Annals of the Institute of Statistical Mathematics (2023) 75:393–424
https://doi.org/10.1007/s10463-022-00849-z

Forward variable selection for ultra-high dimensional quantile regression models

Toshio Honda[1] · **Chien-Tong Lin**[2]

Received: 20 August 2021 / Revised: 15 April 2022 / Accepted: 19 July 2022 /
Published online: 29 August 2022

Abstract
We propose forward variable selection procedures with a stopping rule for feature screening in ultra-high-dimensional quantile regression models. For such very large models, penalized methods do not work and some preliminary feature screening is necessary. We demonstrate the desirable theoretical properties of our forward procedures by taking care of uniformity w.r.t. subsets of covariates properly. The necessity of such uniformity is often overlooked in the literature. Our stopping rule suitably incorporates the model size at each stage. We also present the results of simulation studies and a real data application to show their good finite sample performances.

Keywords Forward procedure · Check function · Sparsity · Screening consistency · Stopping rule

1 Introduction

Suppose that we have n i.i.d. observations of $(Y, \boldsymbol{X})$, $(Y_i, \boldsymbol{X}_i)$, $i = 1, \dots, n$, and that this $(Y, \boldsymbol{X})$ satisfies the following sparse ultra-high-dimensional τth quantile regression model :

$$Y = \boldsymbol{X}^T \boldsymbol{\beta}^* + \epsilon \quad \text{with} \quad \boldsymbol{X} = (X_1, \dots, X_p)^T \in \mathbb{R}^p \tag{1}$$

and

✉ Toshio Honda
t.honda@r.hit-u.ac.jp

Chien-Tong Lin
ctlin@fcu.edu.tw

[1] Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan

[2] Department of Statistics, Feng Chia University, No.100, Wenhua Rd. Xitun Dist., Taichung City 407102, Taiwan, ROC

$$\mathrm{E}\{\psi_\tau(\epsilon)|\boldsymbol{X}\} = 0 \quad \text{and} \quad \boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathrm{E}\{\rho_\tau(Y - \boldsymbol{X}^T\boldsymbol{\beta})\}, \tag{2}$$

where $X_1 \equiv 1$, $\epsilon_i = Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}^*$,

$$\psi_\tau(t) = \tau - I(t \le 0), \quad \text{and} \quad \rho_\tau(t) = t\{\tau - I(t \le 0)\}.$$

Linear quantile regression models are very popular since Koenker and Basset (1978) (see also Koenker, 2005). We denote the set of relevant covariate indexes by $\mathcal{M} = \{j \in [p] \mid \beta_j^* \neq 0\}$ and set $m = |\mathcal{M}|$, where $[p] := \{1, \ldots, p\}$ and $|S|$ is the number of the elements of $S \subset [p]$. In this paper, we deal with the cases where p can be ultra-high-dimensional like $p = O(\exp(n^{c_p}))$ as specified later and m is much smaller than p with a known upper bound K_n. In such ultra-high-dimensional cases, p can be too large for commonly used penalized methods such as the Lasso (cf. Tibshirani, 1996), the SCAD (cf. Fan and Li, 2001), and the MCP (cf. Zhang, 2010). Besides, in some cases, these established penalized methods also miss some of relevant covariates as shown in our simulation studies in Sect. 3.

Therefore, other feasible procedures for feature screening or variable selection are necessary and a lot of authors have proposed them. Forward regression has been recognized as a helpful tool of feature screening for mean regression models since Wang (2009). Recently we have seen some papers on generalized linear models as we describe later in this section. See also Chapter 8 of Fan et al. (2020). However, there have been only a few papers and no rigorous result on forward feature screening for quantile regression models. This is because the randomness of the newly selected variable at each step affects the asymptotics. Therefore, we propose novel model-based forward procedures for quantile regression models and deal with the proposed procedures rigorously from a theoretical point of view by taking the randomness of the newly selected variable at each step into account. Hence, this paper fills this gap by offering effective forward feature screening procedures for quantile regression models with thorough theoretical justification. We also present the results of numerical studies showing the usefulness of the proposed procedures and our contributions range from theoretical to methodological aspects.

Our forward procedures are greedy ones and may choose some irrelevant covariates. This means we should carry out some statistical inference or apply penalized methods like the SCAD after our procedures. As for the penalized procedures like the Lasso, the adaptive Lasso, and the SCAD for quantile regression models, see, e.g., Belloni and Chernozhukov (2011), Wang et al. (2012), Fan et al. (2014), Zheng et al. (2015), Sherwood and Wang (2016), and Honda et al. (2019). See also Bühlmann and van de Geer (2011), Hastie et al. (2015), and Fan et al. (2020) for general results and recent developments on high-dimensional issues.

There are a lot of feature screening procedures based on marginal models or some association measure between the dependent variable and an individual covariate, e.g., Fan and Lv (2008), Fan and Song (2010), He et al. (2013), and Wu and Yin (2015). It is well known that such procedures may miss some relevant covariates if they are applied only once and some authors have proposed these procedure iteratively with no theory. As for forward variable selection procedures, there are Wang (2009), Ing and Lai (2011), Luo

and Chen (2014), and Cheng et al. (2016), to name a few. Liu et al. (2015) is an excellent review paper of feature screening procedures. Feature screening procedures are also called just screening procedures.

In this paper, we consider forward variable selection procedures for ultra-high-dimensional quantile regression models by minimizing $L_n(\boldsymbol{X}_S^T\boldsymbol{\beta}_S)$ defined in (7) as in Wang (2009) and Cheng et al. (2016) for ultra-high-dimensional mean regression models and Pijyan et al. (2020) and Honda and Lin (2021) for ultra-high-dimensional generalized linear models. In addition, we propose simpler forward procedures for ultra-high-dimensional quantile regression models by using a sequentially conditional approach, not fully minimizing (7), as in Zheng et al. (2020) and Honda and Lin (2021) for ultra-high-dimensional generalized linear models. We examine our forward procedures together with our stopping rule in a unified way. There are some other forward quantile regression procedures proposed in Kong et al. (2019) and Tang et al. (2022). Our procedures are based on minimizing the loss function and can be easily extended to varying coefficient quantile regression models as in Cheng et al. (2016) and Honda and Lin (2021).

When we investigate ultra-high-dimensional forward procedures theoretically, we have to take full care of a kind of uniformity w.r.t. $S \subset [p]$. This is because the newly selected variable at each step is not determined in advance except for an unimaginably ideal setup. Then we describe the properties of our procedures including screening consistency in Sect. 2. As far as we know, no other paper on quantile regression models has paid attention to this kind of uniformity for high-dimensional quantile regression. Stopping rules for forward procedures are often constructed from information criteria such as EBIC. See Chen and Chen (2008, 2012) about EBIC. Lee et al. (2014) gave some useful related results on quantile regression. However, their results do not cover the cases where the upper bound K_n increases to infinity. Our proposed stopping rule covers such cases. We also present the results of our numerical studies in Sect. 3. Our simulation results demonstrate that our procedures compete well with the other procedures and show the best performances in some examples.

This paper is organized as follows. In Sect. 2, we describe the notation, our procedures, technical assumptions, and our main results. We present the results of our numerical studies in Sect. 3. We prove our main theoretical results in Sect. 4. The proofs of technical lemmas are relegated to the supplement. Additional numerical results are also given in the supplement.

2 Forward selection procedures

In this subsection, we introduce the notation and give the details of our procedures. Then we state technical assumptions and finally present our theoretical results in a unified way.

2.1 Notation

We can assume that $\mathrm{E}\{X_j\} = 0$ and $\mathrm{E}\{X_j^2\} = 1$ for $j \geq 2$ without loss of generality. Besides, we set $p_n = p \vee n$ and denote the Euclidean norm of a vector v by $\|v\|$.

For $S \subset [p]$, we define $\boldsymbol{X}_S$ and $\boldsymbol{\beta}_S$ by

$$\boldsymbol{X}_S = (X_j)_{j\in S} \in \mathbb{R}^{|S|} \quad \text{and} \quad \boldsymbol{\beta}_S = (\beta_j)_{j\in S} \in \mathbb{R}^{|S|}$$

from $\boldsymbol{X} = (X_1, \dots, X_p)^T$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$, respectively. Similarly we define $\boldsymbol{X}_{iS}$ from $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^T$. Recall that $X_1 \equiv 1$.

Next we define the regression coefficient $\boldsymbol{\beta}_S^* \in \mathbb{R}^{|S|}$ for a possibly misspecified $S \subset [p]$ and $h_{jS}^* \in \mathbb{R}$ for $j \in S^c$ by

$$\boldsymbol{\beta}_S^* = \operatorname*{arg\,min}_{\boldsymbol{\beta}_S \in \mathbb{R}^{|S|}} \mathrm{E}\big\{\rho_\tau\big(Y - \boldsymbol{X}_S^T \boldsymbol{\beta}_S\big)\big\} \tag{3}$$

and

$$h_{jS}^* = \operatorname*{arg\,min}_{h_{jS} \in \mathbb{R}} \mathrm{E}\big\{\rho_\tau\big(Y - \boldsymbol{X}_S^T \boldsymbol{\beta}_S^* - X_j h_{jS}\big)\big\}, \tag{4}$$

respectively. Then we have

$$\mathrm{E}\big\{\boldsymbol{X}_S \psi_\tau\big(Y - \boldsymbol{X}_S^T \boldsymbol{\beta}_S^*\big)\big\} = 0 \tag{5}$$

and

$$\mathrm{E}\big\{X_j \psi_\tau\big(Y - \boldsymbol{X}_S^T \boldsymbol{\beta}_S^* - X_j h_{jS}^*\big)\big\} = 0. \tag{6}$$

When $\mathcal{M} \subset S$, $\boldsymbol{\beta}_S^*$ is exactly from $\boldsymbol{\beta}^*$ as $\boldsymbol{X}_S$ from $\boldsymbol{X}$. However, if $\mathcal{M} \not\subset S$, $\boldsymbol{\beta}_S^*$ is not a subvector of $\boldsymbol{\beta}^*$. This is because this model does not include some relevant covariates and is misspecified.

The sample and population objective functions for $S \subset [p]$ are defined by

$$L_n(\boldsymbol{X}_S^T \boldsymbol{\beta}_S) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau\big(Y_i - \boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S\big) \tag{7}$$

and

$$L_S(\boldsymbol{\beta}_S) = \mathrm{E}\big\{L_n(\boldsymbol{X}_S^T \boldsymbol{\beta}_S)\big\} = \mathrm{E}\big\{\rho_\tau\big(Y - \boldsymbol{X}_S^T \boldsymbol{\beta}_S\big)\big\}, \tag{8}$$

respectively. We estimate $\boldsymbol{\beta}_S^*$ and h_{jS}^* by

$$\widehat{\boldsymbol{\beta}}_S = \operatorname*{arg\,min}_{\boldsymbol{\beta}_S \in \mathbb{R}^{|S|}} L_n\big(\boldsymbol{X}_S^T \boldsymbol{\beta}_S\big) \tag{9}$$

and

$$\widehat{h}_{jS} = \operatorname*{arg\,min}_{h_{jS} \in \mathbb{R}} L_n\big(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j h_{jS}\big), \tag{10}$$

respectively. We repeat that $\boldsymbol{\beta}_S^*$ is not a subvector of $\boldsymbol{\beta}^*$ if $\mathcal{M} \not\subset S$.

We examine the properties of $\widehat{\boldsymbol{\beta}}_S$ in Lemma 2 in Sect. 4 by taking the uniformity w.r.t. S into account. On the other hand, $\widehat{h}_{jS}$ is an auxiliary tool and we do not need any properties of $\widehat{h}_{jS}$ in proving our main theoretical results.

In our procedures, we minimize $L_n(\boldsymbol{X}_S^T\boldsymbol{\beta}_S)$ in (7) and do not use the sample version of (15) in Assumption LB. Therefore, extension to more general models like varying coefficient models are easy and straightforward. Besides, minimizing $L_n(\boldsymbol{X}_S^T\boldsymbol{\beta}_S)$ is directly linked to our stopping rule and we can present the theoretical results for our procedures in a unified way.

2.2 Forward selection procedures

We propose three procedures. We call the first one the full regression procedure (hereafter called FR) since it fully minimizes $L_n(\boldsymbol{X}_S^T\boldsymbol{\beta}_S)$ w.r.t. $\boldsymbol{\beta}_S$ at each step. This full regression procedure may take a little longer time if p is extremely large. The second one minimizes $L_n(\boldsymbol{X}_S\widehat{\boldsymbol{\beta}}_S + X_j h_{jS})$ w.r.t. h_{jS} at each step and will be suitable for extremely large p. We call the second one SC. The third one is a combination of the first two procedures, FR and SC, and chooses a set of candidates for full regression by using the results for $L_n(\boldsymbol{X}_S\widehat{\boldsymbol{\beta}}_S + X_j\widehat{h}_{jS})$. Our main theoretical results focus on the first two procedures. However, the third one enjoys the same properties as the first two as we state in Corollary 1 at the end of Sect. 2.3.

Hereafter we assume that we have a given upper bound K_n for the following procedures. As for ξ_n in (11), almost any ξ_n going to ∞ will work from a theoretical point of view and we set $\xi_n = \log\log n$ in the numerical studies. See (21) and (22) after Theorem 2 about the conditions on ξ_n and K_n.

- Full Regression Procedure (FR):
  Take $S_0 = \{1\}$ and begin with $k = 1$. We stop the procedure if (11) is not satisfied or $k = K_n$.

  (**a**) Set $S = S_{k-1}$. Then define j_k by
  $$j_k = \arg\min_{j\in S^c} \min_{\boldsymbol{\beta}_{S\cup\{j\}}} L_n\big(\boldsymbol{X}_{S\cup\{j\}}^T\boldsymbol{\beta}_{S\cup\{j\}}\big).$$

  (**b**) Check whether we have significantly improved $L_n(\boldsymbol{X}_{S_{k-1}}^T\widehat{\boldsymbol{\beta}}_{S_{k-1}})$ by adding j_k. Specifically, if we have with $S = S_{k-1}$,
  $$L_n\big(\boldsymbol{X}_S^T\widehat{\boldsymbol{\beta}}_S big) - \min_{j\in S^c}\min_{\boldsymbol{\beta}_{S\cup\{j\}}} L_n big(\boldsymbol{X}_{S\cup\{j\}}^T\boldsymbol{\beta}_{S\cup\{j\}} big) > \xi_n|S|\log p_n/n, \tag{11}$$
  set $S_k = S_{k-1}\cup\{j_k\}$ and go to (**a**). If not, set $\widehat{\mathcal{M}} = S_{k-1}$ and end this algorithm. Note that the second term on the LHS of (11) is actually given by $j = j_k$ defined in (**a**). See Remark 2 at the end of this subsection about the stopping rule defined in (11).

Next we propose a simpler and faster forward selection procedure by following Zheng et al. (2020) and call it the sequentially conditional procedure (hereafter called SC) as in Zheng et al. (2020).

- Sequentially Conditional Procedure (SC) : Replace j_k in (**a**) of FR with

$$j_k = \arg\min_{j \in S^c} \min_{h_{jS}} L_n\big(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j h_{jS}\big) \tag{12}$$

and

$$\min_{j \in S^c} \min_{\boldsymbol{\beta}_{S\cup\{j\}}} L_n\big(\boldsymbol{X}_{S\cup\{j\}}^T \boldsymbol{\beta}_{S\cup\{j\}}\big) \ \text{ with } \ \min_{\boldsymbol{\beta}_{S\cup\{j_k\}}} L_n\big(\boldsymbol{X}_{S\cup\{j_k\}}^T \boldsymbol{\beta}_{S\cup\{j_k\}}\big) \tag{13}$$

in (**b**) of FR.

In (12), we choose only the best index among S^c and denote it by j_k there. As in Lin et al. (2022), we can choose more indices for (**b**) based on $\min_{h_{jS}} L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j h_{jS}) = L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j \widehat{h}_{jS})$ for $j \in S^c$. In the next procedure, we choose m_0 indices for (**b**) as in (14) by using $L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j \widehat{h}_{jS})$ for $j \in S^c$. Then we minimize $L_n(\boldsymbol{X}_{S\cup\{j\}}^T \boldsymbol{\beta}_{S\cup\{j\}})$ w.r.t. $\boldsymbol{\beta}_{S\cup\{j\}}$ for every j in M_S defined in (14) and select the best $j \in M_S$ to be added to S_{k-1}.

We can also say that before full minimization w.r.t. $\boldsymbol{\beta}_{S\cup\{j\}}$ for all $j \in S^c$, we apply some preliminary screening to S^c in FR by using the results of $L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j \widehat{h}_{jS})$. Namely, we choose M_S in (14) from S^c with a kind of conditional SIS procedure as in Barut et al. (2016). Then we apply the same procedure as FR with S^c replaced with M_S. This is our greedier variant of SC (hereafter called gSC). See Remark 1 about more details of this procedure.

- A greedier variant of SC (gSC): First we fix an positive integer m_0 and denote gSC with this m_0 by gSC(m_0). Specifically, we construct the index set M_S for S by

$$\begin{aligned} M_S := \big\{ j \in S^c \mid & L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j \widehat{h}_{jS}) \text{ is among the smallest } m_0 \\ & \text{of all } \ L_n\big(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_\ell \widehat{h}_{\ell S}\big), \ \ell \in S^c. \big\} \end{aligned} \tag{14}$$

Then we replace S^c with this M_S in (**a**) and (**b**) of FR. Note that $|M_S| = m_0$. We repeat that $L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j \widehat{h}_{jS})$ is among the smallest m_0 of $\{L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_\ell \widehat{h}_{\ell S}) \mid \ell \in S^c\}$ if $j \in M_S$ and that $|M_S| = m_0$.

Two remarks are in place. Remark 1 is about our motivation to gSC(m_0) and m_0, and Remark 2 is about our stopping rule in (11).

Remark 1 SC, namely gSC(1), carries out one-dimensional quantile regression $(p - k - 1)$ times and only one $(k + 1)$-dimensional quantile regression for j_k as in (13) at the kth step. This is desirable in terms of computational time. However, in some situations, the second or third best index in (12) can actually be better in terms of full $(k + 1)$-dimensional quantile regression. Or the variable minimizing $L_n(\boldsymbol{X}_S^T \boldsymbol{\beta}_{S\cup\{j\}})$ for $j \in S^c$ will be very highly ranked in (**a**) of SC even if it is not j_k there and will be found in M_S for some reasonably large m_0. Thus, gSC(m_0) serves as a desirable combination of FR and SC in terms of computational time and minimization of $L_n(\boldsymbol{X}_S^T \boldsymbol{\beta}_{S\cup\{j\}})$. As we stated, selecting M_S from S^c is a kind of conditional SIS procedure as in Barut et al. (2016) instead of carrying out full minimization w.r.t. $\boldsymbol{\beta}_{S\cup\{j\}}$ for all $j \in S^c$. There

seems to be no optimality theory as to how many covariates should be selected for SIS (this is m_0 here) and $m_n = \lceil n/\log n \rceil$ is one of widely used practical choices for the number of selected covariates. See also a few lines after Steps (1)–(3) on p. 438 in Kong et al. (2019). The authors of Kong et al. (2019) cited three papers on SIS and introduced this choice. In addition, our numerical studies demonstrated that the idea of gSC(m_0) with $m_0 = m_n$ worked well numerically. From a theoretical point of view, this gSC(m_0) enjoys the same properties as the other two as we prove in Corollary 1 at the end of Sect. 2.3.

Remark 2 Note that $|S|$ on the RHS in (11) is the cost for dealing with the uniformity w.r.t. S as $K_n \to \infty$. This kind of uniformity w.r.t. S is necessary to dealing with sequential procedures rigorously and is often overlooked in the literature. When we consider mean regression models, estimators like $\widehat{\boldsymbol{\beta}}_S$ are explicitly available and we can establish the similar uniformity much more easily without $|S|$ in (11). If the upper bound K_n is bounded, we can remove this $|S|$ on the RHS in (11) from a theoretical point of view. In our numerical studies, our stopping rule tend to stop a little too early and we used some practical remedies to this early termination problem as described in Sect. 3.

2.3 Assumptions

Next we present our assumptions before we describe our theoretical results. For quantile regression models, explicit expressions of $\boldsymbol{\beta}_S^*$ and $\widehat{\boldsymbol{\beta}}_S$ are unavailable and we need assumptions like Assumption LB. Hereafter we consider only S satisfying $|S| \le K_n$.

We write $C_1, C_2, \ldots$ for generic positive constants and their values may vary from place to place. Contrary to $C_1, C_2, \ldots$, we use $D_1, D_2, \ldots$ in a similar way with their values fixed; namely, their values do not change from place to place. All of these constants are independent of n. We use $a_n \sim b_n$ for $\{a_n\}$ and $\{b_n\}$ when $a_n < C_1 < b_n$ and $b_n < C_2 < a_n$.

The following assumption is similar to Assumption (E) in Zheng et al. (2020) and this is our basic assumption. It stipulates how large the signal is when $\mathcal{M} \not\subset S$. If the LHS of (15) is small for any $j \in S^c$, the remaining signal is negligible and there will be no need of adding new covariates. In Theorem 1, we relate the LHS of (15) in Assumption LB to our $L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S)$, $L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j \widehat{h}_{jS})$, and $L_n(\boldsymbol{X}_{S\cup\{j\}}^T \widehat{\boldsymbol{\beta}}_{S\cup\{j\}})$.

Assumption LB There is a uniform lower bound κ_{LB} such that

$$\left|\mathrm{E}\left\{X_j \psi_\tau\left(Y - \boldsymbol{X}_S^T \boldsymbol{\beta}_S^*\right)\right\}\right| > \kappa_{LB} \tag{15}$$

for some $j \in \mathcal{M} \cap S^c$ if $\mathcal{M} \not\subset S$. Note that we allow κ_{LB} to decreases to 0 while satisfying the conditions in Theorems 1–3.

We prove in Lemma 1 in Sect. 4 that we can improve $L_S(\boldsymbol{\beta}_S^*)$ sufficiently by adding some $j \in S^c$ if $\mathcal{M} \not\subset S$. Then we use the above assumption together with some technical assumptions given below. This is because theoretical analysis of quantile regression models needs assumptions on conditional density functions and we also have to consider the uniformity in S. In the literature on screening procedures for high-dimensional models, results such as (ii) in Lemma 1 in Sect. 4 are often assumed.

We describe the other technical assumptions for our theoretical results here. We decided to present simpler assumptions and avoid complicated assumptions and these assumptions can be relaxed to some extent as we make comments in the proofs. Since the boundedness of $h^*_{jS}X_j$ is necessary in Assumption FY and the proof of Lemma 1, we state an assumption on this boundedness before Assumption FY on conditional density functions.

Assumption B

(1) $|X_j| < X_M$ uniformly in $j \in [p]$ for some positive constant X_M.
(2) $|h^*_{jS}| < D_h$ uniformly in $S\,(\mathcal{M} \not\subset S)$ and $j \in S^c$ for some positive constant D_h.

(1) of Assumption B can be relaxed a little. See Remark 3 after the proof of Lemma 1 in the supplement.

Assumption FY is about conditional density functions of Y. This kind of assumption is common in the literature on quantile regression. We denote the conditional density function of Y on some random vector $\boldsymbol{W}$ by $f_Y(y|\boldsymbol{W})$.

Assumption FY

(1) There are positive constants C_1 and C_2 and a small positive δ_1 such that

$$C_1 < f_Y\left(y|\left(\boldsymbol{X}_S^T, X_j\right)^T\right) < C_2 \quad \text{on} \quad \left(\boldsymbol{X}_S^T\boldsymbol{\beta}^*_S - D_hX_M - \delta_1, \boldsymbol{X}_S^T\boldsymbol{\beta}^*_S + D_hX_M + \delta_1\right) \tag{16}$$

uniformly in $S\,(\mathcal{M} \not\subset S)$ and $j \in S^c$.
(2) There are positive constants C_3 and C_4 and a small positive δ_2 such that

$$C_3 < f_Y(y|\boldsymbol{X}_S) < C_4 \quad \text{on} \quad \left(\boldsymbol{X}_S^T\boldsymbol{\beta}^*_S - \delta_2, \boldsymbol{X}_S^T\boldsymbol{\beta}^*_S + \delta_2\right) \tag{17}$$

uniformly in $S\,(\mathcal{M} \subset S)$. Recall that we have $\boldsymbol{X}_S^T\boldsymbol{\beta}^*_S = \boldsymbol{X}^T\boldsymbol{\beta}^*$ for such S. This assumption holds automatically if we have (17) with $S = [p]$. As for S such that $\mathcal{M} \not\subset S$, inequalities in (17) for such S follow from those in (16).
(3) For $S\,(\mathcal{M} \subset S)$, $f_Y(y|\boldsymbol{X}_S)$ is uniformly Lipshitz continuous in y on $(\boldsymbol{X}_S^T\boldsymbol{\beta}^*_S - \delta_2, \boldsymbol{X}_S^T\boldsymbol{\beta}^*_S + \delta_2)$, where δ_2 is the same as in (17).

We write $\lambda_m(A)$ and $\lambda_M(A)$ for the minimum and maximum eigenvalues of a symmetric matrix A, respectively. The next assumption is closely related to eigenvalues of $\boldsymbol{X}_S\boldsymbol{X}_S^T/n$ and inevitable to linear regression models.

Assumption X

(1) There are positive constants C_1 and C_2 such that

$$C_1 < \lambda_m\left(\mathrm{E}\left\{\boldsymbol{X}_S\boldsymbol{X}_S^T\right\}\right) \le \lambda_M\left(\mathrm{E}\left\{\boldsymbol{X}_S\boldsymbol{X}_S^T\right\}\right) < C_2$$

uniformly in S.
(2) There are positive constants C_3 and C_4 such that

$$C_3 < \lambda_m\left(\mathrm{E}\left\{\boldsymbol{X}_S\boldsymbol{X}_S^T f_Y\left(\boldsymbol{X}_S^T\boldsymbol{\beta}^*_S|\boldsymbol{X}_S\right)\right\}\right) \le \lambda_M\left(\mathrm{E}\left\{\boldsymbol{X}_S\boldsymbol{X}_S^T f_Y\left(\boldsymbol{X}_S^T\boldsymbol{\beta}^*_S|\boldsymbol{X}_S\right)\right\}\right) < C_4$$

uniformly in S ($\mathcal{M} \subset S$).

We can prove by using the standard arguments that the sample versions of Assumptions X(1)(2) hold with probability tending to 1 under Assumptions X(1)(2). Therefore, we also assume that the sample versions hold with probability tending to 1 for simplicity of presentation. We know Assumption X(2) follows from Assumptions FY(2) and X(1). However, we make Assumption X(2) an independent assumption for better understanding of the proof of Theorem 2.

2.4 Theoretical results

In this subsection, we state Theorems 1–3 which cover both FR and SC in a unified way. We deal with the theoretical properties of gSC(m_0) in Corollary 1 at the end of this subsection.

In the literature on non-iterative feature screening, authors usually take an association measure between Y and X_j, which we denote it by $\rho(Y, X_j)$ here. This $\rho(Y, X_j)$ often comes from marginal models like SIS. Then those authors make an assumption that $\rho(Y, X_j) > \eta_n$ if $j \in \mathcal{M}$ for some suitable η_n. Then they carry out feature screening or variable selection by estimating $\rho(Y, X_j)$ by some estimator denoted by $\rho_n(Y, X_j)$ here. They establish the screening consistency by proving that $\rho_n(Y, X_j) > \eta_n$ for $j \in \mathcal{M}$ with probability tending to 1. In practical situations, we have no idea about η_n and there is no optimality theory as to the number of selected covariates. Therefore, practical rules as in Remark 1 are often used as to how many covariates are selected for SIS-type feature screening. We can say choosing M_S from S^c in gSC(m_0) is a kind of SIS-type feature screening.

If $\mathcal{M} \not\subset S$, Theorem 1 relates Assumption LB to a sufficiently large improvement on $L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S)$. Since we deal with forward procedures for quantile regression models rigorously, there should be S in (15) of Assumption LB. Some technical assumptions are also necessary. Recall that the true coefficient is given by minimizing $L_S(\boldsymbol{\beta}_S)$ for $\mathcal{M} \subset S$.

Theorem 1 *Suppose that Assumptions LB, B*(1)(2), *FY*(1)(2), *and X*(1) *hold. Besides, setting* $S = S_{k-1}$ *for* $k < K_n$, *we assume that*

$$\frac{D_{LB}\kappa_{LB}^2}{2} \geq D_{U1}\left(\sqrt{\frac{|S|\log p_n}{n}} + \frac{|S|^2 \log p_n}{n}\right) + D_{U2}|S|\sqrt{\frac{\log p_n}{n}}, \tag{18}$$

where D_{LB}, D_{U1}, *and* D_{U2} *are given in Lemmas* 1, 3, *and* 4 *in Sect*. 4, *respectively. Then with probability tending to* 1, *we have uniformly in k smaller than* K_n,

$$L_n\left(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S\right) - \min_{j \in S^c} L_n\left(\boldsymbol{X}_{S\cup\{j\}}^T \widehat{\boldsymbol{\beta}}_{S\cup\{j\}}\right) \geq \frac{D_{LB}\kappa_{LB}^2}{2}$$

and

$$L_n\left(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S\right) - \min_{j \in S^c} L_n\left(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j \widehat{h}_{jS}\right) \geq \frac{D_{LB}\kappa_{LB}^2}{2}$$

with $S = S_{k-1}$ *if* $\mathcal{M} \not\subset S_{k-1}$.

The condition in (18) looks complicated. However, (19) is a simple sufficient condition for (18) and it allows ultra-high-dimensional cases.

$$K_n \sqrt{\frac{\log p_n}{n}} = o(\kappa_{LB}^2). \tag{19}$$

Since we propose forward procedures, we need a suitable stopping rule to save computational time and avoid large $|S|$. Even if $K_n < cn$ for some $c \in (0, 1)$, large-dimensional quantile regression may cause some computational problems. As our numerical studies demonstrate, we obtain models of reasonable size due to our stopping rule. In Theorem 2, we establish the theoretical validity of our stopping rule for FR and SC : our procedures do not stop until $\mathcal{M} \subset S_k$ and our procedures stop once $\mathcal{M} \subset S_k$.

Theorem 2 *Suppose that Assumptions LB, B*(1)(2), *FY*(1)(2)(3), *and X*(1)(2) *and* (18) *hold. As long as* $D_{LB}\kappa_{LB}^2 > \xi_n |S_k| \log p_n / n$ *and* $|S_k| < K_n$, *our algorithms FR and SC do not stop while* $\mathcal{M} \not\subset S_k$ *with probability tending to* 1. *Besides, assume*

$$K_n (\log n)^{7/2} \sqrt{\frac{\log p_n}{n}} = o(1). \tag{20}$$

Then once $\mathcal{M} \subset S_k$ *and* $|S_k| < K_n$, *our FR and SC procedures stop at this step with probability tending to* 1.

The first inequality in Theorem 2 is less restrictive than the inequality in (19). The inequality in (20) is similar to (19). Assuming that $\log p_n \sim n^{\gamma_1}$ and $\kappa_{LB} \sim n^{-\gamma_2}$ for some positive γ_1 and γ_2, we give some upper bounds on K_n and ξ_n here. Considering all of (18)–(20) and inequalities in Theorem 2, we obtain this sufficient condition :

$$K_n = o\left(n^{1/2-\gamma_1/2-2\gamma_2}\right) \quad \text{with} \ \ 1/2 - \gamma_1/2 - 2\gamma_2 > 0 \tag{21}$$

and

$$\xi_n K_n = o\left(n^{1-\gamma_1-2\gamma_2}\right) \quad \text{with} \ \ \xi_n \to \infty. \tag{22}$$

In our procedures, we choose only one variable at each step and we need an argument different from many other papers on feature screening to establish screening consistency. Reduction in $L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S)$ at each step should be large enough to find all the members in $\mathcal{M}$ before we reach the upper limit K_n. According to Theorem , both FR and SC enjoy the screening consistency if κ_{LB} in (15) is not very small. Recall that $S_0 = \{1\}$.

Theorem 3 *Suppose that Assumptions LB, B*(1)(2), *FY*(1)(2), *and X*(1) *and* (18) *hold and set*

$$\Delta = L_{S_0}\left(\boldsymbol{\beta}_{S_0}^*\right) - L_{[p]}(\boldsymbol{\beta}^*)$$

Then $\mathcal{M} \subset S_k$ *for some* $k < K_n$ *with probability tending to* 1 *if*

$$\frac{2\Delta}{D_{LB}\kappa_{LB}^2} < K_n - 2. \tag{23}$$

Finally we deal with gSC(m_0). For any m_0, gSC(m_0) has the same desirable properties as FR and SC as we show in Corollary 1. We prove Corollary 1 by exploiting the fact that gSC(m_0) is between FR and SC in terms of minimization.

Corollary 1 *We have the same results for gSC(m_0) as in Theorem 2 under the assumptions of Theorem 2. We also have the same results for gSC(m_0) as in Theorem 3 under the assumptions of Theorem 3.*

3 Numerical studies

In this section, we evaluate the finite sample performances of the proposed procedures through simulation studies and an application to a real gene expression data set. We carried out all the computations by using R.

3.1 Simulation studies

In this subsection, we assess the finite sample performances of gSC(1), gSC(m_0), and FR with or without the stopping rule (11). As mentioned in Remark 1, a larger value of m_0 will borrow more strength from FR at the cost of additional computation; thus, aside from gSC(1), we consider gSC(25) and gSC(m_n) with $m_n = \lceil n/\log n \rceil$ for comparison. This m_n is commonly used in the literature of feature screening and seems moderate under the setting $n = 400$ (so that $m_n = 67$) considered in this subsection.

To terminate the proposed algorithms, we adopt the following rules:
We terminate the algorithm if (11) is not satisfied i times consecutively or the iteration number achieves K_n. The former criterion based on (11) is denoted as T_i, and the proposed algorithms using T_i are denoted as gSC(1)+T_i, gSC(25)+T_i, gSC(m_n)+T_i, and FR+T_i. On the other hand, the algorithms achieving K_n iterations with no stopping rule are denoted as just gSC(1), gSC(25), gSC(m_n), and FR. We tried $i = 1, 2$, and 3 for T_i. Note that FR+T_1 means exactly the full regression procedure in Sect. 2.2, and T_2 and T_3 are our remedies for preventing the early stopping or termination. The same kind of practical rule is also adopted in Cheng et al. (2016) for the same purpose.

The proposed procedures are compared to the penalized quantile regression models with the Lasso penalty (Belloni and Chernozhukov, 2011), the adaptive Lasso (ALasso) penalty, and the non-convex SCAD and MCP penalties. Their tuning parameter λ is chosen by minimizing the BIC for penalized quantile regression (QBIC(λ)) (Lee et al., 2014)

$$\log\left(L_n\left(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}}_\lambda\right)\right) + |\widehat{\boldsymbol{\beta}}_\lambda| C_n \log n/(2n), \tag{24}$$

where $\widehat{\boldsymbol{\beta}}_\lambda$ is the penalized quantile regression estimator with respect to λ. In Lee et al. (2014), the authors suggest using $C_n = \log p_n$ for the purpose of variable selection consistency. Since the goal of this study is to choose a small set of variables with the

sure screening property, we use $C_n = 1/\log n$ for Lasso and use $C_n = \log\log p_n$ for SCAD, MCP and ALasso. Note that the weight of ALasso is determined by Lasso.

In addition to the penalized regression methods, we also compare with the marginal screening method using the conditional quantile utility (CQU) of Wu and Yin (2015) and the associated forward regression using the partial quantile utility (FR-PQU) of Kong et al. (2019). The number of variables selected by CQU is set as $\lceil n/\log(n) \rceil$ following the recommendation of the authors, and the number of iteration for FR-PQU is set as K_n to make a fair comparison with our proposed ones without applying the stopping rule. In addition to FR-PQU, we also choose the model with minimum QBIC(S) from the K_n nested models generated by FR-PQU as suggested in Section 3.2.1. of Kong et al. (2019). This QBIC (S) is defined by

$$\log\left(L_n\left(\boldsymbol{X}_S^T\widehat{\boldsymbol{\beta}}_S\right)\right) + |S|C_n \log n/(2n).$$

We take $C_n = \log\log p_n$ so that the criterion QBIC(S) is comparable to QBIC(λ) for penalized regression models, and denote such combination as FR-PQU+QBIC. It is expected that FR-PQU+QBIC can largely reduce the false positives from FR-PQU but remains containing all relevant variables, so we can say that this FR-PQU+QBIC is the counterpart of our proposed procedures using the stopping rule (11). Note that FR-PQU and FR-PQU+QBIC considered in this subsection correspond to QFR and QFR+QBIC3 in Kong et al. (2019), respectively.

For implementation, we use the package *quantreg* of Koenker (2021) to solve (12) and (13) for our proposed methods, and the package *rqPen* (Sherwood and Maidman, 2020) for penalized regression methods.

We deal with three examples, each of which has three different levels of quantile: $\tau = 0.3, 0.5$ and 0.7. A total of 100 simulation replications are carried out with $(n, p) = (400, 1000)$, and, taken from $n = 400$, the maximum iteration number K_n for the forward-type algorithms is set as 30. The detailed settings for the design matrix $\boldsymbol{X}$, the coefficient vector $\boldsymbol{\beta}^*$, and the error distribution of ϵ are listed below.

Example 1 The response Y is obtained by

$$Y = 1 + 1.5X_7 + 0.7X_{13} + X_{16} - 0.5X_{21} + (1+\gamma)X_2\epsilon,$$

where ϵ follows the t-distribution with degree of freedom 3 and the parameter γ measures the heteroscedasticity. The predictor vector $(X_1, \ldots, X_{p+1})$ is set as $X_1 = 1$ and $X_j = \tilde{X}_j$ for $j \geq 2$, where $(\tilde{X}_2, \ldots, \tilde{X}_{p+1})$ follows the multivariate t-distribution $N_p(\mathbf{0}, \Sigma)$ with degree of freedom 3 and $\Sigma_{jk} = 0.5^{|j-k|}$. The γ is set as $\gamma = 0.5$ so that the quantile coefficient of X_2 is around -0.292 at $\tau = 0.3$, exactly 0 at $\tau = 0.5$, and around 0.292 at $\tau = 0.7$, respectively. Thus, we have $\mathcal{M} = \{7, 13, 16, 21\}$ for $\tau = 0.5$ and $\mathcal{M} = \{2, 7, 13, 16, 21\}$ for $\tau \neq 0.5$. In this example, we deal with unbounded X_j, weak signals on regression coefficients, and the heteroscedastic error terms to check the numerical robustness of our procedures.

Example 2 Adopted from Wu and Yin (2015), the response Y is obtained by

$$Y = X_2 + X_3 + X_4 + X_5 + X_6 + \exp(X_{21})\epsilon,$$

where ϵ follows from a Cauchy distribution. The predictor vector $(X_1, \ldots, X_{p+1})$ is set to X_2 and $\{X_j\}_{j\geq 2}$ follow the multivariate normal distribution $N_p(0, \Sigma)$ with $\Sigma_{jk} = 0.5$ for $j \neq k$ and $\Sigma_{jj} = 1$ for $j, k = 2, \ldots, p+1$. In this example, we have $\mathcal{M} = \{2, 3, 4, 5, 6\}$ for $\tau = 0.5$ but $\mathcal{M} = \{2, 3, 4, 5, 6, 21\}$ for $\tau \neq 0.5$.

Example 3 The response Y is obtained from

$$Y = 2X_2 + 2X_3 + 2X_4 + \epsilon,$$

where ϵ follows the t-distribution with degrees of freedom 3. The predictor vector is generated as follows: $X_1 = 1, X_2 = W_2 - W_3 - W_4, X_3 = W_3 - W_4, X_4 = 2W_4$, and $X_j = W_4 + U_j$ for $j \geq 5$, where variables in $\{W_2, \ldots, W_4, U_5, \ldots, U_{p+1}\}$ are independently generated from $N(0, 1)$. Given this specification, both X_3 and X_4 are uncorrelated with the response Y, but are correlated with irrelevant variables $\{X_j\}_{j\geq 5}$ through W_4 (with $Cor(X_3, X_j) = -0.5$ and $Cor(X_4, X_j) = 0.707$). Both X_3 and X_4 are called marginally weak variables since they are almost impossible to be detected by marginal screening methods like CQU; in addition, the interference of correlation made by irrelevant variables also adds the difficulty in variable screening.

In order to evaluate the performances of each screening procedure, we write $\widehat{\mathcal{M}}^{(b)}$ for the index set constructed by one particular method in the bth simulation replication. Note that we exclude the intercept term $\{1\}$ from $\widehat{\mathcal{M}}^{(b)}$ when evaluating its performance. Based on $\{\widehat{\mathcal{M}}^{(b)}\}_{b=1}^{100}$ and $\mathcal{M}$, we measure the frequency of sure screening (Sure), the averaged number of true positives (TP), and the averaged number of false positives (FP) defined by

$$\text{Sure} = \sum_{b=1}^{100} I\{\mathcal{M} \subset \widehat{\mathcal{M}}^{(b)}\}, \text{TP} = 100^{-1} \sum_{b=1}^{100} \left|\mathcal{M} \bigcap \widehat{\mathcal{M}}^{(b)}\right| \text{ and FP} = 100^{-1} \sum_{b=1}^{100} \left|\mathcal{M}^c \bigcap \widehat{\mathcal{M}}^{(b)}\right|,$$

respectively. Additionally, the averaged computing time (Time) in seconds of each procedure is also recorded.

The simulation results of Example 1–3 are summarized in Tables 1, 2 and 3, from which we make the following observations:

(i) In Tables 1, 2 and 3, gSC(25)+T_i, gSC(m_n)+T_i, and FR+T_i with $i = 2, 3$ give quite satisfactory performances in the balance of high Sure and low FP among all the proposed methods. In particular, gSC(m_n)+T_3 and FR+T_3 compare favorably to CQU, ALasso, SCAD and MCP because gSC(m_n)+T_3 and FR+T_3 have both higher Sure values and lower FP values. Note that only gSC(m_n)+T_i and FR+T_i with $i = 2, 3$ detect X_4 in Table 3 satisfactorily. Besides, just FR+T_3 has perfect Sure values in Table 3.

(ii) FR-PQU+QBIC performs similarly to gSC(m_n)+T_i and FR+T_i with $i = 2, 3$ in Table 1. However, its performances deteriorate seriously in Tables 2 and 3. It seems that QBIC(S) does not work well in Tables 2 and 3 while our T_2 and T_3 work reasonably well. As for FR-PQU, which has no model selection or stopping rule, it works slightly better for X_2 in Table 1 and shows similar performances to gSC(m_n)+T_3

Table 1 Simulation results for Example 1 with $(n, p) = (400, 1000)$

	X_2	X_7	X_{13}	X_{16}	X_{21}	Sure	TP	FP	Time
$\tau = 0.3, \mathcal{M} = \{2, 7, 13, 16, 21\}$									
CQU	5	100	100	100	83	5	3.88	63.12	3.82
Lasso	54	100	100	100	100	54	4.54	49.32	23.62
ALasso	21	100	99	100	100	21	4.20	0.66	23.95
SCAD	18	100	99	99	100	18	4.16	0.35	30.80
MCP	28	100	99	100	99	28	4.26	0.63	31.66
gSC(1)+T_1	8	100	98	100	97	8	4.03	0.12	4.59
gSC(1)+T_2	48	100	100	100	100	48	4.48	0.67	5.71
gSC(1)+T_3	52	100	100	100	100	52	4.52	1.63	6.77
gSC(1)	58	100	100	100	100	58	4.58	25.42	25.75
gSC(25)+T_1	8	100	98	100	97	8	4.03	0.12	4.70
gSC(25)+T_2	52	100	99	100	99	52	4.50	0.65	5.85
gSC(25)+T_3	58	100	100	100	100	58	4.58	1.57	6.97
gSC(25)	59	100	100	100	100	59	4.59	25.41	28.36
gSC(m_n)+T_1	8	100	98	100	97	8	4.03	0.12	4.93
gSC(m_n)+T_2	52	100	99	100	99	52	4.50	0.65	6.11
gSC(m_n)+T_3	58	100	100	100	100	58	4.58	1.57	7.33
gSC(m_n)	63	100	100	100	100	63	4.63	25.37	36.53
FR+T_1	8	100	98	100	97	8	4.03	0.12	5.33
FR+T_2	52	100	99	100	99	52	4.50	0.65	6.77
FR+T_3	59	100	100	100	100	59	4.59	1.56	8.35
FR	61	100	100	100	100	61	4.61	25.39	57.43
FR-PQU+QBIC	63	100	100	100	100	63	4.63	0.63	114.38
FR-PQU	81	100	100	100	100	81	4.81	25.19	114.38
$\tau = 0.5, \mathcal{M} = \{7, 13, 16, 21\}$									
CQU	5	100	100	100	81	81	3.81	63.19	3.84
Lasso	8	100	100	100	100	100	4.00	34.91	11.94
ALasso	0	100	100	100	100	100	4.00	0.05	12.21
SCAD	0	100	100	100	100	100	4.00	0.01	13.65
MCP	0	100	100	100	100	100	4.00	0.06	15.90
gSC(1)+T_1	1	100	100	100	99	99	3.99	0.49	5.02
gSC(1)+T_2	1	100	100	100	100	100	4.00	1.48	6.07
gSC(1)+T_3	2	100	100	100	100	100	4.00	2.48	7.21
gSC(1)	6	100	100	100	100	100	4.00	26.00	25.66
gSC(25)+T_1	2	100	100	100	99	99	3.99	0.49	5.16
gSC(25)+T_2	3	100	100	100	100	100	4.00	1.48	6.33
gSC(25)+T_3	5	100	100	100	100	100	4.00	2.48	7.47
gSC(25)	8	100	100	100	100	100	4.00	26.00	28.55
gSC(m_n)+T_1	2	100	100	100	99	99	3.99	0.49	5.42
gSC(m_n)+T_2	3	100	100	100	100	100	4.00	1.48	6.63
gSC(m_n)+T_3	5	100	100	100	100	100	4.00	2.48	7.82
gSC(m_n)	7	100	100	100	100	100	4.00	26.00	37.03

Table 1 (continued)

	X_2	X_7	X_{13}	X_{16}	X_{21}	Sure	TP	FP	Time
FR+T_1	2	100	100	100	99	99	3.99	0.49	6.07
FR+T_2	3	100	100	100	100	100	4.00	1.48	7.69
FR+T_3	5	100	100	100	100	100	4.00	2.48	9.39
FR	8	100	100	100	100	100	4.00	26.00	58.49
FR-PQU+QBIC	2	100	100	100	100	100	4.00	0.27	115.05
FR-PQU	9	100	100	100	100	100	4.00	26.00	115.05
$\tau = 0.7, \mathcal{M} = \{2, 7, 13, 16, 21\}$									
CQU	16	100	100	100	80	13	3.96	63.04	3.76
Lasso	62	100	100	100	99	62	4.61	50.13	25.88
ALasso	20	100	100	100	99	20	4.19	0.37	26.20
SCAD	16	100	100	100	100	16	4.16	0.22	29.91
MCP	26	100	100	100	100	26	4.26	0.31	33.96
gSC(1)+T_1	5	100	99	100	98	5	4.02	0.06	4.44
gSC(1)+T_2	40	100	100	100	100	40	4.40	0.68	5.56
gSC(1)+T_3	47	100	100	100	100	47	4.47	1.61	6.65
gSC(1)	61	100	100	100	100	61	4.61	25.39	25.13
gSC(25)+T_1	6	100	99	100	97	5	4.02	0.06	4.62
gSC(25)+T_2	48	100	100	100	100	48	4.48	0.60	5.70
gSC(25)+T_3	55	100	100	100	100	55	4.55	1.53	6.85
gSC(25)	64	100	100	100	100	64	4.64	25.36	27.75
gSC(m_n)+T_1	6	100	99	100	97	5	4.02	0.06	4.82
gSC(m_n)+T_2	48	100	100	100	100	48	4.48	0.60	5.99
gSC(m_n)+T_3	55	100	100	100	100	55	4.55	1.53	7.14
gSC(m_n)	63	100	100	100	100	63	4.63	25.37	35.45
FR+T_1	6	100	99	100	97	5	4.02	0.06	5.38
FR+T_2	48	100	100	100	100	48	4.48	0.60	6.92
FR+T_3	55	100	100	100	100	55	4.55	1.53	8.51
FR	64	100	100	100	100	64	4.64	25.36	57.64
FR-PQU+QBIC	67	100	100	100	100	67	4.67	0.63	111.59
FR-PQU	85	100	100	100	100	85	4.85	25.15	111.59

and FR+T_3 in the other results except for X_4 in Table 3. Note that FR-PQU fails to detect X_4 in Table 3 and that it has much higher FP values because it selects exactly K_n covariates. The most significant difference between FR-PQU and our { gSC(m_n), FR} is whether we carry out full minimization w.r.t. $\boldsymbol{\beta}_{S\cup\{j\}}$. We think that the differences in the results for X_4 in Table 3 come from this full minimization w.r.t. $\boldsymbol{\beta}_{S\cup\{j\}}$ in gSC(m_n) + T_3 and $FR + T_3$. As we describe in (iii), increasing m_0 improves the performances of gSC(m_0)+T_i significantly. Recall that we carry out full minimization w.r.t. $\boldsymbol{\beta}_{S\cup\{j\}}$ m_0 times at each step.

(iii) As for m_0, the performances of gSC(25)+T_i are better than those of gSC(1)+T_i in Tables 2 and 3 and almost the same as those of gSC(m_n)+T_i and FR+T_i in Tables 1

Table 2 Simulation results for Example 2 with $(n, p) = (400, 1000)$

	X_2	X_3	X_4	X_5	X_6	X_{21}	Sure	TP	FP	Time
$\tau = 0.3, \mathcal{M} = \{2, 3, 4, 5, 6, 21\}$										
CQU	84	79	81	88	82	1	1	4.15	62.85	3.77
Lasso	94	92	93	95	93	6	6	4.73	19.43	22.68
ALasso	72	69	70	76	68	0	0	3.55	0.10	23.32
SCAD	55	59	54	62	54	2	1	2.86	0.30	49.94
MCP	55	59	57	60	56	2	2	2.89	0.28	65.04
gSC(1)+T_1	67	61	63	76	54	1	0	3.22	0.40	3.22
gSC(1)+T_2	82	75	80	80	79	11	0	4.07	0.55	4.12
gSC(1)+T_3	86	84	85	88	87	35	20	4.65	0.97	4.99
gSC(1)	98	97	94	97	98	73	65	5.57	24.43	26.10
gSC(25)+T_1	70	63	74	74	70	1	0	3.52	0.13	3.35
gSC(25)+T_2	89	85	87	88	92	7	0	4.48	0.17	4.25
gSC(25)+T_3	93	92	93	95	96	56	44	5.25	0.40	5.17
gSC(25)	99	99	98	100	100	80	79	5.76	24.24	27.49
gSC(m_n)+T_1	70	63	74	74	70	1	0	3.52	0.13	3.50
gSC(m_n)+T_2	89	85	87	88	92	7	0	4.48	0.17	4.46
gSC(m_n)+T_3	93	92	93	95	96	56	44	5.25	0.40	5.43
gSC(m_n)	99	99	98	100	100	81	80	5.77	24.23	35.49
FR+T_1	70	63	74	74	70	1	0	3.52	0.13	3.78
FR+T_2	89	85	87	88	92	7	0	4.48	0.17	4.97
FR+T_3	93	92	93	95	96	56	44	5.25	0.40	6.18
FR	99	99	98	100	100	82	80	5.78	24.22	59.49
FR-PQU+QBIC	54	50	49	58	51	2	1	2.64	0.59	113.97
FR-PQU	96	97	95	98	98	76	68	5.60	24.40	113.97
$\tau = 0.5, \mathcal{M} = \{2, 3, 4, 5, 6\}$										
CQU	87	84	84	89	87	5	48	4.31	62.69	3.37
Lasso	99	99	99	99	98	0	98	4.94	22.53	16.63
ALasso	91	89	90	92	90	0	86	4.52	0.01	16.78
SCAD	84	81	77	83	81	0	68	4.06	0.02	28.15
MCP	83	83	77	84	83	0	67	4.10	0.03	41.12
gSC(1)+T_1	83	81	80	85	80	0	12	4.09	0.03	3.70
gSC(1)+T_2	100	100	100	99	98	0	97	4.97	0.15	4.59
gSC(1)+T_3	100	100	100	100	100	0	100	5.00	1.12	5.48
gSC(1)	100	100	100	100	100	5	100	5.00	25.00	25.96
gSC(25)+T_1	83	75	81	86	77	0	5	4.02	0.03	3.74
gSC(25)+T_2	100	99	99	100	99	0	97	4.97	0.08	4.64
gSC(25)+T_3	100	100	100	100	100	0	100	5.00	1.05	5.55
gSC(25)	100	100	100	100	100	6	100	5.00	25.00	27.09
gSC(m_n)+T_1	83	75	81	86	77	0	5	4.02	0.03	3.90
gSC(m_n)+T_2	100	99	99	100	99	0	97	4.97	0.08	4.89
gSC(m_n)+T_3	100	100	100	100	100	0	100	5.00	1.05	5.85
gSC(m_n)	100	100	100	100	100	5	100	5.00	25.00	31.34

Table 2 (continued)

	X_2	X_3	X_4	X_5	X_6	X_{21}	Sure	TP	FP	Time
FR+T_1	83	75	81	86	77	0	5	4.02	0.03	4.17
FR+T_2	100	99	99	100	99	0	97	4.97	0.08	5.37
FR+T_3	100	100	100	100	100	0	100	5.00	1.05	6.65
FR	100	100	100	100	100	5	100	5.00	25.00	58.93
FR-PQU+QBIC	83	78	75	77	76	0	60	3.89	0.35	120.00
FR-PQU	100	100	100	100	100	4	100	5.00	25.00	120.00
$\tau = 0.7, \mathcal{M} = \{2, 3, 4, 5, 6, 21\}$										
CQU	77	63	72	75	72	25	2	3.84	63.16	3.22
Lasso	96	97	95	96	97	72	71	5.53	34.52	21.82
ALasso	86	78	82	82	84	12	8	4.24	0.03	22.13
SCAD	67	61	66	71	65	9	5	3.39	0.37	48.04
MCP	69	59	65	71	62	7	2	3.33	0.59	63.61
gSC(1)+T_1	79	63	68	78	71	8	0	3.67	0.32	3.57
gSC(1)+T_2	92	84	87	93	87	9	0	4.52	0.47	4.46
gSC(1)+T_3	93	91	94	96	96	34	27	5.04	0.95	5.34
gSC(1)	98	96	99	99	98	54	52	5.44	24.56	25.58
gSC(25)+T_1	82	69	73	80	78	7	0	3.89	0.11	3.67
gSC(25)+T_2	96	89	94	99	96	8	0	4.82	0.18	4.59
gSC(25)+T_3	98	97	100	100	99	62	59	5.56	0.44	5.51
gSC(25)	100	99	100	100	100	72	71	5.71	24.29	26.79
gSC(m_n)+T_1	82	69	73	80	78	7	0	3.89	0.11	3.84
gSC(m_n)+T_2	96	89	94	99	96	8	0	4.82	0.18	4.83
gSC(m_n)+T_3	98	97	100	100	99	62	59	5.56	0.44	5.77
gSC(m_n)	100	99	100	100	100	73	72	5.72	24.28	34.83
FR+T_1	82	69	73	80	78	7	0	3.89	0.11	4.18
FR+T_2	96	89	94	99	96	8	0	4.82	0.18	5.37
FR+T_3	98	97	100	100	99	62	59	5.56	0.44	6.62
FR	100	99	100	100	100	72	71	5.71	24.29	57.85
FR-PQU+QBIC	60	55	59	60	58	6	1	2.98	1.03	110.50
FR-PQU	96	97	96	97	97	53	50	5.36	24.64	110.50

and 2. In Table 3, gSC(m_n)+T_i outperforms gSC(25)+T_i and gSC(m_n)+T_3 works as well as FR+T_2 and FR+T_3. Increasing m_0 improves the performances of gSC(m_0) largely and the commonly used practical rule for m_0, $m_n = \lceil n/\log n \rceil$, seems to be a reasonable choice.

All of (i)–(iii) and the columns of computational time imply that both gSC(m_n) + T_i and FR+T_i with $i = 2, 3$ work well in terms of performances and computational time. Therefore, we recommend them and gSC(m_n)+T_i may be suitable for extremely p. We present additional simulation results for $(n, p) = (400, 4000)$ in the supplement and those results also confirm this conclusion. Our remedies for preventing early stopping and the commonly used practical rule for m_0 also show good finite sample properties in the additional simulation results.

Table 3 Simulation results for Example 3 with $(n, p) = (400, 1000)$

	X_2	X_3	X_4	Sure	TP	FP	Time
$\tau = 0.3, \mathcal{M} = \{2, 3, 4\}$							
CQU	100	7	1	0	1.08	65.92	3.63
Lasso	100	100	53	53	2.53	90.64	26.80
ALasso	100	100	40	40	2.40	11.55	27.78
SCAD	100	94	36	36	2.30	7.36	64.65
MCP	100	100	6	6	2.06	10.31	73.44
gSC(1)+T_1	100	46	4	4	1.50	1.54	2.71
gSC(1)+T_2	100	99	4	4	2.03	2.01	3.59
gSC(1)+T_3	100	100	4	4	2.04	3.00	4.47
gSC(1)	100	100	4	4	2.04	27.96	25.75
gSC(25)+T_1	100	84	47	47	2.31	1.16	3.16
gSC(25)+T_2	100	100	57	57	2.57	2.10	4.27
gSC(25)+T_3	100	100	58	58	2.58	3.12	5.20
gSC(25)	100	100	58	58	2.58	27.42	26.92
gSC(m_n)+T_1	100	85	51	51	2.36	1.15	3.35
gSC(m_n)+T_2	100	100	80	80	2.80	2.29	4.90
gSC(m_n)+T_3	100	100	85	85	2.85	3.39	5.99
gSC(m_n)	100	100	85	85	2.85	27.15	30.25
FR+T_1	100	85	51	51	2.36	1.15	3.61
FR+T_2	100	100	85	85	2.85	2.34	5.60
FR+T_3	100	100	100	100	3.00	3.64	7.48
FR	100	100	100	100	3.00	27.00	58.88
FR-PQU+QBIC	100	100	1	1	2.01	12.91	115.77
FR-PQU	100	100	5	5	2.05	27.95	115.77
$\tau = 0.5, \mathcal{M} = \{2, 3, 4\}$							
CQU	100	2	0	0	1.02	65.98	3.37
Lasso	100	100	88	88	2.88	103.71	27.93
ALasso	100	100	79	79	2.79	6.64	29.34
SCAD	100	96	74	74	2.70	3.92	52.88
MCP	100	100	19	19	2.19	8.86	66.53
gSC(1)+T_1	100	41	5	5	1.46	1.59	2.73
gSC(1)+T_2	100	98	5	5	2.03	2.02	3.61
gSC(1)+T_3	100	100	5	5	2.05	3.00	4.50
gSC(1)	100	100	5	5	2.05	27.95	25.72
gSC(25)+T_1	100	86	64	64	2.50	1.14	3.33
gSC(25)+T_2	100	100	65	65	2.65	2.01	4.27
gSC(25)+T_3	100	100	69	69	2.69	3.09	5.30
gSC(25)	100	100	69	69	2.69	27.31	26.81
gSC(m_n)+T_1	100	88	68	68	2.56	1.12	3.54
gSC(m_n)+T_2	100	100	86	86	2.86	2.18	4.86
gSC(m_n)+T_3	100	100	93	93	2.93	3.32	6.02
gSC(m_n)	100	100	95	95	2.95	27.05	29.79

Table 3 (continued)

	X_2	X_3	X_4	Sure	TP	FP	Time
FR+T_1	100	88	68	68	2.56	1.12	3.75
FR+T_2	100	100	88	88	2.88	2.20	5.39
FR+T_3	100	100	100	100	3.00	3.44	7.12
FR	100	100	100	100	3.00	27.00	58.10
FR-PQU+QBIC	100	100	0	0	2.00	13.25	116.58
FR-PQU	100	100	8	8	2.08	27.92	116.58
$\tau = 0.7, \mathcal{M} = \{2, 3, 4\}$							
CQU	100	4	0	0	1.04	65.96	3.22
Lasso	100	100	54	54	2.54	91.55	30.34
ALasso	100	100	38	38	2.38	12.00	31.41
SCAD	100	96	40	40	2.36	8.71	67.03
MCP	100	100	6	6	2.06	10.29	77.78
gSC(1)+T_1	100	38	7	7	1.45	1.62	2.76
gSC(1)+T_2	100	100	7	7	2.07	2.00	3.66
gSC(1)+T_3	100	100	7	7	2.07	3.00	4.59
gSC(1)	100	100	7	7	2.07	27.93	25.49
gSC(25)+T_1	100	56	34	34	1.90	1.44	3.11
gSC(25)+T_2	100	100	41	41	2.41	2.07	4.17
gSC(25)+T_3	100	100	45	45	2.45	3.15	5.19
gSC(25)	100	100	46	46	2.46	27.54	26.64
gSC(m_n)+T_1	100	67	47	47	2.14	1.33	3.36
gSC(m_n)+T_2	100	100	62	62	2.62	2.15	4.63
gSC(m_n)+T_3	100	100	79	79	2.79	3.49	6.12
gSC(m_n)	100	100	81	81	2.81	27.19	29.95
FR+T_1	100	68	48	48	2.16	1.32	3.61
FR+T_2	100	100	68	68	2.68	2.20	5.27
FR+T_3	100	100	100	100	3.00	3.81	7.80
FR	100	100	100	100	3.00	27.00	57.99
FR-PQU+QBIC	100	100	0	0	2.00	13.22	119.53
FR-PQU	100	100	8	8	2.08	27.92	119.53

3.2 Real data analysis

In this subsection, we consider a gene expression dataset reported in Bühlmann et al. (2014) to illustrate the performances of the proposed methods. The data contains 71 independent samples ($n = 71$), from which the logarithm of 4088 gene expression levels ($p = 4088$) and of a response variable riboflavin (vitamin B2) production rate in Bacillus subtilis are measured. This dataset is available in R package *hdi*. Apart from the fact that the number of variables greatly exceeds the number of observations, there are 5.17% out of $\binom{4088}{2}$ pairs whose correlation is greater than 0.7 in absolute value.

Thus, the main objective is now to select predictive genes for different quantiles of the riboflavin production rate using the methods considered in Sect. 3.1, where a small part out of high-dimensional gene expressions are possibly co-expressed. Before analysis, all the genes are rescaled to have mean 0 and variance 1.

To evaluate the prediction performance, we randomly partitioned the 71 samples into two disjoint sets: the training set of size 50 and the testing set of size 21. Under this $n << p$ circumstance, we adopt a two-stage procedure to select relevant genes based on the training set: a screening method in {CQU, gSC(1)+T_3, gSC(m_n)+T_3, FR+T_3, FR-PQU+QBIC} followed by a regularization method in {Lasso, SCAD, MCP}. The proposed procedures gSC(1)+T_3, gSC(m_n)+T_3, and FR+T_3 all have the sure screening property under the assumptions of our Theorems 1–3 as shown in Sect. 2, but they showed different behaviors in our simulation studies and can have their own advantages as screening procedures. Thus, we further consider the method called "Hybrid," which is defined as the union of variable sets selected by gSC(1)+T_3, gSC(m_n)+T_3, and FR+T_3. We call the screening–regularization pairs CQU+Lasso, CQU+SCAD, CQU+MCP, and so on. We also present the results of screening only, namely just gSC(1)+T_3, gCS(m_n)+T_3, FR+T_3, Hybrid, and FR-PQU+QBIC for reference in Table 4. Table 9 in the supplementary material is the counterpart of Table 4 for screening methods with no stopping rule or model selection, namely gSC(1), gSC(m_n), FR, and FR-PQU. In Table 9, SC(1), gSC(m_n), FR, and FR-PQU are followed by regularization methods like Lasso. In addition, we implement Lasso and ALasso, where the weight in ALasso is determined by Lasso. The testing set is applied to evaluate the prediction error (PE) of the τth conditional quantile, defined by $(21)^{-1}\sum_{i=1}^{21}\rho_\tau(y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})$, where $\widehat{\boldsymbol{\beta}}$ is the estimator of $\boldsymbol{\beta}$ obtained from the training set. This procedure is repeated for 50 times, and the median of PE as well as the median model size (Size) are reported in Table 4.

We first observe that, at each level of $\tau, \tau \in \{0.3, 0.5, 0.7\}$, the proposed method in {gSC(1)+T_3, gSC(m_n)+T_3, FR+T_3} has lower PE than FR-PQU+QBIC, and FR-PQU+QBIC have lower PE than CQU, CQU+Lasso, CQU+SCAD, and CQU+MCP. The comparatively high PE values for CQU-based methods partly result from the requirement of independence assumption, which is not satisfied in this dataset. Second, by taking $m_n = \lceil 50/\log(50) \rceil = 13$, both PE and Size values of gSC(m_n)+T_3 is close to that of FR+T_3, and is comparable to that of gSC(1)+T_3: the PE value of gSC(m_n)+T_3 is lower (higher) than that of gSC(1)+T_3 at $\tau = 0.3$ and 0.7 (at $\tau = 0.5$). We further observe that the results of screening only do not seem to be significantly different from those combined with a regularization method in terms of PE and Size. Finally, the Hybrid+Lasso method has the smallest PE value at $\tau = 0.3$, the ALasso method has the smallest PE value at $\tau = 0.5$, and the Hybrid method has the smallest PE value at $\tau = 0.7$. Note that Hybrid has slightly larger Size values but has apparently smaller PE values than those of gSC(1)+T_3, gSC(m_n)+T_3, or FR+T_3. It means the methods in {gSC(1)+T_3, gSC(m_n)+T_3, FR+T_3} yield similar results, and the predictive genes missed by one can be selected by the others.

Next, we proceed our analysis by comparing genes selected from methods in Table 4 based on the full data ($n = 71$). Note that the follow-up regularization step in {Lasso, SCAD, MCP} to the forward-type screening method in {gSC(1)+T_3,

Table 4 Prediction analysis for the gene expression dataset. Values in parentheses are estimated standard deviation

Screen	Regularization	$\tau = 0.3$		$\tau = 0.5$		$\tau = 0.7$	
		PE	Size	PE	Size	PE	Size
Lasso	ALasso	0.226 (0.060)	4 (1.7)	0.225 (0.047)	4 (1.7)	0.217 (0.056)	4 (1.7)
CQU	Lasso	0.278 (0.067)	5 (2.5)	0.264 (0.075)	5 (2.5)	0.243 (0.048)	5 (2.5)
	SCAD	0.275 (0.057)	3 (1.6)	0.263 (0.054)	3 (1.6)	0.241 (0.044)	3 (1.6)
	MCP	0.281 (0.058)	3 (1.3)	0.259 (0.056)	3 (1.3)	0.241 (0.042)	3 (1.3)
gSC(1)+T_3	–	0.251 (0.049)	4 (0.0)	0.239 (0.054)	4 (0.0)	0.211 (0.038)	4 (0.0)
	Lasso	0.249 (0.049)	4 (0.0)	0.238 (0.054)	4 (0.0)	0.211 (0.038)	4 (0.0)
	SCAD	0.248 (0.049)	4 (0.0)	0.240 (0.054)	4 (0.1)	0.211 (0.038)	4 (0.1)
	MCP	0.249 (0.049)	4 (0.1)	0.240 (0.054)	4 (0.1)	0.211 (0.039)	4 (0.1)
gSC(m_n)+T_3	–	0.235 (0.067)	4 (0.0)	0.247 (0.052)	4 (0.0)	0.200 (0.048)	4 (0.0)
	Lasso	0.237 (0.067)	4 (0.0)	0.244 (0.051)	4 (0.0)	0.202 (0.048)	4 (0.0)
	SCAD	0.237 (0.066)	4 (0.0)	0.245 (0.051)	4 (0.1)	0.199 (0.048)	4 (0.1)
	MCP	0.238 (0.067)	4 (0.1)	0.245 (0.051)	4 (0.1)	0.200 (0.048)	4 (0.1)
FR+T_3	–	0.238 (0.070)	4 (0.0)	0.247 (0.047)	4 (0.0)	0.202 (0.047)	4 (0.0)
	Lasso	0.238 (0.070)	4 (0.1)	0.244 (0.047)	4 (0.1)	0.204 (0.047)	4 (0.1)
	SCAD	0.238 (0.070)	4 (0.1)	0.245 (0.047)	4 (0.1)	0.201 (0.047)	4 (0.1)
	MCP	0.238 (0.070)	4 (0.1)	0.246 (0.047)	4 (0.1)	0.202 (0.047)	4 (0.1)
Hybrid	–	0.218 (0.059)	7 (1.1)	0.237 (0.050)	7 (1.1)	0.187 (0.040)	7 (1.1)
	Lasso	0.216 (0.060)	6 (1.2)	0.239 (0.049)	6 (1.2)	0.191 (0.041)	6 (1.2)
	SCAD	0.217 (0.058)	5 (1.1)	0.241 (0.047)	5 (1.1)	0.200 (0.041)	5 (1.1)
	MCP	0.216 (0.059)	5 (1.1)	0.240 (0.047)	5 (1.1)	0.200 (0.041)	5 (1.1)
FR-PQU+QBIC	–	0.278 (0.066)	6 (2.7)	0.253 (0.050)	6 (2.7)	0.220 (0.039)	6 (2.7)
	Lasso	0.277 (0.066)	6 (2.6)	0.255 (0.050)	6 (2.6)	0.218 (0.039)	6 (2.6)
	SCAD	0.276 (0.067)	6 (2.5)	0.254 (0.048)	6 (2.5)	0.217 (0.040)	6 (2.5)
	MCP	0.275 (0.065)	6 (2.5)	0.257 (0.050)	6 (2.5)	0.217 (0.040)	6 (2.5)

gSC(m_n)+ T_3, FR+T_3, FR-PQU+QBIC} does not remove any gene, so we only present their screening results in Table 5. Since the genes YXLC, YXLD and YXLE are located in the same operon and they are highly correlated with the gene YXLJ, the genes in set {YXLC, YXLD, YXLE, YXLJ} are likely to be co-expressed and involved in a similar cellular functions. We denote that the set {YXLC, YXLD, YXLE, YXLJ} is selected if at least one gene within this set is selected by a specific method. The gene sets {XHLA ,XHLB, XTRA} and {ARGF,ARGJ} are denoted in a similar manner. We present the correlation in Table 10 in the supplemantary material and the complete result in Table 5, and refer the readers to *SubtiWiki* at http://www.subtiwiki.uni-goettingen.de/ for more details about the annotation of genes and operon in Bacillus subtilis.

As shown in Table 5, the genes sets {YXLC, YXLD, YXLE, YXLJ} and {XHLA, XHLB, XTRA} are selected by all the methods except for FR-PQU at all considered

Table 5 A comparison of genes selected by various methods

Gene	ALasso	CQU+SCAD	CQU+MCP	gSC(1)+T_3	gSC(m)+T_3	FR+T_3	FR-PQU+QBIC
$\tau = 0.3$							
ARGJ					√	√	
PHRI_r		√					
YCGN	√						
YHZA							√
YKOC					√	√	
YTRP				√			
YOAB		√	√	√			
YXLC				√			
YXLD		√	√		√	√	
YXLE	√						
YXLJ							
XHLA				√	√	√	
XHLB	√						
$\tau = 0.5$							
IOLA				√	√	√	
XHLB	√						
XTRA				√	√	√	
YCGN		√	√				
YDDR					√	√	
YTGB	√	√	√				
YTOQ				√			
YVNB							√
YWFO		√	√				
YXLD	√	√	√		√	√	
YXLJ				√			
$\tau = 0.7$							
DPPC							√
IOLA				√	√	√	
XKDM	√						
XTRA				√	√	√	
YCGN				√	√	√	
YCKE		√	√				
YDAO		√					
YKUH	√						
YTGB							
YXLD		√					
YXLE	√		√				
YXLJ				√	√	√	

levels of τ, the gene set {ARGF, ARGJ} is selected by gSC(m_n)+T_3 and FR+T_3 at $\tau = 0.3$. The importance of these gene sets have been certified in Bühlmann et al. (2014) and Das et al. (2019) based on mean regression models, where the gene YXLD and the gene ARGF have been discovered associated with the riboflavin production rate directly, and the gene XHLA has been identified as a stable gene (potentially) having a causal effect on the riboflavin production rate. We also observe that the gene YCGN is selected by ALasso at $\tau = 0.3$, by CQU_SCAD and CQU_MCP at $\tau = 0.5$, and by our methods at $\tau = 0.7$; the gene IOLA is only identified by our methods at $\tau = 0.5$ and $\tau = 0.7$. These genes have been overlooked in the literature of using the mean regression models to analyze the riboflavin dataset and may deserve more attention for further study.

4 Assumptions and proofs

In this section, we prove Theorems 1–3 and Corollary 1. Before the proofs, we present technical lemmas for the proofs of Theorems 1–3 and Corollary 1. The technical lemmas are verified in the supplement.

Recall that $|S|$ and $|S \cup \{j\}|$ are less than or equal to K_n in this section and the supplement, too.

4.1 Technical lemmas

In this subsection, we state technical lemmas for the proofs of Theorems 1–3 and Corollary 1. The proofs of these lemmas are given in the supplement.

Lemma 1 relates Assumption LB in Sect. 2 to the improvement in $L_S(\boldsymbol{\beta}_S)$ when $\mathcal{M} \not\subset S$.

Lemma 1 *Suppose that Assumptions LB, B*(1)(2), *and FY*(1) *hold. Then there are positive constants* D_1, D_2, *and* D_{LB} *for* (*i*) *and* (*ii*).

(*i*) *If* $\mathcal{M} \not\subset S$, *we have for some* $j \in \mathcal{M} \cap S^c$,

$$|h^*_{jS}| \geq D_1 \kappa_{LB}.$$

(*ii*) *If* $\mathcal{M} \not\subset S$, *we have for some* $j \in \mathcal{M} \cap S^c$,

$$\mathrm{E}\{\rho_\tau(Y - \boldsymbol{X}_S^T \boldsymbol{\beta}^*_S)\} - \mathrm{E}\{\rho_\tau(Y - \boldsymbol{X}_S^T \boldsymbol{\beta}^*_S - X_j h^*_{jS})\} \geq D_2 |h^*_{jS}|^2 \geq D_{LB} \kappa^2_{LB}.$$

In Lemma 2, we consider the uniform convergence rate of $\widehat{\boldsymbol{\beta}}_S$ in (9). Note that $|S|^{1/2}$ in (25) is the cost for the uniformity in S. The proof of this lemma is based on the standard arguments in the literature on quantile regression. For example, see the proofs of Theorem 1 in Fan et al. (2014), Proposition 1 in Honda et al. (2019), and Lemma C in Kong et al. (2019). However, none of them deals with the uniformity in S for $|S| \leq K_n$ or gives the rate such as given in (25) for quantile regression models.

Lemma 2 *Suppose that Assumptions B*(1)(2), *FY*(1)(2), *and X*(1) *hold. Then we have for some positive constant* D_R,

$$\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\| \leq D_R |S|^{1/2} \sqrt{\frac{|S| \log p_n}{n}} \tag{25}$$

uniformly in S with probability tending to 1.

In Lemma 3, we evaluate the difference between $L_S(\boldsymbol{\beta}_S^*)$ and its estimator uniformly in S such that $\mathcal{M} \not\subset S$. Lemma 4 deals with a similar problem. The rate in Lemma 4 dominates that in Lemma 3.

Lemma 3 *Suppose that Assumptions B*(1)(2) *and FY*(1) *hold. Then we have for some positive constant* D_{U1},

$$\left|L_n\left(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S\right) - L_S\left(\boldsymbol{\beta}_S^*\right)\right| \leq D_{U1} \left(\sqrt{\frac{|S| \log p_n}{n}} + \frac{|S|^2 \log p_n}{n} \right) \tag{26}$$

uniformly in S ($\mathcal{M} \not\subset S$) *with probability tending to* 1.

As we mentioned earlier, the properties of $\widehat{h}_{jS}$ are not necessary to the proofs of Theorems 1–3 and Corollary 1 and we deal with h_{jS}^* in Lemma 4.

Lemma 4 *Suppose that Assumptions B*(1)(2), *FY*(1), *and X*(1) *hold. Then we have for some positive constant* D_{U2},

$$|L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S + X_j h_{jS}^*) - L_S((\boldsymbol{\beta}_S^{*T}, h_{jS}^*)^T)| \leq D_{U2} |S| \sqrt{\frac{\log p_n}{n}} \tag{27}$$

uniformly in S ($\mathcal{M} \not\subset S$) *and* $j \in S^c$ *with probability tending to* 1.

4.2 Proofs of Theorems 1–3

We prove Theorems 1–3 by using Lemmas 1–4. We put the proof of Theorem 2 after that of Theorem 3 because that of Theorem 2 is long and complicated. We verify Theorem 2 by following the proof of Theorem 2 of Honda et al. (2019) and the proof of Theorem 2 of Honda and Lin (2021). The former deals with the cases where K_n is bounded and the uniformity in S is trivial. The latter is about generalized varying coefficient models, not quantile regression models.

Corollary 1 follows from the proofs of Theorems 2 and 3 by just noting two inequalities and the uniformity w.r.t. S. We give the proof at the end of this subsection.

Proof of Theorem 1 Recall that $\mathcal{M} \not\subset S$ in this theorem. Then by the definitions of $\widehat{\boldsymbol{\beta}}_S$ and $\widehat{h}_{jS}$, we have

$$L_n\left(X_{S\cup\{j\}}^T\widehat{\boldsymbol{\beta}}_{S\cup\{j\}}\right) \le L_n\left(X_S^T\widehat{\boldsymbol{\beta}}_S + X_j\widehat{h}_{jS}\right) \le L_n\left(X_S^T\widehat{\boldsymbol{\beta}}_S + X_j h_{jS}^*\right) \tag{28}$$

uniformly in $j \in S^c$ and S.

By Lemma 4, we have with probability tending to 1,

$$L_n\left(X_S^T\widehat{\boldsymbol{\beta}}_S + X_j h_{jS}^*\right) \le L_{S\cup\{j\}}\left(\left(\boldsymbol{\beta}_S^{*T}, h_{jS}^*\right)^T\right) + D_{U2}|S|\sqrt{\frac{\log p_n}{n}} \tag{29}$$

uniformly in $j \in S^c$ and S.

By Lemma 1, we have for some $j \in \mathcal{M} \cap S^c$,

$$L_{S\cup\{j\}}\left(\left(\boldsymbol{\beta}_S^{*T}, h_{jS}^*\right)^T\right) \le L_S\left(\boldsymbol{\beta}_S^*\right) - D_{LB}\kappa_{LB}^2. \tag{30}$$

With probability tending to 1, we have for that j,

$$\begin{aligned} L_n\left(X_{S\cup\{j\}}^T\widehat{\boldsymbol{\beta}}_{S\cup\{j\}}\right) &\le L_n\left(X_S^T\widehat{\boldsymbol{\beta}}_S + X_j\widehat{h}_{jS}\right) \le L_S\left(\boldsymbol{\beta}_S^*\right) - D_{LB}\kappa_{LB}^2 + D_{U2}|S|\sqrt{\frac{\log p_n}{n}} \\ &\le L_n\left(X_S^T\widehat{\boldsymbol{\beta}}_S\right) - D_{LB}\kappa_{LB}^2 + D_{U2}|S|\sqrt{\frac{\log p_n}{n}} \\ &\quad + D_{U1}\left(\sqrt{\frac{|S|\log p_n}{n}} + \frac{|S|^2\log_n}{n}\right). \end{aligned} \tag{31}$$

We used (28)–(30) and Lemma 3 here.

By combining (31) and the assumption in (18), we obtain

$$L_n\left(X_{S\cup\{j\}}^T\widehat{\boldsymbol{\beta}}_{S\cup\{j\}}\right) \le L_n\left(X_S^T\widehat{\boldsymbol{\beta}}_S + X_j\widehat{h}_{jS}\right) \le L_n\left(X_S^T\widehat{\boldsymbol{\beta}}_S\right) - \frac{D_{LB}\kappa_{LB}^2}{2}. \tag{32}$$

Hence, the proof is complete. □

Proof of Theorem 3 Notice that if $k \le K_n$ and $\mathcal{M} \not\subset S_k$, we have by Lemma 3,

$$L_{S_k}\left(\boldsymbol{\beta}_{S_k}^*\right) - \frac{D_{LB}\kappa_{LB}^2}{2} \le L_n\left(X_{S_k}^T\widehat{\boldsymbol{\beta}}_{S_k}\right) \le L_n\left(X_{S_0}^T\widehat{\boldsymbol{\beta}}_{S_0}\right) \le L_{S_0}\left(\boldsymbol{\beta}_{S_0}^*\right) + \frac{D_{LB}\kappa_{LB}^2}{2} \tag{33}$$

and

$$L_n(X_{S_0}^T\widehat{\boldsymbol{\beta}}_{S_0}) - L_n(X_{S_k}^T\widehat{\boldsymbol{\beta}}_{S_k}) \ge k\frac{D_{LB}\kappa_{LB}^2}{2}. \tag{34}$$

By (33) and (34), we have

$$(k-2)\frac{D_{LB}\kappa_{LB}^2}{2} \le L_{S_0}\left(\boldsymbol{\beta}_{S_0}^*\right) - L_{S_k}\left(\boldsymbol{\beta}_{S_k}^*\right) \le \Delta \tag{35}$$

if $k \le K_n$ and $\mathcal{M} \not\subset S_k$. If $k = K_n$, this contradicts the assumption in (23). Then $\mathcal{M} \subset S_l$ for some $l < K_n$.

Hence, the proof is complete. □

Proof of Theorem 2 The result for S_k such that $\mathcal{M} \not\subset S_k$ follows from Theorem 1 straightforwardly.

We concentrate on the cases where $\mathcal{M} \subset S_k$ and we prove that our algorithms stop once $\mathcal{M} \subset S_k$. The proof consists of two steps.

Hereafter we drop the subscript k since we have to consider all the S containing $\mathcal{M}$. Then $\boldsymbol{\beta}_S^*$ is exactly a subvector of $\boldsymbol{\beta}^*$ and we have for such S,

$$\boldsymbol{X}_S^T \boldsymbol{\beta}_S^* = \boldsymbol{X}^T \boldsymbol{\beta}^* \quad \text{and} \quad \epsilon = Y - \boldsymbol{X}_S^T \boldsymbol{\beta}_S^*. \tag{36}$$

Step 1 First we derive an explicit expression of $L_n(\boldsymbol{X}_S^T \widehat{\boldsymbol{\beta}}_S) - L_n(\boldsymbol{X}_S^T \boldsymbol{\beta}_S^*)$ in (50) and we prove the desired result for S such that $\mathcal{M} \subset S$ by exploiting the expression.

To get the expression in (50), we closely examine

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n G\big(\boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S\big) &:= L_n\big(\boldsymbol{X}_S^T \boldsymbol{\beta}_S\big) - L_n\big(\boldsymbol{X}_S^T \boldsymbol{\beta}_S^*\big) + \frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_{iS}^T \big(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\big) \{\tau - I(\epsilon_i \le 0)\} \\ &\quad - \mathrm{E}_\epsilon \big\{ L_n\big(\boldsymbol{X}_S^T \boldsymbol{\beta}_S\big) - L_n\big(\boldsymbol{X}_S^T \boldsymbol{\beta}_S^*\big) \big\}, \end{aligned} \tag{37}$$

where $G(\cdot)$ is clearly defined, $\mathrm{E}_\epsilon\{\cdot\}$ is the conditional expectation of $\{\epsilon_1, \ldots, \epsilon_n\}$ on $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, and $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \le \sqrt{\eta_n |S| \log p_n / n}$. We specify η_n going to ∞ later in this proof. Note that $\sqrt{\eta_n |S| \log p_n / n}$ is large enough here although this can be smaller than the rate in Lemma 2. Notice that

$$G\big(\boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S\big) = \overline{G}\big(\boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S\big) - \mathrm{E}_\epsilon \big\{ \overline{G}\big(\boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S\big) \big\}, \tag{38}$$

where $b_i = \boldsymbol{X}_{iS}^T (\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)$ and

$$\begin{aligned} \overline{G}\big(\boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S\big) &:= \rho_\tau\big(Y_i - \boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S\big) - \rho_\tau\big(Y_i - \boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S^*\big) + b_i \{\tau - I(\epsilon_i \le 0)\} \\ &= -(\epsilon_i - b_i)\{I(\epsilon_i \le b_i) - I(\epsilon_i \le 0)\}. \end{aligned}$$

We evaluate $n^{-1} \sum_{i=1}^n G(\boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S)$ by repeated use of Bernstein's inequality. Before we apply the inequality, note

$$\max_{1 \le i \le n} \|\boldsymbol{X}_S\| \le |S|^{1/2} X_M, \tag{39}$$

$$\max_{1 \le i \le n} |\overline{G}\big(\boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S\big)| \le 2|b_i| \le 2|S| X_M (n^{-1} \eta_n \log p_n)^{1/2}, \tag{40}$$

and

$$\begin{aligned} \sum_{i=1}^m \mathrm{E}_\epsilon \big[\big\{ \overline{G}\big(\boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S\big) \big\}^2 \big] &\le C_1 \sum_{i=1}^n |b_i|^3 \\ &\le C_2 |S| X_M (n^{-1} \eta_n \log p_n)^{1/2} \times n (n^{-1} \eta_n |S| \log p_n) \\ &\le C_3 |S|^2 X_M n^{-1/2} (\log p_n)^{3/2} \eta_n^{3/2}, \end{aligned} \tag{41}$$

where C_1, C_2, and C_3 are suitable positive constants. We used Assumption FY(2) in evaluating $\mathrm{E}_\epsilon[\{\overline{G}(\boldsymbol{X}_{iS}^T\boldsymbol{\beta}_S)\}^2]$ and Assumption X(1) in evaluating $\sum_{i=1}^n |b_i|^3$.

By (40), (41), and Bernstein's inequality (Lemma 2.2.9 of van der Vaart and Wellner (1996)), we have for any fixed $\boldsymbol{\beta}_S$ satisfying $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \le \sqrt{\eta_n |S| \log p_n / n}$,

$$\mathrm{P}_\epsilon\left(\left| \frac{1}{n} \sum_{i=1}^n G\big(\boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S\big) \right| \ge \frac{|S| \log p_n}{\zeta_n n} \right) \le 2 \exp\left\{ - \frac{(\log p_n)^{1/2} n^{1/2}}{3 \zeta_n^2 \eta_n^{3/2} X_M} \right\} \tag{42}$$

where $\mathrm{P}_\epsilon(\cdot)$ is the conditional probability of $\{\epsilon_1, \ldots, \epsilon_n\}$ on $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ and we specify ζ_n going to ∞ later in the proof.

Recall that in Assumption X(1), we also assume that the sample version holds uniformly with probability tending to 1. This means that (42) is true when the sample version of Assumption X(1) holds and that we have (42) uniformly with probability tending to 1.

To establish the uniformity in $\boldsymbol{\beta}_S$ and S, we exploit the small-block argument and divide the region $\{\boldsymbol{\beta}_S \in \mathbb{R}^{|S|} \mid \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \le \sqrt{\eta_n |S| \log p_n / n}\}$ into sufficient small blocks. Then the number of such small blocks are less than $n^{D_{SB}|S|}$ for some large fixed D_{SB}. If

$$\sum_{q \le K_n} \exp\{(D_{SB}+1) q \log p_n\} \exp\left\{ - \frac{(\log p_n)^{1/2} n^{1/2}}{3 \zeta_n^2 \eta_n^{3/2} X_M} \right\} \to 0, \tag{43}$$

we can establish a uniform evaluation of (37) in both S and $\boldsymbol{\beta}_S$. A sufficient condition of (43) is

$$K_n (\log p_n)^{1/2} \eta_n^{3/2} \zeta_n^2 X_M = o(n^{1/2}). \tag{44}$$

This is satisfied with $\eta_n = \zeta_n = \log n$ due to (20). Hereafter we take $\eta_n = \zeta_n = \log n$.

Hence, we have uniformly in both $\boldsymbol{\beta}_S$ and S,

$$\begin{aligned} L_n\big(\boldsymbol{X}_S^T \boldsymbol{\beta}_S\big) - L_n\big(\boldsymbol{X}_S^T \boldsymbol{\beta}_S^*\big) = & -\big(\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\big)^T \frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_{iS} \{\tau - I(\epsilon_i \le 0)\} \\ & + \mathrm{E}_\epsilon\big\{ L_n\big(\boldsymbol{X}_S^T \boldsymbol{\beta}_S\big) - L_n\big(\boldsymbol{X}_S^T \boldsymbol{\beta}_S^*\big) \big\} + O_p\left(\frac{|S| \log p_n}{\zeta_n n} \right). \end{aligned} \tag{45}$$

We calculate the conditional expectation in (45) by employing Knight's identity (see (59) in the supplement) with $u = Y_i - \boldsymbol{X}_{iS}^T \boldsymbol{\beta}_S^*$ and $v = \boldsymbol{X}_{iS}^T (\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*)$ there. Then by following the standard argument in the quantile regression literature and also using Assumption FY(2)(3), we obtain uniformly in both $\boldsymbol{\beta}_S$ and S,

$$
\begin{aligned}
&\mathrm{E}_\epsilon\left\{L_n\left(\boldsymbol{X}_S^T\boldsymbol{\beta}_S\right)-L_n\left(\boldsymbol{X}_S^T\boldsymbol{\beta}_S^*\right)\right\}\\
&\quad=\mathrm{E}_\epsilon\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_{iS}^T\left(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\right)\{I(\epsilon_i\le 0)-\tau\}\right]\\
&\qquad+\mathrm{E}_\epsilon\left[\frac{1}{n}\sum_{i=1}^n\int_0^{\boldsymbol{X}_{iS}^T\left(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\right)}\left\{I\left(Y_i-\boldsymbol{X}_{iS}^T\boldsymbol{\beta}_S^*\le s\right)-I\left(Y_i-\boldsymbol{X}_{iS}^T\boldsymbol{\beta}_S^*\le 0\right)\right\}\mathrm{d}s\right]\\
&\quad=\frac{1}{2}\left(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\right)^T\widehat{\Sigma}_S\left(\boldsymbol{\beta}_S-\boldsymbol{\beta}_S^*\right)+O_p\left(\frac{|S|\log p_n}{n\zeta_n}\right),
\end{aligned}
\tag{46}
$$

where

$$
\widehat{\Sigma}_S:=\frac{1}{n}\sum_{i=1}^n f_Y\left(\boldsymbol{X}_{iS}^T\boldsymbol{\beta}_S^*|\boldsymbol{X}_{iS}\right)\boldsymbol{X}_{iS}\boldsymbol{X}_{iS}^T.
$$

Define $\boldsymbol{a}_S$ by

$$
\boldsymbol{a}_S:=\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_{iS}\{\tau-I(\epsilon_i\le 0)\}.
$$

By applying Bernstein's inequality componentwise, we have

$$
\|\boldsymbol{a}_S\|^2=O_p\left(\frac{|S|\log p_n}{n}\right) \tag{47}
$$

uniformly in S. We define $\overline{\boldsymbol{\beta}}_S$ by

$$
\overline{\boldsymbol{\beta}}_S:=\boldsymbol{\beta}_S^*+\widehat{\Sigma}_S^{-1}\boldsymbol{a}_S
$$

and consider $\boldsymbol{\delta}_S$ satisfying

$$
\|\overline{\boldsymbol{\beta}}_S+\boldsymbol{\delta}_S-\boldsymbol{\beta}_S^*\|\le\sqrt{\frac{\eta_n|S|\log p_n}{n}}. \tag{48}
$$

Note that $\boldsymbol{\delta}_S=0$ satisfies (48) due to (47) and Assumption X(2).

We put $\overline{\boldsymbol{\beta}}_S+\boldsymbol{\delta}_S$ and (46) into (45) and then we obtain uniformly in S,

$$
\begin{aligned}
&L_n\left(\boldsymbol{X}_S^T\left(\overline{\boldsymbol{\beta}}_S+\boldsymbol{\delta}_S\right)\right)-L_n\left(\boldsymbol{X}_S^T\boldsymbol{\beta}_S^*\right)\\
&\qquad=-\frac{1}{2}\boldsymbol{a}_S^T\widehat{\Sigma}_S^{-1}\boldsymbol{a}_S+\frac{1}{2}\boldsymbol{\delta}_S^T\widehat{\Sigma}_S\boldsymbol{\delta}_S+O_p\left(\frac{|S|\log p_n}{\zeta_n n}\right).
\end{aligned}
\tag{49}
$$

By the optimality of $\widehat{\boldsymbol{\beta}}_S$, the convexity of $L_n(\boldsymbol{X}_S^T\boldsymbol{\beta}_S)$, and (49), we obtain uniformly in S,

$$
L_n\left(\boldsymbol{X}_S^T\widehat{\boldsymbol{\beta}}_S\right)-L_n\left(\boldsymbol{X}_S^T\boldsymbol{\beta}_S^*\right)=-\frac{1}{2}\boldsymbol{a}_S^T\widehat{\Sigma}_S^{-1}\boldsymbol{a}_S+O_p\left(\frac{|S|\log p_n}{\zeta_n n}\right). \tag{50}
$$

Step 2 By exploiting the expression in (50), we can proceed as in Step 3 of the proof of Theorem 2 in Honda and Lin (2021) although the paper deals with generalized varying coefficient models. We borrow the notation from the paper.

Hereafter we write $S_+ = S \cup \{j\}$ for S and j such that $\mathcal{M} \subset S$, $|S| < K_n$, and $j \notin S$. Then we evaluate $L_n(\widehat{\boldsymbol{\beta}}_{S_+}) - L_n(\widehat{\boldsymbol{\beta}}_S)$ by using (50). Here note that $L_n(\boldsymbol{X}_{S_+}^T \boldsymbol{\beta}_{S_+}^*) = L_n(\boldsymbol{X}_S^T \boldsymbol{\beta}_S^*)$ since $\mathcal{M} \subset S$.

We write

$$\widehat{\Sigma}_{S_+} = \begin{pmatrix} \widehat{\Sigma}_S & \widehat{\boldsymbol{\sigma}}_{jS}^T \\ \widehat{\boldsymbol{\sigma}}_{jS} & \widehat{\sigma}_{jj} \end{pmatrix} \quad \text{and} \quad \boldsymbol{a}_{S_+} = \begin{pmatrix} \boldsymbol{a}_S \\ a_j \end{pmatrix}. \tag{51}$$

Note that $\widehat{\sigma}_{jj} \in \mathbb{R}$ and $\widehat{\boldsymbol{\sigma}}_{jS}^T \in \mathbb{R}^{|S|}$.

The expression in (50) shows we have only to closely examine

$$\boldsymbol{a}_{S_+}^T \widehat{\Sigma}_{S_+}^{-1} \boldsymbol{a}_{S_+} - \boldsymbol{a}_S^T \widehat{\Sigma}_S^{-1} \boldsymbol{a}_S = (\widehat{\boldsymbol{\sigma}}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{a}_S)^2 \widehat{\sigma}_S^{jj} - 2\widehat{\boldsymbol{\sigma}}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{a}_S a_j \widehat{\sigma}_S^{jj} + a_j^2 \widehat{\sigma}_S^{jj} \tag{52}$$

where $\widehat{\sigma}_S^{jj} = (\widehat{\sigma}_{jj} - \widehat{\boldsymbol{\sigma}}_{jS} \widehat{\Sigma}_S^{-1} \widehat{\boldsymbol{\sigma}}_{jS}^T)^{-1}$.

If we show that the RHS of (52) has the stochastic order of $|S| O_p(n^{-1} \log p_n)$ uniformly in S and j, Theorem 2 for S_k containing $\mathcal{M}$ follows from this fact and (50). This is because the RHS of (11) is stochastically larger than (52).

Assumption X(2) implies that

$$\widehat{\sigma}_S^{jj} = O_p(1) \quad \text{and} \quad \|\widehat{\boldsymbol{\sigma}}_{jS}^T\| = O_p(1) \tag{53}$$

uniformly in j and S.

Besides, we recall that $a_j^2 = O_p(n^{-1} \log p_n)$ uniformly in j as in (47). This and (53) imply that the third term of the RHS of (52) has the stochastic order of $O_p(n^{-1} \log p_n)$ uniformly in S and j.

Next we deal with the first and second terms of the RHS of (52). Then we should evaluate $\widehat{\boldsymbol{\sigma}}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{a}_S$, which is rewritten as

$$\widehat{\boldsymbol{\sigma}}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{a}_S = \frac{1}{n} \sum_{i=1}^n \widehat{\boldsymbol{\sigma}}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{X}_{iS} \{\tau - I(\epsilon_i \le 0)\}. \tag{54}$$

Since by Assumption X(1)(2) we have for some positive constants C_4, C_5, and C_6,

$$|\widehat{\boldsymbol{\sigma}}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{X}_{iS}| \le C_4 X_M \|\widehat{\boldsymbol{\sigma}}_{jS}^T\| |S|^{1/2}$$

and

$$\frac{1}{n} \sum_{i=1}^n |\widehat{\boldsymbol{\sigma}}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{X}_{iS}|^2 \le C_5 \|\widehat{\boldsymbol{\sigma}}_{jS}^T\|^2 \le C_6$$

uniformly in i, j, and S with probability tending to 1, we employ the standard argument based on Bernstein's inequality conditionally on $\widehat{\boldsymbol{\sigma}}_{jS} \widehat{\Sigma}_S^{-1} \boldsymbol{X}_{iS}$ and obtain

$$\widehat{\boldsymbol{\sigma}}_{jS}\widehat{\Sigma}_S^{-1}\boldsymbol{a}_S = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\sigma}}_{jS}\widehat{\Sigma}_S^{-1}\boldsymbol{X}_{iS}\{\tau - I(\epsilon_i \le 0)\} = O_p(\{n^{-1}|S|\log p_n\}^{1/2}) \tag{55}$$

uniformly in m, S, and j.

Note that $|S|$ in $O_p(\{(nL)^{-1}|S|\log p_n\}^{1/2})$ above is necessary because $\widehat{\boldsymbol{\sigma}}_{jS}\widehat{\Sigma}_S^{-1}\boldsymbol{X}_{iS}$ depends on i, j, and S and we have to take into account all S and j satisfying $\mathcal{M} \subset S$, $|S| < K_n$, and $j \notin S$. The same kind of argument is also given in the proof of Lemma 2 in the supplement.

It follows from (55) that the first and second terms of the RHS of (52) has the stochastic order of $|S|O_p(n^{-1}\log p_n)$ uniformly in S and j.

Hence, the proof of Theorem 2 is complete. □

Proof of Corollary 1 If $S = S_{k-1}$ is common to FR, SC, and gSC(m_0), we have at the kth step :

$$\min_{j\in S^c}\min_{\boldsymbol{\beta}_{S\cup\{j\}}} L_n\big(\boldsymbol{X}_{S\cup\{j\}}^T\boldsymbol{\beta}_{S\cup\{j\}}\big) \le \min_{j\in M_S}\min_{\boldsymbol{\beta}_{S\cup\{j\}}} L_n\big(\boldsymbol{X}_{S\cup\{j\}}^T\boldsymbol{\beta}_{S\cup\{j\}}\big) \text{ and} \tag{56}$$

$$\min_{j\in M_S}\min_{\boldsymbol{\beta}_{S\cup\{j\}}} L_n\big(\boldsymbol{X}_{S\cup\{j\}}^T\boldsymbol{\beta}_{S\cup\{j\}}\big) \le \min_{\boldsymbol{\beta}_{S\cup\{j_k\}}} L_n\big(\boldsymbol{X}_{S\cup\{j_k\}}^T\boldsymbol{\beta}_{S\cup\{j_k\}}\big) \le L_n\big(\boldsymbol{X}_S^T\widehat{\boldsymbol{\beta}}_S + X_{j_k}\widehat{h}_{j_kS}\big), \tag{57}$$

where j_k in (57) is from SC. Recall the definition of M_S in (14) and the uniformity w.r.t. S in the proofs of the lemmas and theorems.

In the proof of Theorem 2, we have proved that if $S \subset \mathcal{M}$, we have uniformly in $j \in S^c$,

$$0 \le L_n\big(\boldsymbol{X}_S^T\widehat{\boldsymbol{\beta}}_S\big) - L_n\big(\boldsymbol{X}_{S\cup\{j\}}^T\widehat{\boldsymbol{\beta}}_{S\cup\{j\}}\big) = O_p\left(\frac{|S|\log p_n}{n}\right).$$

This and (56) imply the latter half of Theorem 2.

In the proof of Theorem 1, we have proved that if $S \not\subset \mathcal{M}$,

$$L_n\big(\boldsymbol{X}_S^T\widehat{\boldsymbol{\beta}}_S\big) - \min_{j\in S^c} L_n\big(\boldsymbol{X}_S^T\widehat{\boldsymbol{\beta}}_S + X_j\widehat{h}_{jS}\big) \ge \frac{D_{LB}\kappa_{LB}^2}{2}$$

for SC. This and (57) imply the former half of Theorem 2 and Theorem 3.

Hence, the proof of Corollary 1 is complete. □

5 Conclusions

In this paper, we proposed three forward variable selection procedures with a stopping rule for ultra-high-dimensional sparse quantile regression models. We established their desirable properties such as screening consistency by taking care of necessary uniformity w.r.t. covariate index sets in Sect. 2. Such uniformity has been

often overlooked in the literature on forward variable selection procedures for high-dimensional models. As we noted before, our procedures are greedy ones and statistical inference or some other procedures should follow our procedures.

We also carried out some numerical studies in Sect. 3. In Sect. 3.1, we compared our procedures with the Lasso, adaptive Lasso, SCAD, and MCP and some other procedures. Our procedures worked very well compared to all the other procedures as variable selection and screening procedures in our three examples. In Sect. 3.2, we applied our procedures and the other procedures to the riboflavin data set in Bühlmann et al. (2014).

In conclusion, we recommend gSC(m_n)+T_i and FR+T_i with $i = 2, 3$. As our numerical studies also show, there seems to be no perfect variable selection procedure in high-dimensional setups. Researchers should try several procedures for ultra-high-dimensional sparse quantile regression model including ours if necessary.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-022-00849-z.

Acknowledgements We appreciate valuable comments from the two reviewers very much. We also appreciate comments and help from Prof. Ching-Kang Ing very much. Honda's research was supported in part by JSPS KAKENHI Grant Number JP 20K11705, Japan. Lin's research was supported by grant 111-2118-M-035-007-MY2 from the National Science and Technology Council, Taiwan.

References

Barut, E., Fan, J., Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association, 111*, 1266–1277.

Belloni, A., Chernozhukov, V. (2011). ℓ1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics, 39*, 82–130.

Bühlmann, P., van de Geer, S. (2011). *Statistics for high-dimensional data*: *Methods, theory and applications*. New York: Springer.

Bühlmann, P., Kalisch, M., Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application, 1*, 255–278.

Chen, J., Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika, 95*, 759–771.

Chen, J., Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica, 22*, 555–574.

Cheng, M. Y., Honda, T., Zhang, J. T. (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association, 111*, 1209–1221.

Das, D., Gregory, K., Lahiri, S. N. (2019). Perturbation bootstrap in adaptive lasso. *The Annals of Statistics, 47*, 2080–2116.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 95*, 1348–1360.

Fan, J., Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B, 70*, 849–911.

Fan, J., Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics, 38*, 3567–3604.

Fan, J., Fan, Y., Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics, 42*, 324–351.

Fan, J., Li, R., Zhang, C. H., Zou, H. (2020). *Statistical foundations of data science*. Boca Raton: Chapman and Hall/CRC.

Hastie, T., Tibshirani, R., Wainwright, M. (2015). *Statistical learning with sparsity*: *The lasso and generalizations*. Boca Raton: Chapman & Hall/CRC.

He, X., Wang, L., Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics, 41*, 342–369.

Honda, T., Lin, C. T. (2021). Forward variable selection for sparse ultra-high-dimensional generalized varying coefficient models. *Japanese Journal of Statistics and Data Science, 4*, 151–179.
Honda, T., Ing, C. K., Wu, W. Y. (2019). Adaptively weighted group Lasso for semiparametric quantile regression models. *Bernoulli, 25*, 3311–3338.
Ing, C. K., Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica, 21*, 1473–1513.
Koenker, R. (2005). *Quantile regression*. New York: Cambridge University Press.
Koenker, R. (2021). *quantreg*: Quantile regression. R Package version 5.86. https://cran.r-project.org/web/packages/quantreg/index.html.
Koenker, R., Basset, G. (1978). Regression quantiles. *Econometrica, 46*, 33–50.
Kong, Y., Li, Y., Zerom, D. (2019). Screening and selection for quantile regression using an alternative measure of variable importance. *Journal of Multivariate Analysis, 173*, 435–455.
Lee, E. R., Noh, H., Park, B. U. (2014). Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association, 109*, 216–229.
Lin, C. T., Cheng, Y. J., Ing, C. K. (2022). Greedy variable selection for high-dimensional Cox models. *Statistica Sinica, 34*.
Liu, J., Zhong, W., Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics, 58*, 1–22.
Luo, S., Chen, Z. (2014). Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association, 109*, 1229–1240.
Pijyan, A., Zheng, Q., Hong, H. G., Li, Y. (2020). Consistent estimation of generalized linear models with high dimensional predictors via stepwise regression. *Entropy, 22*, 965.
Sherwood, B., Maidman A. (2020). *rqPen*: Penalized quantile regression. R Package version 2.2.2. https://cran.r-project.org/web/packages/rqPen/index.html.
Sherwood, B., Wang, L. (2016). Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics, 44*, 288–317.
Tang, Y., Wang, Y., Wang, H. J., Pan, Q. (2022). Conditional marginal test for high dimensional quantile regression. *Statistica Sinica, 32*, 869–892.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*: *Series B, 58*, 267–288.
van der Vaart, A. W., Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.
Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association, 104*, 1512–1524.
Wang, L., Wu, Y., Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association, 107*, 214–222.
Wu, Y., Yin, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika, 102*, 65–76.
Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*, 894–942.
Zheng, Q., Hong, H. G., Li, Y. (2020). Building generalized linear models with ultrahigh dimensional features: A sequentially conditional approach. *Biometrics, 76*, 47–60.
Zheng, Q., Peng, L., He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics, 43*, 2225–2258.