	0.94		0.06	0.00
$\beta_{21} = 1.5$	−0.00	0.00	0.04	0.16	0.17	0.97	0.98	1.00	1.00
$\beta_{41} = 2$	0.01	−0.02	−0.01	0.15	0.18	0.95	1.00	1.00	1.00
$\beta_{61} = 2.5$	0.01	0.00	0.07	0.18	0.18	0.94	1.00	1.00	1.00
				$n = 700, p = 2000$					
$\beta_2 = \phi_1(\tau)$	−0.01	0.05	0.13	0.32	0.32	0.93	0.93	0.07	0.00
	0.10	0.26	0.59	0.46	0.44	0.91		0.09	0.02
	0.05	−0.07	0.15	0.53	0.46	0.87		0.74	0.58
$\beta_5 = \phi_4(\tau)$	0.10	0.10	0.25	0.45	0.35	0.84	0.90	1.00	0.83
	−0.03	−0.18	−0.31	0.41	0.36	0.89		0.76	0.01
	−0.00	−0.01	−0.13	0.36	0.34	0.85		0.15	0.00
$\beta_{21} = 1.5$	0.01	0.01	0.03	0.18	0.21	0.98	0.98	1.00	1.00
$\beta_{41} = 2$	0.01	0.02	−0.07	0.22	0.20	0.91	0.99	1.00	0.98
$\beta_{61} = 2.5$	−0.01	−0.01	−0.05	0.25	0.20	0.94	1.00	1.00	0.98

NOTE: See the footnote of Tables 2 and 3; For β_2 and β_5, the numbers are shown at $\tau = .25, .5, .75$ from the top to the bottom and, for the other β's, at $\tau = 0.5$.

Table 5. Comparisons of computing time (on average per dataset in seconds) when performing Example 1.

	Fused	Z18	W12	F14	M09
$(n, p) = (300, 1000)$	888	853	509	390	170
$(n, p) = (700, 1000)$	3,108	1,812	2,230	1,231	440

NOTE: See the footnote of Table 2.

Table 6. Patients' characteristics in the BLCSC samples. ($n = 674$)

		Mean (SD)
Age		60 (10.8)
		Count (%)
Female		259 (38)
Education level	≤ High school	264 (39)
	> High school	410 (61)
Smoking	Nonactive	418 (62)
	Active	256 (38)
Cancer type	Adenocarcinoma	283 (42)
	Squamous cell	110 (16)
	Other	281 (42)
Cancer stage	1	283 (42)
	2	110 (16)
	3	256 (38)
	4	25 (4)

estimated SEs are closer to the empirical ones as n increases. Figure 1 shows that the Fused-HDCQR estimates in general agree with the oracle estimates and the truth, except at the non-smooth change points, and have narrower confidence intervals with a larger n, where the vertical bars are the average confidence intervals of the τ grid points.

In regards to the choice of B in the variance computation, our numerical experience suggests that it may be sufficient to use a B that is of the same order of the sample size, even when n is less than p. This coincides with the note under (6) that B is only required to be of order n to reduce the Monte Carlo noise below the sampling noise.

Finally, we compare the computation intensity among Z18, M09, W12, F14, and Fused-HDCQR under Example 1 and report in Table 5 the computing time on average per dataset. Our method is the most computationally intensive, because it involves multiple data-splittings and draws inferences on all of the p coefficients. However, by utilizing parallel computing, we have managed to reduce the computational time to the same order of Z18, W12, and F14 that are based on penalized regression. The R code used for generating the simulation results can be accessed via *https://github.com/feizhe/HDCQR_Paper*.

6. Application to the Boston Lung Cancer Survival Cohort (BLCSC)

Detection of molecular profiles related to cancer survival can aid personalized treatment in prolonging patients' survival and improving their quality of life. In a subset of BLCSC samples, 674 lung cancer patients were measured with survival times, along with 40,000 SNPs and clinical indicators, such as lung cancer subtypes (adenocarcinoma, squamous cell carcinoma, or others), cancer stages (1–4), age, gender, education level (≤ high school or > high school), and smoking status (active or non-active smokers); see Table 6 for the patients' characteristics. The censoring rate was 23% and a total of 518 deaths were observed
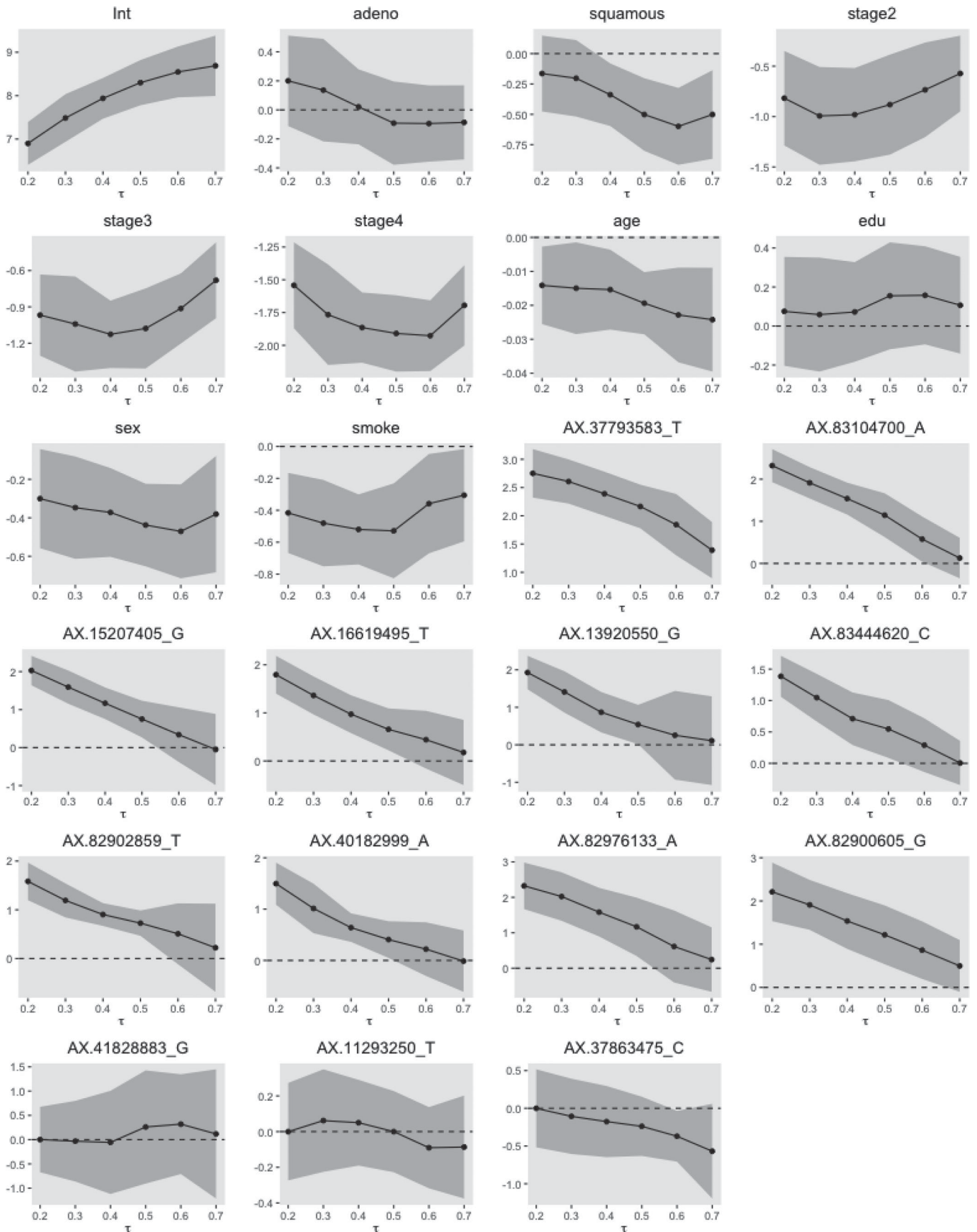
Figure 2. Estimated quantile-specific coefficients of the predictors in Table 7.

Table 7. Analysis of the BLCSC data with Fused-HDCQR. The SNPs are sorted by their *p*-values at τ = 0.2, corresponding to the high-risk groups. Results for the top 10 and the bottom 3 are presented.

τ	Estimator 0.2	SE	p-value	Estimator 0.3	SE	p-value	Estimator 0.4	SE	p-value
Int	6.90	0.25	1.4E-165	7.48	0.28	4.3E-157	7.94	0.24	3.2E-241
Adeno	0.20	0.16	2.1E-01	0.14	0.18	4.5E-01	0.02	0.13	8.7E-01
Squamous	-0.16	0.16	3.0E-01	-0.20	0.16	2.1E-01	-0.34	0.13	1.0E-02
Stage2	-0.82	0.24	6.3E-04	-0.99	0.25	6.0E-05	-0.98	0.24	3.2E-05
Stage3	-0.97	0.17	1.6E-08	-1.04	0.20	2.0E-07	-1.13	0.14	2.0E-15
Stage4	-1.54	0.17	3.0E-20	-1.77	0.20	1.7E-19	-1.86	0.14	2.2E-42
Age	-0.01	0.01	1.5E-02	-0.01	0.01	3.0E-02	-0.02	0.01	1.0E-02
Edu	0.08	0.14	6.0E-01	0.06	0.15	6.9E-01	0.07	0.13	5.8E-01
Female	-0.30	0.13	2.2E-02	-0.35	0.14	1.0E-02	-0.37	0.12	1.6E-03
Smoke	-0.42	0.13	1.1E-03	-0.48	0.14	5.0E-04	-0.52	0.11	3.4E-06
AX.37793583_T	2.75	0.22	3.0E-36	2.61	0.20	4.6E-39	2.39	0.20	3.7E-33
AX.83104700_A	2.32	0.20	4.0E-31	1.91	0.19	6.3E-24	1.54	0.19	1.5E-15
AX.15207405_G	2.03	0.20	1.0E-24	1.59	0.22	9.8E-13	1.17	0.21	3.7E-08
AX.16619495_T	1.79	0.20	3.3E-19	1.36	0.20	1.3E-11	0.97	0.20	1.2E-06
AX.13920550_G	1.93	0.23	2.5E-17	1.41	0.28	5.3E-07	0.87	0.27	1.6E-03
AX.83444620_C	1.39	0.17	7.4E-17	1.05	0.19	6.6E-08	0.71	0.21	8.8E-04
AX.82902859_T	1.58	0.20	8.7E-16	1.19	0.18	2.0E-11	0.90	0.12	3.4E-14
AX.40182999_A	1.50	0.21	9.6E-13	1.01	0.25	3.9E-05	0.64	0.14	6.5E-06
AX.82976133_A	2.32	0.33	3.8E-12	2.02	0.35	6.7E-09	1.58	0.35	6.1E-06
AX.82900605_G	2.21	0.35	1.6E-10	1.91	0.29	9.1E-11	1.54	0.33	2.9E-06
⋮									
AX.41828883_G	1.4E-03	0.34	1.00	-3.2E-02	0.42	0.94	-5.7E-02	0.54	0.92
AX.11293250_T	-3.6E-04	0.14	1.00	6.2E-02	0.15	0.67	5.0E-02	0.12	0.68
AX.37863475_C	-3.1E-04	0.26	1.00	-1.1E-01	0.25	0.68	-1.8E-01	0.24	0.46

τ	Estimator 0.5	SE	p-value	Estimator 0.6	SE	p-value	Estimator 0.7	SE	p-value
Int	8.30	0.27	4.8E-214	8.55	0.30	4.9E-180	8.69	0.35	2.8E-132
Adeno	-0.09	0.15	5.3E-01	-0.09	0.13	4.8E-01	-0.09	0.13	5.1E-01
Squamous	-0.50	0.15	1.0E-03	-0.60	0.16	2.1E-04	-0.50	0.19	7.1E-03
Stage2	-0.88	0.25	5.0E-04	-0.73	0.24	2.1E-03	-0.57	0.19	2.8E-03
Stage3	-1.08	0.17	1.7E-10	-0.91	0.15	6.4E-10	-0.68	0.16	2.0E-05
Stage4	-1.91	0.15	7.0E-38	-1.93	0.14	1.7E-44	-1.69	0.16	2.1E-27
Age	-0.02	0.00	3.3E-05	-0.02	0.01	1.3E-03	-0.02	0.01	1.9E-03
Edu	0.15	0.14	2.7E-01	0.16	0.13	2.2E-01	0.11	0.13	4.0E-01
Female	-0.44	0.11	6.4E-05	-0.47	0.12	1.6E-04	-0.38	0.15	1.3E-02
Smoke	-0.53	0.15	4.9E-04	-0.36	0.16	2.4E-02	-0.31	0.15	3.8E-02
AX.37793583_T	2.16	0.20	4.1E-28	1.84	0.28	2.8E-11	1.39	0.25	4.2E-08
AX.83104700_A	1.15	0.27	1.6E-05	0.58	0.27	3.5E-02	0.13	0.25	6.0E-01
AX.15207405_G	0.75	0.25	2.3E-03	0.34	0.37	3.5E-01	-0.05	0.48	9.2E-01
AX.16619495_T	0.66	0.22	3.1E-03	0.44	0.31	1.5E-01	0.18	0.35	6.1E-01
AX.13920550_G	0.54	0.27	4.3E-02	0.26	0.60	6.7E-01	0.11	0.60	8.6E-01
AX.83444620_C	0.55	0.23	2.0E-02	0.29	0.22	1.8E-01	0.01	0.18	9.7E-01
AX.82902859_T	0.73	0.13	4.2E-08	0.51	0.32	1.1E-01	0.22	0.46	6.3E-01
AX.40182999_A	0.41	0.18	2.6E-02	0.22	0.27	4.1E-01	-0.01	0.30	9.6E-01
AX.82976133_A	1.17	0.42	5.4E-03	0.61	0.52	2.4E-01	0.24	0.46	6.0E-01
AX.82900605_G	1.22	0.35	4.5E-04	0.86	0.34	1.1E-02	0.50	0.31	1.0E-01
⋮									
AX.41828883_G	0.26	0.60	0.66	0.32	0.52	0.54	0.12	0.68	0.86
AX.11293250_T	-0.00	0.12	1.00	-0.09	0.12	0.44	-0.09	0.15	0.56
AX.37863475_C	-0.24	0.20	0.23	-0.37	0.17	0.03	-0.57	0.32	0.08

during the followup period, with the observed followup time varying from 13 to 8,584 days.

We could have included all 40,000 SNPs in our analysis. However, for more statistical power, we opt for the targeted gene approach by focusing on 2,002 SNPs residing in 14 genes identified to be cancer related, namely, ALK, BRAF, BRCA1, EGFR, ERBB2, ERCC1, KRAS, MET, PIK3CA, RET, ROS1, RRM1, TP53, and TYMS (Brose et al. 2002; Toyooka, Tsuda, and Gazdar 2003; Paez et al. 2004; Soda et al. 2007). Pinpointing the effects of individual loci within the targeted genes is helpful for understanding disease mechanisms (Evans et al. 2011; D'Antonio et al. 2019) and designing gene therapies (Pâques and Duchateau 2007; Hanawa et al. 2004). We also adjust for the patients' clinical and environmental characteristics listed in Table 6, which gives a total of $p = 2,011$ predictors.

We apply Fused-HDCQR to compute the point estimates (3) and the variance estimates (6). We set the quantile interval to be $[0.2, 0.7]$, which is wide enough to cover high- and low-risk groups and, in the meantime, ensures the quantile parameters be estimable in the presence of censoring (Zheng, Peng, and He 2015). We choose the lower bound $\tau_0 = \nu = 0.1$ to circumvent the singularity problem with CQR at $\tau = 0$, because few ($< 2\%$) observations are censored below the 0.1th quantile. With $\epsilon_n = 0.01$, we form the τ-grid Γ_m of length $m = 61$. We set $B = 750$ as the number of re-samples, which is sufficiently large and comparable to the sample size. To determine the tuning parameter λ_n in L-HDCQR for selection, we use 5-fold cross-validation as specified in Remark 3.

For ease of presentation, we summarize the results evaluated at 6 quantile levels, $\tau = 0.2, 0.3, \ldots, 0.7$, instead of the whole grid Γ_m. To highlight the findings of the high-risk group, we rank all SNPs based on their p-values at $\tau = 0.2$. In particular, after Bonferroni correction for multiple testing, there are 83 significant SNPs for $\tau = 0.2$ with the overall Type I error of $\alpha = 0.05$. Our method estimates the coefficients and the p-values for *all* predictors, and we only present the results for the patient characteristics, the top 10 significant SNPs, and the 3 least significant SNPs in Figure 2 and Table 7. The estimated coefficient of active smoking drops from -0.42 ($p = 0.0011$) to -0.53 ($p = 0.0005$) as τ changes from 0.2 to 0.5, and then increases to -0.31 ($p = 0.038$) as τ changes to 0.7, suggesting that active smoking might be more harmful to the high- and median-risk groups than the low-risk group of patients. The most significant SNP at $\tau = 0.2$ is AX.37793583_T, which remains significant throughout $\tau = 0.2$ to $\tau = 0.7$. However, its estimated coefficient decreases from 2.75 ($\tau = 0.2$) to 1.39 ($\tau = 0.7$), indicating its heterogeneous impacts on survival, that is, stronger protective effect at lower quantiles and vice versa.

The effects of some SNPs are nearly zero for higher quantiles. For example, the estimated coefficient of AX.15207405_G decreases from 2.03 ($\tau = 0.2$; $p = 10^{-24}$) to -0.05 ($\tau = 0.7$; $p = 0.92$), with the estimated SE increasing from 0.20 to 0.48. Similarly, the estimated coefficient of AX.40182999_A decreases from 1.5 ($\tau = 0.2$; $p = 9.6 \times 10^{-13}$) to -0.01 ($\tau = 0.7$; $p = 0.96$). The results again hint at heterogeneous SNP effects in various risk groups, which cannot be detected using traditional Cox models.

Finally, our results shed light on the roles of SNPs in the high-risk group (i.e., lower quantiles). Specifically, we map the 83 SNPs with significant effects at the 0.2th quantile by Fused-HDCQR to the corresponding genes and rank the genes by the number of significant SNPs (over total number of SNPs for each gene in the parenthesis), which are TP53 (14/321), RRM1 (14/174), ERCC1 (10/167), BRCA1 (10/114), ALK (8/163), ROS1 (5/294), EGFR (5/261), ERBB2 (4/167), and 6 other genes with the number of significant SNPs less than 4. While these genes were reported to be associated with lung cancer (Toyooka, Tsuda, and Gazdar 2003; Takeuchi et al. 2012; Rosell et al. 2011; Lord et al. 2002; Zheng et al. 2007; Sasaki et al. 2006; Brose et al. 2002), our analysis provides more detailed information as to which SNPs and locations of the genes are jointly associated with the lung cancer survival, as well as the estimated effects and uncertainties. Analysis of heterogeneous SNP effects has been gaining increasing attention in lung cancer research (McKay et al. 2017; Dong et al. 2012; Huang et al. 2009), and beyond it (Garcia-Closas et al. 2008; Cheng et al. 2010; Gulati et al. 2014).

7. Conclusions

Our proposed procedure involves repeated estimates from low-dimensional CQRs, which are computationally straightforward and can be efficiently implemented with parallel computing. We require the variable selection to possess a sure screening property as in Condition (A4). This seems to be supported by our simulations, which find our procedure works well when the variable selection method can select a superset of the true model with high probability. Our condition is much weaker than a condition of selection consistency as specified in Fei et al. (2019).

For the selection of B, we recommend B to be in the same order of the sample size n. Smaller B might not affect coefficient estimation much; but it might yield inaccurately estimated SEs, leading to incorrect inferences. In addition, we opt to define Γ_m by setting the grid as $n/\log p$ equally spaced points between τ_0 and τ_U. This may cover the quantile interval well, with reasonable computation efficiency.

There are open questions to be addressed. First, substantial work is needed for handling highly correlated predictors as the performance of our method, like the other competing methods, deteriorates when correlations among predictors become stronger. Second, it is of interest to investigate an alternative method when the sparsity condition fails. For example, it is challenging to find an effective strategy to draw inferences when a non-negligible portion of predictors have small but nonzero effects. We will pursue them elsewhere.

Supplementary Materials

The supplementary materials contain additional simulations to compare our method with various ad hoc approaches under the same settings as in Example 2. Also included are the proofs of Theorems 1–3 and additional lemmas.

Acknowledgments

Funding

ORCID

Yi Li http://orcid.org/0000-0003-1720-2760

References

Belloni, A., Chernozhukov, V., and Kato, K. (2019), "Valid Post-Selection Inference in High-Dimensional Approximately Sparse Quantile Regression Models," *Journal of the American Statistical Association*, 114, 749–758. [2]

Berk, R., Brown, L., Buja, A., Zhang, K., and L. Zhao (2013), "Valid Post-Selection Inference," *The Annals of Statistics*, 41, 802–837. [2]

Brose, M. S., Volpe, P., Feldman, M., Kumar, M., Rishi, I., Gerrero, R., Einhorn, E., Herlyn, M., Minna, J., Nicholson, A., Roth, J. A., Albelda, S. M., Davies, H., Cox, C., Brignell, G., Stephens, P., Futreal, P. A., Wooster, R., Stratton, M. R., Weber, B. L. (2002), "BRAF and RAS Mutations in Human Lung Cancer and Melanoma," *Cancer Research*, 62, 6997–7000. [12]

Bühlmann, P., Kalisch, M., and Meier, L. (2014), "High-Dimensional Statistics With a View Toward Applications in Biology," *Annual Review of Statistics and Its Application*, 1, 255–278. [2]

Cheng, I., Plummer, S. J., Neslund-Dudas, C., Klein, E. A., Casey, G., Rybicki, B. A., and Witte, J. S. (2010), "Prostate Cancer Susceptibility Variants Confer Increased Risk of Disease Progression," *Cancer Epidemiology and Prevention Biomarkers*, 19, 2124–2132. [12]

Christiani, D. C. (2017), "The Boston Lung Cancer Survival Cohort," available at http://grantome.com/grant/NIH/U01-CA209414-01A1. [1]

D'Antonio, M., Reyna, J., Jakubosky, D., Donovan, M. K., Bonder, M.-J., Matsui, H., Stegle, O., Nariai, N., D'Antonio-Chronowska, A., Frazer, K. A. (2019), "Systematic Genetic Analysis of the MHC Region Reveals Mechanistic Underpinnings of HLA Type Associations With Disease," *eLife*, 8, e48476. [12]

Dong, J., Hu, Z., Shu, Y., Pan, S., Chen, W., Wang, Y., Hu, L., Jiang, Y., Dai, J., Ma, H., Jin, G., Shen, H. (2012), "Potentially Functional Polymorphisms in DNA Repair Genes and Non-Small-Cell Lung Cancer Survival: A Pathway-Based Analysis," *Molecular Carcinogenesis*, 51, 546–552. [1,2,12]

Efron, B. (2014), "Estimation and Accuracy After Model Selection," *Journal of the American Statistical Association*, 109, 991–1007. [2,5]

Evans, D. M., Spencer, C. C., Pointon, J. J., Su, Z., Harvey, D., Kochan, G., Oppermann, U., Dilthey, A., Pirinen, M., Stone, M. A., Appleton, L., Moutsianas, L., Leslie, S., Wordsworth, T., Kenna, T. J., Karaderi, T., Thomas, G. P., Ward, M. M., Weisman, M. H., Farrar, C., Bradbury, L. A., Danoy, P., Inman, R. D., Maksymowych, W., Gladman, D., Rahman, P., Spondyloarthritis Research Consortium of Canada (SPARCC); Morgan, A., Marzo-Ortega, H., Bowness, P., Gaffney, K., Hill Gaston, J. S., Smith, M., Bruges-Armas, J., Couto, A.-R., Sorrentino, R., Paladini, F., Ferreira, M. A., Xu, H., Liu, Y., Jiang, L., Lopez-Larrea, C., Díaz-Peña, R., López-Vázquez, A., Zayats, T., Band, G., Bellenguez, Cé., Blackburn, H., Blackwell, J. M., Bramon, E., Bumpstead, S. J., Casas, J. P., Corvin, A., Craddock, N., Deloukas, P., Dronov, S., Duncanson, A., Edkins, S., Freeman, C., Gillman, M., Gray, E., Gwilliam, R., Hammond, N., Hunt, S. E., Jankowski, J., Jayakumar, A., Langford, C., Liddle, J., Markus, H. S., Mathew, C. G., McCann, O. T., McCarthy, M. I., Palmer, C. N. A., Peltonen, L., Plomin, R., Potter, S. C., Rautanen, A., Ravindrarajah, R., Ricketts, M., Samani, N., Sawcer, S. J., Strange, A., Trembath, R. C., Viswanathan, A. C., Waller, M., Weston, P., Whittaker, P., Widaa, S., Wood, N. W., McVean, G., Reveille, J. D., Wordsworth, B. P., Brown, M. A., Donnelly, P., Australo-Anglo-American Spondyloarthritis Consortium (TASC); Wellcome Trust Case Control Consortium 2 (WTCCC2) (2011). "Interaction Between ERAP1 and HLA-B27 in Ankylosing Spondylitis Implicates Peptide Handling in the Mechanism for HLA-B27 in Disease Susceptibility," *Nature Genetics*, 43, 761–767. [12]

Fan, J., Fan, Y., and Barut, E. (2014), "Adaptive Robust Variable Selection," *The Annals of Statistics*, 42, 324–351. [1,5,8]

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 2013–2038. [6,8]

Fang, E. X., Ning, Y., and Liu, H. (2017), "Testing and Confidence Intervals for High Dimensional Proportional Hazards Models," *Journal of the Royal Statistical Society*, Series B, 79, 1415–1437. [2]

Fei, Z., and Li, Y. (2021), "Estimation and Inference for High Dimensional Generalized Linear Models: A Splitting and Smoothing Approach," *Journal of Machine Learning Research*, 22, 1–32. [2,4,5]

Fei, Z., Zhu, J., Banerjee, M., and Li, Y. (2019), "Drawing Inferences for High-Dimensional Linear Models: A Selection-Assisted Partial Regression and Smoothing Approach," *Biometrics*, 75, 551–561. [2,4,12]

Fleming, T. R., and Harrington, D. P. (2011), *Counting Processes and Survival Analysis*, Vol. 169. John Wiley & Sons. [2]

Garcia-Closas, M., Hall, P., Nevanlinna, H., Pooley, K., Morrison, J., Richesson, D. A., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K., Arias, J. I., Milne, R. L., Ribas, G., González-Neira, A., Benítez, J., Zamora, P., Brauch, H., Justenhoven, C., Hamann, U., Ko, Y. D., Bruening, T., Haas, S., D ork, T., Sch urmann, P., Hillemanns, P., Bogdanova, N., Bremer, M., Karstens, J. H., Fagerholm, R., Aaltonen, K., Aittom aki, K., von Smitten, K., Blomqvist, C., Mannermaa, A., Uusitupa, M., Eskelinen, M., Tengstr om, M., Kosma, V. M., Kataja, V., Chenevix-Trench, G., Spurdle, A. B., Beesley, J., Chen, X., Australian Ovarian Cancer Management Group; Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer, Devilee, P., van Asperen, C. J., Jacobi, C. E., Tollenaar, R. A., Huijts, P. E., Klijn, J. G., Chang-Claude, J., Kropp, S., Slanger, T., Flesch-Janys, D., Mutschelknauss, E., Salazar, R., Wang-Gohrke, S., Couch, F., Goode, E. L., Olson, J.E., Vachon, C., Fredericksen, Z. S., Giles, G. G., Baglietto, L., Severi, G., Hopper, J. L., English, D. R., Southey, M. C., Haiman, C. A., Henderson, B. E., Kolonel, L. N., Le Marchand, L., Stram, D. O., Hunter, D. J., Hankinson, S. E., Cox, D. G., Tamimi, R., Kraft, P., Sherman, M. E., Chanock, S. J., Lissowska, J., Brinton, L. A., Peplonska, B., Klijn, J. G., Hooning, M. J., Meijers-Heijboer, H., Collee, J. M., van den Ouweland, A., Uitterlinden, A. G., Liu, J., Lin, L. Y., Yuqing, L., Humphreys, K., Czene, K., Cox, A., Balasubramanian, S. P., Cross, S. S., Reed, M. W., Blows, F., Driver, K., Dunning, A., Tyrer, J., Ponder, B. A., Sangrajrang, S., Brennan, P., McKay, J., Odefrey, F., Gabrieau, V., Sigurdson, A., Doody, M., Struewing, J. P., Alexander, B., Easton, D. F., Pharoah, P. D. (2008), "Heterogeneity of Breast Cancer Associations With Five Susceptibility Loci by Clinical and Pathological Characteristics," *PLoS Genetics*, 4, e1000054. [12]

Gulati, S., Martinez, P., Joshi, T., Birkbak, N. J., Santos, C. R., Rowan, A. J., Pickering, L., Gore, M., Larkin, J., Szallasi, Z., Bates, P.A., Swanton, C., Gerlinger, M. (2014), "Systematic Evaluation of the Prognostic Impact and Intratumour Heterogeneity of Clear Cell Renal Cell Carcinoma Biomarkers," *European Urology*, 66, 936–948. [12]

Hanawa, H., Hargrove, P. W., Kepes, S., Srivastava, D. K., Nienhuis, A. W., and Persons, D. A. (2004), "Extended β-Globin Locus Control Region Elements Promote Consistent Therapeutic Expression of a γ-Globin Lentiviral Vector in Murine β-Thalassemia," *Blood*, 104, 2281–2290. [12]

He, X., Wang, L., and Hong, H. G. (2013), "Quantile-Adaptive Model-Free Variable Screening for High-Dimensional Heterogeneous Data," *The Annals of Statistics*, 41, 342–369. [1,3,4]

Ho, D. S. W., Schierding, W., Wake, M., Saffery, R., and O'Sullivan, J. (2019), "Machine Learning SNP Based Prediction for Precision Medicine," *Frontiers in Genetics*, 10, 267. [1]

Hong, H. G., Christiani, D. C., and Li, Y. (2019), "Quantile Regression for Survival Data in Modern Cancer Research: Expanding Statistical Tools for Precision Medicine," *Precision Clinical Medicine*, 2, 90–99. [2]

Huang, Y.-T., Heist, R. S., Chirieac, L. R., Lin, X., Skaug, V., Zienolddiny, S., Haugen, A., Wu, M. C, Wang Z, Su L, Asomaning K, Christiani DC. (2009), "Genome-Wide Analysis of Survival in Early-Stage Non–Small-Cell Lung Cancer," *Journal of Clinical Oncology*, 27, 2660–2667. [1,2,12]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [2]

Kelley, M. J., Li, S., and Harpole, D. H. (2001), "Genetic Analysis of the β-Tubulin Gene, Tubb, in Non-Small-Cell Lung Cancer," *Journal of the National Cancer Institute*, 93, 1886–1888. [1]

Klee, V., and Minty, G. J. (1972), "How Good is the Simplex Algorithm," *Inequalities*, 3, 159–175. [3]

Koenker, R. (2005). *Quantile Regression*, Cambridge: Cambridge University Press. [3]

Koenker, R., and Bassett, G. Jr (1978), "Regression Quantiles," *Econometrica: Journal of the Econometric Society*, 46, 33–50. [1]

Kong, S., Yu, Z., Zhang, X., and Cheng, G. (2021), "High Dimensional Robust Inference for COX Regression Models Using De-Sparsified Lasso," *Scandinavian Journal of Statistics*. doi: 10.1111/sjos.12543. [2]

Korpanty, G. J., Graham, D. M., Vincent, M. D., and Leighl, N. B. (2014), "Biomarkers That Currently Affect Clinical Practice in Lung Cancer: EGFR, ALK, MET, ROS-1, and KRAS," *Frontiers in Oncology*, 4, 204. [1]

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016), "Exact Post-Selection Inference, With Application to the Lasso," *The Annals of Statistics*, 44, 907–927. [2]

Lord, R. V., Brabender, J., Gandara, D., Alberola, V., Camps, C., Domine, M., Cardenal, F., Sánchez, J. M., Gumerlock, P. H., Tarón, M., Sánchez, J. J., Danenberg, K. D., Danenberg, P. V., Rosell, R. (2002), "Low ERCC1 Expression Correlates With Prolonged Survival After Cisplatin Plus Gemcitabine Chemotherapy in Non-Small Cell Lung Cancer," *Clinical Cancer Research*, 8, 2286–2291. [12]

Lovász, L., and Vempala, S. (2007), "The Geometry of Logconcave Functions and Sampling Algorithms," *Random Structures & Algorithms*, 30, 307–358. [4]

McKay, J. D., Hung, R. J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D. C., Caporaso, N. E., Johansson, M., Xiao, X., Li, Y., Byun, J., Dunning, A., Pooley, K. A., Qian, D. C., Ji, X., Liu, G., Timofeeva, M. N., Bojesen, S. E., Wu, X., Le Marchand, L., Albanes, D., Bickeb oller, H., Aldrich, M. C., Bush, W. S., Tardon, A., Rennert, G., Teare, M. D., Field, J. K., Kiemeney, L. A., Lazarus, P., Haugen, A., Lam, S., Schabath, M. B., Andrew, A. S., Shen, H., Hong, Y. C., Yuan, J. M., Bertazzi, P. A., Pesatori, A. C., Ye, Y., Diao, N., Su, L., Zhang, R., Brhane, Y., Leighl, N., Johansen, J. S., Mellemgaard, A., Saliba, W., Haiman, C. A., Wilkens, L. R., Fernandez-Somoano, A., Fernandez-Tardon, G., van der Heijden, H. F. M., Kim, J. H., Dai, J., Hu, Z., Davies, M. P. A., Marcus, M. W., Brunnstr om, H., Manjer, J., Melander, O., Muller, D. C., Overvad, K., Trichopoulou, A., Tumino, R., Doherty, J. A., Barnett, M. P., Chen, C., Goodman, G. E., Cox, A., Taylor, F., Woll, P., Br uske, I., Wichmann, H. E., Manz, J., Muley, T. R., Risch, A., Rosenberger, A., Grankvist, K., Johansson, M., Shepherd, F. A., Tsao, M. S., Arnold, S. M., Haura, E. B., Bolca, C., Holcatova, I., Janout, V., Kontic, M., Lissowska, J., Mukeria, A., Ognjanovic, S., Orlowski, T. M., Scelo, G., Swiatkowska, B., Zaridze, D., Bakke, P., Skaug, V., Zienolddiny, S., Duell, E. J., Butler, L. M., Koh, W. P., Gao, Y. T., Houlston, R. S., McLaughlin, J., Stevens, V. L., Joubert, P., Lamontagne, M., Nickle, D. C., Obeidat, M., Timens, W., Zhu, B., Song, L., Kachuri, L., Artigas, M. S., Tobin, M. D., Wain, L. V.; SpiroMeta Consortium, Rafnar, T., Thorgeirsson, T. E., Reginsson, G. W., Stefansson, K., Hancock, D. B., Bierut, L. J., Spitz, M. R., Gaddis, N. C., Lutz, S. M., Gu, F., Johnson, E. O., Kamal, A., Pikielny, C., Zhu, D., Lindstr oem, S., Jiang, X., Tyndale, R. F., Chenevix-Trench, G., Beesley, J., Boss,é Y., Chanock, S., Brennan, P., Landi, M. T., Amos, C. I. (2017), "Large-Scale Association Analysis Identifies New Lung Cancer Susceptibility Loci and Heterogeneity in Genetic Susceptibility Across Histological Subtypes," *Nature Genetics*, 49, 1126–1132. [1,2,12]

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "P-values for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681. [2,5,8]

Moon, C., Oh, Y., and Roth, J. A. (2003), "Current Status of Gene Therapy for Lung Cancer and Head and Neck Cancer," *Clinical Cancer Research*, 9, 5055–5067. [1]

Ning, Y., and Liu, H. (2017), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," *The Annals of Statistics*, 45, 158–195. [2]

Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F. J., Lindeman, N., Boggon, TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M. (2004), "EGFR Mutations in Lung Cancer: Correlation With Clinical Response to Gefitinib Therapy," *Science*, 304, 1497–1500. [12]

Pâques, F., and Duchateau, P. (2007), "Meganucleases and DNA Double-Strand Break-Induced Recombination: Perspectives for Gene Therapy," *Current Gene Therapy*, 7, 49–66. [12]

Peng, L., and Huang, Y. (2008), "Survival Analysis With Quantile Regression Models," *Journal of the American Statistical Association*, 103, 637–649. [1,2,4]

Portnoy, S. (2003), "Censored Regression Quantiles," *Journal of the American Statistical Association*, 98, 1001–1012. [1]

Powell, J. L. (1986), "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155. [1]

Risch, A., and C. Plass (2008), "Lung Cancer Epigenetics and Genetics," *International Journal of Cancer*, 123, 1–7. [1]

Rosell, R., Molina, M. A., Costa, C., Simonetti, S., Gimenez-Capitan, A., Bertran-Alamillo, J., Mayo, C., Moran, T., Mendez, P., Cardenal, F., Isla, D., Provencio, M., Cobo, M., Insa, A., Garcia-Campelo, R., Reguart, N., Majem, M., Viteri, S., Carcereny, E., Porta, R., Massuti, B., Queralt, C,, de Aguirre, I., Sanchez, J.M., Sanchez-Ronco, M., Mate, J.L., Ariza, A., Benlloch, S., Sanchez, J. J., Bivona, T. G., Sawyers, C. L., and Taron, M. (2011), "Pretreatment EGFR T790M Mutation and BRCA1 mRNA Expression in Erlotinib-Treated Advanced Non–Small-Cell Lung Cancer Patients With EGFR Mutations," *Clinical Cancer Research*, 17, 1160–1168. [12]

Sasaki, H., Shimizu, S., Endo, K., Takada, M., Kawahara, M., Tanaka, H., Matsumura, A., Iuchi, K., Haneda, H., Suzuki, E., Kobayashi, Y., Yano, M., Fujii, Y. (2006), "EGFR and erbB2 Mutation Status in Japanese Lung Cancer Patients," *International Journal of Cancer*, 118, 180–184. [12]

Shows, J. H., Lu, W., and Zhang, H. H. (2010), "Sparse Estimation and Inference for Censored Median Regression," *Journal of Statistical Planning and Inference*, 140, 1903–1917. [2]

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., Bando, M., Ohno, S., Ishikawa, Y., Aburatani, H., Niki, T., Sohara, Y., Sugiyama, Y., Mano, H. (2007), "Identification of the Transforming EML4–ALK Fusion Gene in Non-Small-Cell Lung Cancer," *Nature*, 448, 561–566. [12]

Takeuchi, K., Soda, M., Togashi, Y., Suzuki, R., Sakata, S., Hatano, S., Asaka, R., Hamanaka, W., Ninomiya, H., Uehara, H., Lim Choi, Y., Satoh, Y., Okumura, S., Nakagawa, K., Mano, H., and Ishikawa, Y. (2012), "RET, ROS1 and ALK Fusions in Lung Cancer," *Nature Medicine*, 18, 378–381. [12]

Toyooka, S., Tsuda, T., and Gazdar, A. F. (2003), "The TP53 Gene, Tobacco Exposure, and Lung Cancer," *Human Mutation*, 21, 229–239. [12]

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [2]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [2,5]

Wager, S., Hastie, T., and Efron, B. (2014), "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife," *Journal of Machine Learning Research*, 15, 1625–1651. [5]

Wang, L., Wu, Y., and Li, R. (2012), "Quantile Regression for Analyzing Heterogeneity in Ultra-High Dimension," *Journal of the American Statistical Association*, 107, 214–222. [1,4,5,8]

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. [2]

Yamamoto, H., Shigematsu, H., Nomura, M., Lockwood, W. W., Sato, M., Okumura, N., Soh, J., Suzuki, M., Wistuba, I. I., Fong, K. M., Lee, H., Toyooka, S., Date, H., Lam, W. L., Minna, J. D., Gazdar, A. F. (2008), "Pik3ca Mutations and Copy Number Gains in Human Lung Cancers," *Cancer Research*, 68, 6913–6921. [1]

Zhang, C.-H., and Huang, J. (2008), "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594. [4]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [1,2]

Zhao, S. D., and Li, Y. (2012), "Principled Sure Independence Screening for Cox Models With Ultra-High-Dimensional Covariates," *Journal of Multivariate Analysis*, 105, 397–411. [2]

Zheng, Q., Gallagher, C., and Kulasekera, K. (2013), "Adaptive Penalized Quantile Regression for High Dimensional Data," *Journal of Statistical Planning and Inference*, 143, 1029–1038. [1]

Zheng, Q., Peng, L., and He, X. (2015), "Globally Adaptive Quantile Regression With Ultra-High Dimensional Data," *The Annals of Statistics*, 43, 2225–2258. [2,4,12]

——— (2018), "High Dimensional Censored Quantile Regression," *The Annals of Statistics*, 46, 308–343. [1,2,3,4,5,8]

Zheng, Z., Chen, T., Li, X., Haura, E., Sharma, A., and Bepler, G. (2007), "DNA Synthesis and Repair Genes RRM1 and ERCC1 in Lung Cancer," *New England Journal of Medicine*, 356, 800–808. [12]

Zhu, Q.-G., Zhang, S.-M., Ding, X.-X., He, B., and Zhang, H.-Q. (2017), "Driver Genes in Non-Small Cell Lung Cancer: Characteristics, Detection Methods, and Targeted Therapies," *Oncotarget*, 8, 57680–57692. [1]