Inference for High-Dimensional Censored Quantile Regression

Zhe Feia, QiZheng b, Hyokyoung G. Hongc, and Yi LId

aDepartment of Biostatistics, University of California, Los Angeles, CA; bDepartment of Bioinformatics and Biostatistics, University of Louisville, KY; cDepartment of Statistics and Probability, Michigan State University, East Lansing, MI; dDepartment 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 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 fusion 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 problem, by assessing the covariate effects on the survival distribution, from the accelerated failure time model (Wei 1992) and off effects.

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 Z, a (p − 1)-dimensional vector of covariates (p ≥ 1). Let \( X = \min(T, C), \Delta = 1(T \leq C), \) and \( Z = (1, Z^T)^T, \) where 1\( \cdot \) is the binary indicator function. The observed data, \( D^{(i)} = \{(X_i, \Delta_i, Z_i), i = 1, \ldots, n\}, \) are n identical and independently distributed (iid) copies of \( (X, \Delta, Z) \). With \( Y = \log T \), let \( Q_y(\tau|Z) = \inf\{t: P(Y \leq t|Z) \geq \tau\} \) be the \( \tau \)th conditional quantile of \( Y \) given \( Z \). A global CQR model stipulates

\[
Q_y(\tau|Z) = Z^T \beta^*(\tau), \quad \tau \in (0, 1),
\]

where \( \beta^*(\tau) \) is a \( p \)-dimensional vector of coefficients at \( \tau \). 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 \( N(t) = 1(\log X \leq t, \Delta = 1), A_T(t|Z) = -\log(1 - P(\log T \leq t|Z)), \) and \( H(u) = -\log(1 - u) \). Then, \( M(t) = N(t) - A_T(t \wedge \log X|Z) \) is a martingale process under model (1) (Fleming and Harrington 2011) and hence \( E(M(t)|Z) = 0 \). 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 \( \theta_i(\tau) = Z_i^T \beta(\tau) \) and

\[
U_n(\beta, \tau) = -n^{-1} \sum_{i=1}^n \left( N_i(\theta_i(\tau)) - \int_0^\tau 1(\log X_i \geq \theta_i(u))dH(u) \right).
\]

We denote by \( \mathbf{u}(\beta, \tau) \) the expectation of \( U_n(\beta, \tau) \).

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

\[
n^{-1/2} U_n(\beta, \tau) = n^{-1/2} \sum_{i=1}^n Z_i \left[ N_i(\theta_i(\tau)) - \int_0^\tau 1(\log X_i \geq \theta_i(u))dH(u) \right] = 0.
\]

The stochastic integral in Equation (2) naturally suggests sequential estimation with respect to \( \tau \). 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, \( \nu \) is chosen such that only a small proportion of observations are censored below the \( \nu \)th quantile.

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

\[
n^{-1/2} \sum_{i=1}^n Z_i \left( N_i(\hat{\theta}_i(\tau_k)) - \int_{\tau_k}^{\tau_{k+1}} 1(\log X_i \geq \hat{\theta}_i(\tau))dH(u) \right) = 0,
\]

where \( \hat{\theta}_i(\tau_k) = Z_i^T \hat{\beta}(\tau_k) \). Due to the monotonicity of \( \theta_i(\tau) \) in \( \tau \), \( \hat{\beta}(\tau_k) \) can be solved efficiently via L1-minimization. And \( \hat{\beta}(\tau), \tau \in [\nu, \tau_U] \), is defined as a right-continuous piece-wise constant function that only jumps at the grid points. It can be shown that \( \hat{\beta}(\tau) \) is uniformly consistent and converges weakly.
to a mean zero Gaussian process for $\tau \in [v, \tau_U]$ when $p = o(n)$. More importantly, $\hat{\beta}(\tau)$ provides a comprehensive understanding of the covariate effects on the conditional survival distribution over the quantile interval $[v, \tau_U]$. 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 the penalized estimation and selection procedure developed by He, Wang, and Hong (2013) incorporates 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) $\lambda_n$ of $S$, 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, $D^b_{1}$ and $D^b_{2}$;
   (ii) on $D^b_{1}$, apply $S$ with $\lambda_n$ on $[v, \tau_U]$, to select a subset of predictors, denoted by $S^b_{1}$ or $S^b_{0}$ for short;
   (iii) on $D^b_{2}$, for each $j = 1, \ldots, p$, append $j$ to $S^b_{0}$ such that $S^b_{1} = \{j\} \cup S^b_{0}$, fit a partial CQR on the covariates indexed by $S^b_{1}$, and denote their coefficient estimates by $\hat{\beta}_{S^b_{1}}(\tau), \tau \in [v, \tau_U]$. Here, $\hat{\beta}_{S^b_{1}}(\tau)$ is a right-continuous piecewise-constant function with jumps only at the grid points of $\tau_k \in \Gamma_m$;
   (iv) denote by $\hat{\beta}_{S^b_{0}}^b(\tau) = \left(\hat{\beta}_{S^b_{1}}(\tau)\right)_j$ the entry of $\hat{\beta}_{S^b_{1}}(\tau)$ corresponding to $Z_j$.

3. Fusing: the final estimate of $\hat{\beta}_{S^b_{1}}(\tau), \tau \in [v, \tau_U], j = 1, \ldots, p$ is

$$\hat{\beta}_j(\tau) = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_{S^b_{0}}^b(\tau).$$

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

Remark 2. Our procedure needs a variable selection procedure to reduce dimension. For example, L-HDCQR selects a subset: $\{j \in \{2, \ldots, p\} : \max_k |\hat{\beta}_j(\tau_k)| > a_0, \tau_k \in \Gamma_m\}$, where $\hat{\beta}_j(\tau_k)$’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 $[v, \tau_U]$, and choose AL-HDCQR if we opt to select fewer predictors.

Remark 3. We select $\lambda_n$ 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 $\hat{\beta}_{S^b_{0}}^{(k)}(\tau)$ be the penalized estimate of $\beta(\tau)$ using all of the data excluding the $k$th fold with a tuning parameter $\lambda$ and $\tau \in [v, \tau_U]$, where $k = 1, \ldots, K$. Under the global CQR model (1), we define the cross-validation error as

$$\text{CV Error}(\lambda) = \sum_{k=1}^{K} \sum_{\tau \in [v, \tau_U]} \int_{\tau_U}^{\tau} |D_k[\hat{\beta}_{S^b_{0}}^{(k)}(\tau)]| \, dt,$$

where

$$D_k[\beta(\tau)] = \text{sign}\left\{M_k(\beta(\tau))\right\} \sqrt{-2M_k(\beta(\tau)) + \Delta_k \log \left\{\Delta_k - M_k(\beta(\tau))\right\}}$$

with $M_k(\beta(\tau)) = N_k(Z^T_k \beta(\tau)) - \int_{\tau}^{\tau_U} 1 \log|X| \geq N_k(Z^T_k \beta(\tau)) \log|H(u)| - \nu$. Here, $H(u) = -\log(1 - u)$, $N_k(\cdot)$ is the counting process, and $M_k(\beta(\tau))$ 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 $O(n^2p)$ (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 $O(n^4p)$.

3. Theoretical Studies

3.1. Notation and Regularity Conditions

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

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

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

holds for sufficiently large $n$.

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

(A3) (Bounded densities) Let $F_T(t|Z) = P(\log T \leq t|Z)$, $\Lambda_T(t|Z) = -\log(1 - F_T(t|Z))$, $F(t|Z) = P(\log X \leq t|Z)$, and $G(t|Z) = P(\log X \leq t, \Delta = 1|Z)$. Also, define $f(t|Z) = dF(t|Z)/dt$, and $g(t|Z) = dG(t|Z)/dt$. 
(a) There exist constants $f, \tilde{f}, g$ and $\tilde{g}$ such that
\[
\begin{align*}
    f & \leq \inf_{z, \tau \in \psi_{[\nu, \tau]}} f(z^T \beta^*(\tau) | z) \leq \sup_{z, \tau \in \psi_{[\nu, \tau]}} f(z^T \beta^*(\tau) | z) \leq \tilde{f}, \\
g & \leq \inf_{z, \tau \in \psi_{[\nu, \tau]}} g(z^T \beta^*(\tau) | z) \leq \sup_{z, \tau \in \psi_{[\nu, \tau]}} g(z^T \beta^*(\tau) | z) \leq \tilde{g}.
\end{align*}
\]

(b) There exist constants $\kappa > 0$ and $A$ such that, when $|t| \leq \kappa$,
\[
\begin{align*}
    \sup_{z, \tau \in \psi_{[\nu, \tau]}} |f(z^T \beta^*(\tau) + t | z) - f(z^T \beta^*(\tau) | z)| & \leq A|t|, \\
    \sup_{z, \tau \in \psi_{[\nu, \tau]}} |g(z^T \beta^*(\tau) + t | z) - g(z^T \beta^*(\tau) | z)| & \leq A|t|.
\end{align*}
\]

(A4) (Sparsity) Assume $\log p = o(n^{1/2})$, and let
\[
S^* = \bigcup_{\tau \in \nu_{[\nu, \tau]}} \{j : |\beta_j^*(\tau)| \neq 0\}, \quad S^* = \left\{ j : \sup_{\tau \in \nu_{[\nu, \tau]}} |\beta_j^*(\tau)| > 0 \right\}, \quad \text{and} \quad q = |S^*|.
\]

Let $\hat{S}$ be the index set of covariates selected by $S$ with a tuning parameter $\lambda$. There exist constants $0 \leq c_1 < 1/3$, $c_2$, $K_1$, $K_2 > 0$ such that $q \leq K_1 n^{\nu_1}$, $|\hat{S}| \leq K_1 n^{\nu_2}$, and $P(S^* \subseteq \hat{S}) \geq 1 - K_2(p \wedge n)^{-1/2}$.  

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

(A6) (Bounded eigenvalues) $\delta^2 E[|Z|, Z^T] / \|\delta\|^2$ is bounded below and above by $\lambda_{\min}$ and $\lambda_{\max}$ respectively, over $\|\delta\| \leq K_1 n^{\nu_1}$, $\delta \neq 0$, where $0 < \lambda_{\min} < \lambda_{\max}$.

(Nonlinear impact) $\epsilon_2 := \inf_{\|\delta\| \leq K_1 n^{\nu_1}, \delta \neq 0} E[(Z^T \delta)^2]^{3/2} / E[Z^T \delta^2] > 0$.

(A7) $\Gamma_m$ is equally gridded with $\tau_k - \tau_{k-1} = c_n = c_0 n^{-1}$ 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 $\nu$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), $\nu$ is chosen such that only a small proportion of the observed survival times below the $\nu$th 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 $f(t | Z)$ between $Z^T \beta^*(\nu)$ and $Z^T \beta^*(\tau)$, which is essential for the identifiability of $\beta^*(\tau)$ for $\tau < \tau_v$. (A4) restricts the order of data dimensions, as well as the sparsity of $\beta^*(\tau)$, 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 for $S$. 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
\[
\inf_{\mu \leq \nu} \sup_{\tau \in \nu_{[\nu, \tau]}} |\beta_j^*(\tau)| > C_1 \exp(C_2 q \tau_v) \sqrt{q \log(p \vee n) / n} + L \sqrt{q \tau_v}.
\]

- 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 $\beta^*(\tau)$. 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 minorization 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 $Z_k$ 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 $\Gamma_m$, 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 $\hat{S}$’s in the fusing procedure vary around $S^*$. Specifically, we assume a subset $S \subset \{1, \ldots, p\}$ in Theorems 1 and 2, where $|S| \leq K_1 n^{\nu_1}$, $0 \leq c_1 < 1/3$ and $K_1 > 0$. Let $\hat{\beta}(\tau), \tau \in [\nu, \tau_v]$ be the estimator from Peng and Huang (2008) of fitting the CQR with $Z_k$, the covariates indexed by $S$, over the $\tau$-grid $\Gamma_m$.

**Theorem 1.** (Consistency with a diverging number of covariates) Under Conditions (A1)–(A7) and given a subset $S \subset \{1, \ldots, p\}$ such that $S^* \subseteq S$ and $|S| \leq K_1 n^{\nu_1}$, there exist positive constants $\xi_1$ and $\xi_2$ such that
\[
\sup_{\nu \leq \tau \leq \tau_v} \|\hat{\beta}(\tau) - \beta^*(\tau)\| \leq \xi_1 \exp(\xi_2(K_1 n^{\nu_1-1} \log n)^{1/2})
\]
with probability at least $1 - 20c_n^{-2}K_1 n^{\nu_1-2}$.

**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 $\xi_1$ and $\xi_2$ do not depend on the choice of $S$ or $n$. Thus, $\xi_1$ and $\xi_2$ are universal for all possible $S$ satisfying $S^* \subseteq S$ and $|S| \leq K_1 n^{\nu_1}$.

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

Theorem 2. (Weak convergence with a diverging number of covariates) Under Conditions (A1)–(A7) and given a \( S \subseteq \{1, \cdots, p\} \) such that \( S^* \subseteq S \) and \( |S| \leq K_1 n^{1/4} \), it holds that \[
abla \left( \hat{\beta}_j(\tau) - \beta_j^*(\tau) \right) \]
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 \( j \in \{1, \cdots, p\} \),
\[
\sqrt{n} \left( \hat{\beta}_j(\tau) - \beta_j^*(\tau) \right)
\]
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 \( \hat{\beta}_K(\tau) \), a \( K \)-dimensional subvector of the oracle estimator, still holds for \( \tau \in [\nu, \tau_U] \), that is, \( \sqrt{n} (\hat{\beta}_K(\tau) - \beta_K^*(\tau)) \), \( \tau \in [\nu, \tau_U] \) converges to a \( K \)-dimensional Gaussian distribution at any \( \tau \in [\nu, \tau_U] \). We only need to replace \( \hat{\beta}_j(\tau) \) by \( \hat{\beta}_K(\tau) \) 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 \( \hat{\beta}_K(\tau) \), that is, \( \sqrt{n} (\hat{\beta}_K(\tau) - \beta_K^*(\tau)) \) 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 \( \hat{\beta}_j(\tau) \) depends on the unknown active set \( S^* \), the unknown conditional density functions \( f(t|Z) \) and \( g(t|Z) \), 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 \( i \)th observation is in the \( i \)th sub-sample \( D_b^i \), and \( J_i = B^{-1} \sum_{b=1}^B J_{bi} \). For each \( i = 1, \cdots, n \), we define the resampling covariance between \( J_{bi} \) and \( \bar{\beta}_j^B(t_k) \) at \( t_k \in \Gamma_m \) as
\[
s_j(t_k) = \frac{1}{B} \sum_{b=1}^B (J_{bi} - J_i) \left( \hat{\beta}_j^B(t_k) - \hat{\beta}_j(t_k) \right).
\]
Define \( S_j(t_k) = (s_{1j}(t_k), s_{2j}(t_k), \ldots, s_{nj}(t_k))^T \) and let \( n_1 = |D_1^B| \). It follows that the covariance between \( \hat{\beta}_j(t_k) \) and \( \hat{\beta}_j(t_k) \) can be consistently estimated by
\[
\widehat{\text{Cov}}(\hat{\beta}_j(t_k), \hat{\beta}_j(t_\ell)) = \frac{n_1 - 1}{n} \left( \frac{n}{n_1} \right)^2 \sum_{i=1}^n s_{ij}(t_k) s_{ij}(t_\ell),
\]
\[
= \frac{n(n-1)}{(n_1 - 1)^2} \left( \hat{\beta}_j^B(t_k) - \hat{\beta}_j(t_k) \right)^T S_j(t_k) \hat{\beta}_j(t_k),
\]
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 \( t_\ell = t_k \), a variance estimator for \( \hat{\beta}_j(t_k) \) is
\[
\widehat{\text{Var}}(\hat{\beta}_j(t_k)) = \frac{n(n-1)}{(n_1 - 1)^2} \left( \hat{\beta}_j^B(t_k) - \hat{\beta}_j(t_k) \right)^T S_j(t_k) \hat{\beta}_j(t_k).
\]

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

Since \( \hat{\beta}_j(\tau) \) 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
\[
\left( \hat{\beta}_j(\tau_k) - \Phi^{-1}(1 - \alpha/2) \sqrt{\hat{\text{Var}}_j^B(\tau_k)} \right),
\]
\[
\hat{\beta}_j(\tau_k) + \Phi^{-1}(1 - \alpha/2) \sqrt{\hat{\text{Var}}_j^B(\tau_k)},
\]
where \( \hat{\text{Var}}_j^B(\tau_k) \) is as defined in Equation (6), and \( \Phi \) is the standard normal cumulative distribution function. The \( p \)-value of testing \( H_0: \beta_j^*(\tau_k) = 0 \) for a \( \tau_k \in \Gamma_m \) as
\[
2 \times \left[ 1 - \Phi \left( \frac{\hat{\beta}_j(\tau_k)}{\sqrt{\hat{\text{Var}}_j^B(\tau_k)}} \right) \right].
\]

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
The event times are generated by a heterogeneous effect. We study the procedure not being sensitive to tuning parameters as long as they can ensure sure screening. In practical settings, we recommend the variable selection method (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 \( \tau = 0.25, 0.5 \). As a benchmark for comparisons, we also compute the oracle estimates based on the true model (with \( S^0 \) known). Since W12, F14, and Z18 only provide coefficient estimates without standard errors (SEs), we report the selection 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 smaller 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 \( \beta_4 \) varying with \( \tau \). 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 5, we report the selection frequencies of each signal variable in Table 3 (and also in Table 4), and observe that the selection frequencies increase 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 \( \beta_2 \) and \( \beta_3 \) that vary with \( \tau \), the estimation biases of Fused-HDCQR become smaller and the
Figure 1. Estimated heterogeneous effects and confidence intervals of Fused-HDCQR using Example 3: $\beta^{\ast}_2(\cdot)$ (left panel) and $\beta^{\ast}_5(\cdot)$ (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.

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

NOTE: CR, average censoring rate; q, average true positives; TP, 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.

<table>
<thead>
<tr>
<th>Bias</th>
<th>Oracle</th>
<th>Fused</th>
<th>Z18</th>
<th>F14</th>
<th>W12</th>
<th>EmpSD</th>
<th>SE</th>
<th>Cov</th>
<th>M09</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{21} = 0.5$</td>
<td>0.02</td>
<td>0.02</td>
<td>-0.38</td>
<td>-0.50</td>
<td>-0.48</td>
<td>0.14</td>
<td>0.13</td>
<td>0.93</td>
<td>0.97</td>
<td>0.06</td>
</tr>
<tr>
<td>$\beta_{41} = 1$</td>
<td>-0.00</td>
<td>-0.00</td>
<td>-0.03</td>
<td>-0.68</td>
<td>-0.32</td>
<td>0.14</td>
<td>0.12</td>
<td>0.92</td>
<td>0.00</td>
<td>0.19</td>
</tr>
<tr>
<td>$\beta_{61} = 1.5$</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.25</td>
<td>0.13</td>
<td>0.14</td>
<td>0.95</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Bias</th>
<th>Oracle</th>
<th>Fused</th>
<th>Z18</th>
<th>EmpSD</th>
<th>SE</th>
<th>Cov</th>
<th>M09</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{21} = 1$</td>
<td>0.01</td>
<td>0.13</td>
<td>0.29</td>
<td>0.32</td>
<td>0.31</td>
<td>0.88</td>
<td>0.73</td>
<td>0.82</td>
</tr>
<tr>
<td>0.05</td>
<td>-0.07</td>
<td>0.06</td>
<td>0.33</td>
<td>0.29</td>
<td>0.90</td>
<td>0.11</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>$\beta_{41} = 1$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>-0.53</td>
<td>-0.17</td>
<td>0.09</td>
<td>0.08</td>
<td>0.91</td>
</tr>
<tr>
<td>$\beta_{61} = 1.5$</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>-0.49</td>
<td>-0.19</td>
<td>0.09</td>
<td>0.09</td>
<td>0.90</td>
</tr>
</tbody>
</table>

NOTE: Each $\beta$ has three rows corresponding to $\tau = 0.25, 0.5, 0.75$ 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).
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 \( \tau \) 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.

<table>
<thead>
<tr>
<th>Estimated (n, p) and the numbers shown at ( \tau = 0.5 ). Table 5. Comparisons of computing time (on average per dataset in seconds) when performing Example 1.</th>
<th>Fused</th>
<th>Z18</th>
<th>W12</th>
<th>F14</th>
<th>M09</th>
</tr>
</thead>
<tbody>
<tr>
<td>((n, p) = (300, 1000))</td>
<td>3,108</td>
<td>1,812</td>
<td>2,230</td>
<td>1,231</td>
<td>440</td>
</tr>
<tr>
<td>((n, p) = (700, 1000))</td>
<td>3,108</td>
<td>1,812</td>
<td>2,230</td>
<td>1,231</td>
<td>440</td>
</tr>
</tbody>
</table>

NOTE: See the footnote of Table 2.

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 (\( \leq \) 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.

<table>
<thead>
<tr>
<th>Table 6. Patients’ characteristics in the BLCSC samples. ((n = 674))</th>
<th>Age</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count (%)</td>
<td>60 (10.8)</td>
</tr>
<tr>
<td>Female</td>
<td>259 (38)</td>
<td></td>
</tr>
<tr>
<td>Education level</td>
<td>264 (39)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td>410 (61)</td>
<td></td>
</tr>
<tr>
<td>Nonactive</td>
<td>418 (62)</td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td>256 (38)</td>
<td></td>
</tr>
<tr>
<td>Cancer type</td>
<td>283 (42)</td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>110 (16)</td>
<td></td>
</tr>
<tr>
<td>Squamous cell</td>
<td>281 (42)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>256 (38)</td>
<td></td>
</tr>
<tr>
<td>Cancer stage</td>
<td>110 (16)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>283 (42)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>283 (42)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>110 (16)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>256 (38)</td>
<td></td>
</tr>
</tbody>
</table>
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 $\tau = 0.2$, corresponding to the high-risk groups. Results for the top 10 and the bottom 3 are presented.

<table>
<thead>
<tr>
<th></th>
<th>Estimator</th>
<th>SE</th>
<th>p-value</th>
<th>Estimator</th>
<th>SE</th>
<th>p-value</th>
<th>Estimator</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2</td>
<td></td>
<td></td>
<td>0.3</td>
<td></td>
<td></td>
<td>0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Int</td>
<td>6.90</td>
<td>0.25</td>
<td>1.4E-165</td>
<td>7.48</td>
<td>0.28</td>
<td>4.3E-157</td>
<td>7.94</td>
<td>0.24</td>
<td>3.2E-241</td>
</tr>
<tr>
<td>Adeno</td>
<td>0.20</td>
<td>0.16</td>
<td>2.1E-01</td>
<td>0.14</td>
<td>0.18</td>
<td>4.5E-01</td>
<td>0.02</td>
<td>0.13</td>
<td>8.7E-01</td>
</tr>
<tr>
<td>Squamous</td>
<td>-0.16</td>
<td>0.16</td>
<td>3.0E-01</td>
<td>-0.20</td>
<td>0.16</td>
<td>2.1E-01</td>
<td>-0.34</td>
<td>0.13</td>
<td>1.0E-02</td>
</tr>
<tr>
<td>Stage2</td>
<td>-0.82</td>
<td>0.24</td>
<td>6.3E-04</td>
<td>-0.99</td>
<td>0.25</td>
<td>6.0E-05</td>
<td>-0.98</td>
<td>0.24</td>
<td>3.2E-05</td>
</tr>
<tr>
<td>Stage3</td>
<td>-0.97</td>
<td>0.17</td>
<td>1.6E-08</td>
<td>-1.04</td>
<td>0.20</td>
<td>2.0E-07</td>
<td>-1.13</td>
<td>0.14</td>
<td>2.0E-15</td>
</tr>
<tr>
<td>Stage4</td>
<td>-1.54</td>
<td>0.17</td>
<td>3.0E-20</td>
<td>-1.77</td>
<td>0.20</td>
<td>2.1E-19</td>
<td>-1.86</td>
<td>0.14</td>
<td>2.2E-42</td>
</tr>
<tr>
<td>Age</td>
<td>-0.01</td>
<td>0.01</td>
<td>1.5E-02</td>
<td>-0.01</td>
<td>0.01</td>
<td>3.0E-02</td>
<td>-0.02</td>
<td>0.01</td>
<td>1.0E-02</td>
</tr>
<tr>
<td>Edu</td>
<td>0.08</td>
<td>0.14</td>
<td>6.0E-01</td>
<td>0.06</td>
<td>0.15</td>
<td>6.9E-01</td>
<td>0.07</td>
<td>0.13</td>
<td>5.8E-01</td>
</tr>
<tr>
<td>Female</td>
<td>-0.30</td>
<td>0.13</td>
<td>2.2E-02</td>
<td>-0.35</td>
<td>0.14</td>
<td>1.0E-02</td>
<td>-0.37</td>
<td>0.12</td>
<td>1.3E-03</td>
</tr>
<tr>
<td>Smoke</td>
<td>0.55</td>
<td>0.22</td>
<td>2.0E-02</td>
<td>0.29</td>
<td>0.32</td>
<td>1.1E-02</td>
<td>0.31</td>
<td>0.31</td>
<td>2.9E-06</td>
</tr>
<tr>
<td></td>
<td>2.21</td>
<td>0.35</td>
<td>1.6E-10</td>
<td>0.91</td>
<td>0.29</td>
<td>9.1E-11</td>
<td>0.54</td>
<td>0.33</td>
<td>2.9E-06</td>
</tr>
</tbody>
</table>

|                | 0.5       |     |            | 0.6       |     |            | 0.7       |     |         |
|                | 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-18    | -1.93     | 0.14| 1.7E-14    | -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 |
|                | 2.16      | 0.24| 4.1E-28    | 1.84      | 0.28| 2.8E-11    | 1.39      | 0.25| 4.2E-08 |
|                | 1.15      | 0.27| 1.6E-05    | 0.58      | 0.27| 3.5E-02    | 0.13      | 0.25| 6.0E-01 |
|                | 0.75      | 0.27| 2.3E-03    | 0.34      | 0.37| 3.5E-01    | -0.05     | 0.48| 9.2E-01 |
|                | 0.66      | 0.22| 3.1E-03    | 0.44      | 0.31| 1.5E-01    | 0.18      | 0.35| 6.1E-01 |
|                | 0.54      | 0.27| 4.3E-02    | 0.26      | 0.60| 6.7E-01    | 0.11      | 0.60| 8.6E-01 |
|                | 0.55      | 0.22| 2.0E-02    | 0.29      | 0.22| 1.8E-01    | 0.01      | 0.18| 9.7E-01 |
|                | 0.73      | 0.13| 4.2E-08    | 0.51      | 0.32| 1.1E-01    | 0.22      | 0.46| 6.3E-01 |
|                | 0.41      | 0.18| 2.6E-02    | 0.27      | 0.27| 4.1E-01    | -0.01     | 0.30| 9.6E-01 |
|                | 1.17      | 0.42| 5.4E-03    | 0.61      | 0.52| 2.4E-01    | 0.24      | 0.46| 6.0E-01 |
|                | 1.22      | 0.35| 4.5E-04    | 0.86      | 0.34| 1.1E-02    | 0.50      | 0.31| 1.0E-01 |

|                | 0.26      | 0.60| 0.66      | 0.32      | 0.52| 0.54       | 0.12      | 0.68| 0.86    |
|                | -0.00     | 0.12| 1.00      | -0.09     | 0.12| 0.44       | -0.09     | 0.15| 0.56    |
|                | -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 $\tau$-grid $\Gamma_m$ 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 $\lambda_m$ 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 $\Gamma_m$. 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 $\tau$ changes from 0.2 to 0.5, and then increases to $-0.31$ ($p = 0.038$) as $\tau$ 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 $\Gamma_m$ by setting the grid as $n/\log p$ equally spaced points between $\tau_0$ and $\tau_U$. 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

We are deeply grateful toward the Editor, the AE and the two referees for their constructive comments and suggestions that have improved the article substantially. We thank our long-term collaborator, David Christiani of Harvard Medical School, for providing the BLGSC data.
Funding
The work was partially supported by grants from NIH (grant nos. 5R01CA249096 and U01CA209414).

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

References


