On the Analysis of Quantile Forward Regression

Hongqi Chen[*†]

October 2022

Abstract

In this paper, we study the theoretical properties of K-step and t-threshold quantile forward regressions in a linear quantile regression model with high-dimensional covariates. The model under our investigation is assumed potentially misspecified and we consider the best sparse approximation using quantile forward regressions. We provide non-asymptotic prediction bounds for both methods, and show asymptotic convergence results and the asymptotic efficacy of K-step quantile forward regression. In our asymptotic framework, we allow the number of covariates and the number of steps to diverge at different rates with the sample size. We demonstrate superior finite sample performance of quantile forward regressions to commonly-used penalization methods in terms of prediction accuracy and variable selection through extensive Monte Carlo simulations. We illustrate the usage of quantile forward regressions by two empirical applications: growth-at-risk forecasting and testing the convergence hypothesis of international economic growth.

Keywords: Forward regression; High dimensionality; Quantile regression

JEL classification: C01, C21, C61

[*]Department of Economics, University of Illinois Urbana-Champaign. Email:hongqic2@illinois.edu

[†]The author acknowledges extremely helpful comments from Ji Hyung Lee, Eunyi Chung, Marcelo C. Medeiros, Xiaofeng Shao, Zhentao Shi, Zhan Gao and Yanhao Wang. All remaining errors are my own.

1 Introduction

In econometrics and statistics, quantile regression plays an important role in characterizing the conditional distribution of a response variable by a combination of many independent variables. It is a useful tool to analyze the heterogeneity of the data or to evaluate potential distributional risks of research interests. Since the seminal work by Koenker and Bassett (1978), extensive studies have been devoted to advancing the theory of quantile regression. Several excellent references, e.g., Koenker (2005), Koenker et al. (2017), and Koenker (2017), summarize that development with other related topics. In the past four decades, quantile regression is widely applied in many fields of economics. To name a few, in labor economics, researchers use quantile regression to capture the changes in the earning distributions and measure the wage inequalities, see Chamberlain (1994), Chernozhukov and Hansen (2004), Angrist et al. (2006), Arellano and Bonhomme (2017) and many others. In health economics, Abrevaya (2002) uses quantile regression to estimate the quantiles of the infant birth-weight distribution based on demographic and maternal variables. In the causal inference framework, Abadie et al. (2002) use quantile regression to estimate the quantile treatment effects of subsidized training on trainee's salaries. In macroeconomics, Adrian et al. (2019) and Brownlees and Souza (2021) utilize quantile regression to measure the potential national or global economic recession levels. In finance, quantile estimates can act as a measure of Value-at-Risk (VaR) and capture the volatility of financial assets, see plenty of literature in Xiao et al. (2015). Overall, quantile regression might reveal more distributional information on the data and can ensure robustness for the estimates compared with the mean regression.

In quantile regression, a typical model is the linear quantile regression model, which is in the form of

$$Q_\tau\left(y_i|\boldsymbol{x}_i\right) = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_\tau \qquad for\ i = 1, \ldots, n$$

where $Q_\tau\left(y_i|\boldsymbol{x}_i\right)$ denotes the τ-th conditional quantile of a response variable y_i given the covariates $\boldsymbol{x}_{i_{p\times 1}}$. The p-dimensional $\boldsymbol{\beta}_\tau$ is the corresponding coefficients. The heterogeneous or distributional effect can be interpreted as $\boldsymbol{\beta}_\tau$ for different τ. Therefore, this type of quantile regression model can capture the heterogeneous effect of covariates, conduct robust estimation, and make the distributional prediction.

Nowadays, the advent of big data creates unprecedented opportunities and challenges for researchers. Along with the development of quantile regression, high dimensional statistical models have become popular among researchers and empirical practitioners over the past two decades. There are some famous review papers and references summarizing the recent development of high dimensional economic models and methods, to name a few, Fan et al. (2011), Athey and Imbens (2019), and Giannone et al. (2021). High dimensionality often exists in the dimension of covariates or predictors. It means that the number of covariates or predictors, p, is larger or comparably larger than the sample size n. Under this setup, two main obstacles arise. First, the traditional estimation and inference procedure might break down when $n < p$. For example, the direct solution of the ordinary least square (OLS) estimator becomes infeasible. The same issue appears for the quantile

regression. Second, with the abundance of variables, economic researchers also desire to distinguish the important covariates from other potentially irrelevant covariates. This idea can prevent those unnecessary covariates from masking true signals. Based on the selection results, they can make further explanations and modifications to the economic models. Therefore, a model selection or variable selection procedure is essential for high dimensional models.

While incorporating high dimensionality issue ($p > n$) into quantile regression, simplification of the model assumption is needed. Typically, many models assume the sparsity level, i.e., the number of non-zero coefficients s_0, which is smaller compared with n ($s_0 < n$). To accommodate this high dimensionality problem, a common approach is using the penalization method to select informative covariates. For example, Belloni and Chernozhukov (2011) first propose the l_1-penalized quantile regression, which is an analog of the famous least absolute shrinkage and selection operator (LASSO) in the mean regression. The l_1-penalized quantile regression applies shrinkage regularization on $\boldsymbol{\beta}_\tau$ and controls the number of non-zero coefficients by solving linear programming optimizations. Alternatively, we can switch the penalization scheme to l_0-penalized quantile regressions, or its variant best subset selection in quantile regressions, as in Chen and Lee (2020). Both approaches are attractive but suffer certain disadvantages, however. For the l_1-penalization method, researchers need to choose a suitable tuning parameter (penalization parameter), which is only indirectly related to the number of selected variables. In addition, the computational cost of searching for a proper choice of tuning parameter is not low. For the second approach, both best subset selection and l_0-penalization suffer an issue of high computational burden, which is NP-hard in nature, even if some modern optimization techniques such as mixed integer optimization (MIO) solver are suggested.

There is another classical approach to solving the high dimensional issue and overcoming the computational difficulty of finding the tuning parameter or the optimization solution. This approach relies on the forward selection procedure, i.e., greedy algorithm, to select one additional variable together with previously selected variables by reducing the loss function most in each step. Forward-selection type approach is widely-applied in plenty of model, variable, and sample selection problems. It shows significant effects on both the selection of objectives and the computational efficiency; see Miller (2002) and Bühlmann and Van De Geer (2011) for textbook treatments. In the classical least square model, forward selection has been applied not only in variable selection but also in variable screening for ultra-high dimensional data. For example, Zhang (2009) shows the asymptotic properties of forward least square regression, while Wang (2009) investigates the performance of forward selection in the screening framework. Although there is a huge success of forward selection under least square framework, the theoretical guarantee is still unknown in the context of the linear quantile regression model, to the best of our knowledge.

In this paper, we investigate several theoretical properties of two types of quantile forward regression in the high dimensional linear quantile regression framework. Our analysis contains K-step quantile forward regression and t-threshold quantile forward regression. This paper first studies non-asymptotic prediction bounds for both types of quantile forward regressions. These non-asymptotic bounds complement previous results about forward selection in the least square literature. The

bound for K-step quantile forward regression is a quantile analog of the results in Elenberg et al. (2018), while the bound for t-threshold quantile forward regression inherits the ideas in Kozbur (2020). Second, this paper considers an asymptotic framework that allows the number of covariates to diverge faster than the sample size. Under this framework, this paper shows the convergence results on coefficients, prediction consistency, and asymptotic efficacy for K-step quantile forward regression. Our asymptotic framework is modified from Shi and Huang (2021), which allows the number of steps K diverges at a relatively slow rate. Our coefficients convergence rate achieves the same $O\left(\sqrt{\frac{\log p}{n}}\right)$ rate as penalization methods except for an additional $\sqrt{K}$ term, which is asymptotically negligible to the $\sqrt{\frac{\log p}{n}}$ term in both theory and practice. Our results on asymptotic efficacy show good attributes about the approximation to the optimality in terms of quantile loss function asymptotically. In our numerical simulation part, we compare the performance of K-step quantile forward regression with l_1-penalized quantile regression under different data generating process (DGP) settings. Our Monte Carlo simulation results show that the K-step quantile forward regression outperforms l_1-penalized quantile regression under both sparse and dense models in terms of the average quantile prediction accuracy and the effectiveness in variable selection. To show the promising usage of our method, we apply K-step quantile forward regression into two empirical examples with real macroeconomic data. The first application considers a growth-at-risk (GaR) prediction example with high dimensional macroeconomic series. Our result suggests that K-step quantile forward regression can provide more accurate quantile forecasting than l_1-penalized quantile regression. In the second application, we examine the famous international growth convergence hypothesis from the quantile perspective with relatively high dimensional, country-specific, and cross-sectional data. We first select relevant conditioning covariates by our K-step quantile forward regression and then obtain the estimate of the effect across different quantiles. Our result confirms the hypothesis that poor countries have high potential growth rate and rich countries have a lower rate. This finding is in coherence with existing literature, e.g., Chernozhukov et al. (2016) and Koenker and Machado (1999) but provides more information on variable selection over the growth rate distribution.

The rest of the paper is organized as follows. We finish the literature review about forward regression techniques and variable selection methods for quantile regression in this section. Section 2 introduces the model framework for quantile forward selection as well as different types of forward selection algorithms. We also present several key definitions called submodularity ratio, strong concavity, and strong smoothness in the same section. Section 3 presents the non-asymptotic bounds for both types of quantile forward regressions. In section 4, we show the asymptotic consistency and efficacy for K-step quantile forward regression. We show Monte Carlo simulations of K-step quantile forward regressions and compare those finite sample numerical results with l_1-penalized method in section 5. Section 6 demonstrates two empirical applications: the Growth-at-Risk prediction example and the international growth convergence example for quantile forward regression. The last section concludes. Detailed proofs and other complementary materials are provided in the appendix.

Relevant literature

Since this paper contributes to the literature about quantile forward regression under high dimensional data, the first part of related literature is about forward selection, i.e., greedy algorithms. In the context of ordinary least square estimation, Das and Kempe (2011, 2018) show the non-asymptotic guarantee bounds of forward selection in the least square estimator using properties from the submodularity ratio. Elenberg et al. (2018) consider the connection between forward selections and weak submodularity ratio and show the performance bounds for forward selection and orthogonal matching pursuit of both linear regression and logistic regression. Wang (2009) considers asymptotic properties of forward regression under ultra-high dimensional variable screening scenario. Sancetta et al. (2016) consider various type of greedy algorithms for prediction under mixing condition. Ing (2020) investigates the least square forward selection with time series data. Shi and Huang (2021) utilize forward selection with panel data into a program evaluation framework. Moreover, Kozbur (2020) considers a testing-based forward selection procedure to choose the best model in the least square forward regression.

For the completeness of our literature review, we summarize other variable selection techniques in the context of linear quantile regression. Belloni and Chernozhukov (2011) consider the quantile regression with l_1 penalty on coefficients. Wang (2013) also shows the l_1-penalized least absolute deviation (LAD) estimator under a high dimensional setup. Besides the traditional l_1 norm regularization, there are other various penalization methods in quantile regression framework. For example, Kato (2011) proposes the group LASSO estimator for quantile regression. Zheng et al. (2013) and Zheng et al. (2015) consider the adaptive penalized quantile regression in high dimensional data. Fan et al. (2021) apply the adaptive LASSO to predictive quantile regression with mixed root covariates. Wu and Liu (2009) demonstrate the theoretical property of smoothly clipped absolute deviation (SCAD) in quantile regression.

Notations

In this paper, we denote plain letters "x" and "X" as scalars, a boldface lowercase letter "$\boldsymbol{x}$" as a vector, and a boldface uppercase letter "$\boldsymbol{X}$" as a matrix. We use $[p]$ denote the index set $\{1, \ldots, p\}$ and let $2^{[p]}$ as the set of all subsets in $[p]$. We denote $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ for the minimal and maximal eigenvalues for matrix $\boldsymbol{A}$. We let $\|\cdot\|$ as the L_1-norm and $\|\cdot\|_p$ as the L_p-norm for vectors and $\langle \cdot, \cdot \rangle$ as the inner product of two vectors in Euclidean space. We use $E(\cdot)$ and $E_n(\cdot)$ as the population and empirical expectation. Generically, we denote ∇ as the (sub)gradient operator. For any $u \in (0,1)$, $\rho_\tau(u) = u(\tau - 1(u < 0))$ acts as the traditional quantile loss function and its subgradient is $\psi_\tau(u) = \tau - 1(u < 0)$.

2 Model framework and quantile forward regression

In the section, we introduce the linear quantile regression model with high dimensional covariates. We also explain two different types of quantile forward regression algorithms, namely K-step quantile

forward regression and t-threshold quantile forward regression. In addition, some relevant definitions for further theoretical analysis are discussed.

2.1 Linear quantile regression model and notations

Suppose that the data $\{y_i, \boldsymbol{x}_i\}_{i=1}^n \in \{\mathbb{R} \times \mathbb{R}^p\}_1^n$ of sample size n is randomly generated from a joint distribution $P(y, \boldsymbol{x})$. For covariates, we allow their cardinality can diverge to infinity, i.e., $p = 1, \ldots, \infty$. We consider the traditional linear quantile regression model as the following form:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i = \sum_{j=1}^{\infty} x_{ij} \beta_j + \varepsilon_i$$

for all observations $i = 1, \ldots, n$. We denote $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j(\tau)$ as the quantile regression coefficients and $\varepsilon_i = \varepsilon_i(\tau) = y_i - Q_y(\tau|\boldsymbol{x}_i)$ as the error term for any fixed $\tau \in (0,1)$, where $Q_y(\tau|\boldsymbol{x})$ denotes the conditional quantile of y given $\boldsymbol{x}$. Under this setup, we have the conditional quantile of the population error term, ε_0, is 0 at quantile level τ given $\boldsymbol{x}$, i.e., $P(\varepsilon_0 < 0|\boldsymbol{x}) = \tau$. Later we drop the notation τ in coefficients and variables for simplicity. Similarly, we write the model in the population level as

$$y = \boldsymbol{x}^T \boldsymbol{\beta} + \varepsilon_0 = x_1 \beta_1 + x_2 \beta_2 + \cdots + \varepsilon_0$$

As we mentioned above, this model allows high dimensionality structure. Therefore, assuming a sparse structure in the coefficients is a common approach for the convenience of inference. If we impose a sparsity specification, then we call the model sparse. However, in this paper, we do not assume the linear quantile model is necessarily sparse. So if the model is dense, we consider a sparse linear quantile structure as an approximation of the true model. Therefore, we denote $\boldsymbol{\beta}_0$ as the true sparse coefficients under the sparse model or the best sparse coefficients to approximate the dense model. We will assume its cardinality is less than the sample size, i.e., $|\boldsymbol{\beta}_0| = s_0 < n$. In addition, we use $S^* \subset [p]$ as the optimal index set for $\boldsymbol{\beta}_0$ in the sense that $S^* = \{j \in [p] : \beta_{0j} \neq 0\}$.

Quantile regression estimator $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ in the population level is defined as

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} E\left(\rho_\tau\left(y - \boldsymbol{x}^T \boldsymbol{\beta}\right)\right)$$

as well as the sample analog

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} E_n\left(\rho_\tau\left(y - \boldsymbol{x}^T \boldsymbol{\beta}\right)\right)$$

For any given $\boldsymbol{\beta} \in \mathbb{R}^p$, we define the loss function $l(\boldsymbol{\beta}) : \mathbb{R}^p \to \mathbb{R}$ and empirical loss function $l_n(\boldsymbol{\beta}) : \mathbb{R}^p \to \mathbb{R}$ with respect to $\boldsymbol{\beta}$ as

$$l(\boldsymbol{\beta}) = E\left(\rho_\tau\left(y - \boldsymbol{x}^T \boldsymbol{\beta}\right)\right)$$
$$l_n(\boldsymbol{\beta}) = E_n\left(\rho_\tau\left(y - \boldsymbol{x}^T \boldsymbol{\beta}\right)\right)$$

For further analysis, we denote the negative loss function $f(S), f_n(S) : 2^{[p]} \to \mathbb{R}$ with respect to the index sets as

$$f(S) = -l(\boldsymbol{\beta}_S)$$
$$f_n(S) = -l_n(\boldsymbol{\beta}_S)$$

for any set $S \in 2^{[p]}, S \subseteq [p]$. We use $\boldsymbol{\beta}_S$ denoting the quantile regression estimator with the covariates' index set S. Similarly, $\boldsymbol{x}_S$ denotes the sub-vector of $\boldsymbol{x}$ with respect to S.

2.2 Quantile forward regression algorithms

Forward regression algorithms denote a class of model, variable, and sample selection techniques. It includes K-step forward regression, $t-$threshold forward regression, orthogonal matching pursuit, testing-based forward regression, as well as many others. In this paper, we mainly focus on the K-step forward regression and t-threshold forward regression. These two are the most commonly-used algorithms. We illustrate these two quantile forward regression algorithms in the following paragraphs. Besides, we discuss the testing-based quantile forward regression briefly in the appendix.

K-step quantile forward regression

The K-step quantile forward regression chooses one variable greedily at a time until a fixed number of steps K is reached. In every selection step, it maximizes the difference of negative quantile loss functions. The algorithm 1 below illustrates its procedure.

Algorithm 1 K-step quantile forward regression

1. Set $S_0 = \emptyset$.
2. For $1 \leq k \leq K$,
   (a) Select the variable index $s = \arg\max_{j \in [p] \setminus S_{k-1}} f(S_{k-1} \cup \{j\}) - f(S_{k-1})$.
   (b) Set $S_k = S_{k-1} \cup \{s\}$.
3. Return S_K and the corresponding estimator $\hat{\boldsymbol{\beta}}_{S_K}$.

t-threshold quantile forward regression

Under the same model setup, the t-threshold quantile forward regression selects covariates by setting a threshold. For any fixed choice of t as the performance improvement threshold, we hope to make the improvement of each step larger than t. Among those variables which can satisfy this threshold condition, we choose the best one that maximizes the improvement. Details are in the algorithm 2 below.

Algorithm 2 t-threshold quantile forward regression

1. Set $S_0 = \emptyset$.
2. For $k = 1, 2, \ldots$
   (a) If $f(S_{k-1} \cup \{j\}) - f(S_{k-1}) > t$ for some $j \in [p] \setminus S_{k-1}$
      i. Select the variable index
      $$s = \underset{j \in [p] \setminus S_{k-1}}{\arg\max} \{f(S_{k-1} \cup \{j\}) - f(S_{k-1}) \,|\, f(S_{k-1} \cup \{j\}) - f(S_{k-1}) > t\}$$
      ii. Set $S_k = S_{k-1} \cup \{s\}$.
   (b) Else, break the algorithm and let $K = k$
3. Return $\hat{S} = S_K$ with final step K. $\hat{S}$ satisfies $f\left(\hat{S} \cup \{j\}\right) - f\left(\hat{S}\right) < t$ for all $j \in [p] \setminus \hat{S}$, and the corresponding estimator is $\hat{\boldsymbol{\beta}}_{\hat{S}}$.

As we can notice, both algorithms above choose one important variable in every step but follow different criteria. For K-step quantile forward regression, the choice of K is crucial. In the later simulation and applications parts, K is determined by information criteria, such as AIC or BIC. This approach is intuitive since classical information criteria depend on the number of selected covariates, which is K exactly in the algorithm 1. In addition, K is also important in the asymptotic analysis because we need to allow $K \to \infty$ as $p \to \infty$ in our high-dimensionality framework. It is intuitive that the more steps we experience, the higher probability we can select important covariates. For the t-threshold quantile forward regression, selecting a suitable t is important. An advantage of the t-threshold approach is that t is a continuous tuning parameter while K can only be integers. Since the choice of t depends on data, however, there is no criterion for theoretical guidance so far. Although this drawback limits the usage of the algorithm 2, one variant of $t-$threshold forward selection called testing-based forward selection can provide some meaningful explanation of the threshold, see Kozbur (2020). The testing-based forward selection procedure incorporates a statistical hypothesis to decide important covariates; see more discussion in the appendix.

2.3 Related definitions

In this subsection, we introduce a key definition named submodularity ratio. As mentioned by Das and Kempe (2018, 2011), the submodularity ratio plays an important role in analyzing sparse selection problems, especially for forward selection algorithms. We borrow the definition of submodularity ratio from definition 2 in Das and Kempe (2018):

Definition 2.1. The submodularity ratio of a monotone function f with respect to a set U and an

integer $k \geq 1$ is defined as

$$\gamma_{U,k}(f) = \min_{L \subseteq U, S: |S| \leq k, S \cap L = \emptyset} \frac{\sum_{x \in S} f(L \cup \{x\}) - f(L)}{f(L \cup S) - f(L)}$$

In the next section, we will show the lower bound of the submodularity ratio for a transformation of negative quantile loss function f and use this term to derive the non-asymptotic prediction bounds. This approach is similar to Das and Kempe (2018), who utilize the submodularity ratio in the least square forward regression estimator.

Moreover, it is worth mentioning that if f is submodular[1], then $\gamma_{U,k}(f) \geq 1$. In the forward selection algorithm, we usually have $0 \leq \gamma_{U,k}(f) \leq 1$ and researchers call it the weak submodularity. This means that even if the function f is not submodular, we can still use its properties to provide performance bounds.

To show the weak submodularity, we also mention two relevant definitions from Elenberg et al. (2018): restricted strong concavity and restricted smoothness of a function f. These two properties imply weak submodularity.

Definition 2.2. A function $f : \mathbb{R}^p \to \mathbb{R}$ satisfies the restricted strong concavity with parameter m_Ω and restricted smoothness with parameter M_Ω on a domain $\Omega \subset \mathbb{R}^p \times \mathbb{R}^p$ if for all $(\boldsymbol{a}, \boldsymbol{b}) \in \Omega$,

$$-\frac{m_\Omega}{2} \|\boldsymbol{b} - \boldsymbol{a}\|_2^2 \geq f(\boldsymbol{b}) - f(\boldsymbol{a}) - \langle \nabla f(\boldsymbol{a}), \boldsymbol{b} - \boldsymbol{a} \rangle \geq -\frac{M_\Omega}{2} \|\boldsymbol{b} - \boldsymbol{a}\|_2^2$$

An intuitive explanation for these two definitions is that restricted strong concavity and restricted smoothness parameters control the upper and lower bounds of the remainder of the first order Taylor expansion of function f. Good behavior of the remainder indicates decent control of the first order approximation of f. Therefore, when the greedy algorithm applies, it will eventually make control on every greedy step; see more discussion in Elenberg et al. (2018). For example, in the least square forward regression, restricted strong concavity and restricted smoothness parameters of the mean-square loss function are apparent and related to the Gram matrix of the data $E\boldsymbol{x}\boldsymbol{x}^T$. In the next section, we will also provide these two parameters for the quantile forward regressions.

3 Non-asymptotic properties for quantile forward regressions

In this section, we analyze prediction error bounds for algorithms 1 and 2 above. We first analyze the quantile loss prediction bound for K-step quantile forward regression. The primary tool we use is from Elenberg et al. (2018), which shows that the non-asymptotic prediction bound is related to the submodularity ratio of the forward selection algorithms. Next, we illustrate a theoretical analysis of $t-$threshold quantile forward regression. We apply the techniques from Das and Kempe (2018) and Kozbur (2020), which show the prediction bounds for least square forward regression.

[1] A function f is submodular if for all subset $L \in [p]$ and $x, y \in [p] \setminus A$, $f(A \cup \{x\}) - f(A) \geq f(A \cup \{x, y\}) - f(A \cup \{y\})$.

3.1 K-step quantile forward regression

To consider the prediction bound of K-step quantile forward regression, we state the following regularity assumptions.

Assumption 1. *Let* $\{y_i, \boldsymbol{x}_i\}_{i=1}^n \in \{\mathbb{R} \times \mathbb{R}^p\}_1^n$ *be i.i.d observations randomly drawn from the same joint distribution* $P(y, \boldsymbol{x})$.

Assumption 2. *(Boundedness conditions) (i) Assume* $\sup_{j\in[p]} |x_j| < \infty$. *(ii) For any* $\tilde{\boldsymbol{\beta}}_S = \arg\min_{\boldsymbol{\beta}_S} E\left(\rho_\tau\left(y - \boldsymbol{x}_S^T\boldsymbol{\beta}_S\right)\right)$ *with* $S \subseteq [p]$ *and* $|S| \leq 2K$, *let* $\sup_S \left|\boldsymbol{x}_S^T\tilde{\boldsymbol{\beta}}_S\right| < \infty$. *(iii) In addition, let* $\boldsymbol{\beta}_S \in \Omega_S$ *and* $\Omega_S \subset \mathbb{R}^{|S|}$ *is compact for all* $S \subseteq [p]$.

Assumption 3. *(Density conditions) (i) For any set* $S \subseteq [p]$ *with* $|S| \leq 2K$, *the conditional density function* $f_y(y|\boldsymbol{x}_S)$ *has bounded support and is continuously differentiable for any* $\boldsymbol{x}_S$ *in its support. (ii) For any* $S \subseteq [p]$ *with* $|S| \leq 2K$, *let* $0 < \underline{c}_f \leq f_y(y|\boldsymbol{x}_S) \leq \bar{c}_f < \infty$ *for some universal constants* $\underline{c}_f$ *and* $\bar{c}_f$, *i.e., the conditional density function is bounded below and above by some universal constants.*

Assumption 4. *(Eigenvalues condition) For any set* $S \subseteq [p]$ *with* $|S| \leq 2K$, *the matrix* $\boldsymbol{J}_S = Ef_y(y|\boldsymbol{x}_S)\boldsymbol{x}_S\boldsymbol{x}_S^T$ *satisfies the following eigenvalue condition*

$$0 < m_S \leq \lambda_{\min}(\boldsymbol{J}_S) \leq \lambda_{\max}(\boldsymbol{J}_S) \leq M_S < \infty$$

with some universal constants m_S *and* M_S.

Moreover, we denote $\phi_{\min}(s)(\boldsymbol{J}) = \min_{S\subseteq[p]:|S|\leq s} \lambda_{\min}(\boldsymbol{J}_S)$ *as the minimal eigenvalue over all sets* S *and* $\phi_{\max}(s)(\boldsymbol{J}) = \max_{S\subseteq[p]:|S|\leq s} \lambda_{\max}(\boldsymbol{J}_S)$ *as the maximal eigenvalue respectively.*

Before we move into the detailed proofs, we make some remarks on the above assumptions. Assumption 1 is a regular assumption in data sampling. This *i.i.d* assumption can be relaxed if we only consider the non-asymptotic properties. Assumption 2 imposes the boundedness conditions on the covariates and their linear combinations with respect to the coefficients. The boundedness conditions are common in high dimensional analysis, see Ma et al. (2017) and can simplify our theoretical analysis. Intuitively, we can replace those boundedness conditions by imposing control on the tail of the random variables. However, this will make the technical proofs much more complicated so we will get rid of this approach. Assumption 3 on the conditional density of response variable is standard in quantile regression literature, even for the high dimensional quantile literature; see Belloni and Chernozhukov (2011). This assumption rules out irregular cases about the conditional density of the data. In addition, a well-behaved shape on the conditional density implies the restricted strong concavity and restricted smoothness condition of the quantile loss function. As for the eigenvalues assumption, the assumption 4 is modified from Lee and Shin (2021) and Lu and Su (2015). This restricted eigenvalue assumption is crucial in deriving the following bounds. In addition, the definitions of $\phi_{\min}(s)(\boldsymbol{J})$ and $\phi_{\max}(s)(\boldsymbol{J})$ imply the monotone properties of $\phi_{\min}(s)(\boldsymbol{J})$ and $\phi_{\max}(s)(\boldsymbol{J})$ with respect to s, i.e., $\phi_{\min}(s_2)(\boldsymbol{J}) < \phi_{\min}(s_1)(\boldsymbol{J})$ and $\phi_{\max}(s_1)(\boldsymbol{J}) < \phi_{\max}(s_2)(\boldsymbol{J})$ if $s_1 < s_2$.

In algorithm 1, we shall minimize quantile loss functions in every step with respect to certain index set S, i.e., $\min_{\boldsymbol{\beta}_S} E\left[\rho_\tau\left(y-\boldsymbol{x}_S^T\boldsymbol{\beta}_S\right)\right]$. An equivalent formulation is to maximize the quantile-type goodness of fit $R_\tau^1\left(\boldsymbol{\beta}_S\right)$ in Koenker and Machado (1999):

$$R_\tau^1\left(\boldsymbol{\beta}_S\right)=1-\frac{E\left[\rho_\tau\left(y-\boldsymbol{x}_S^T\boldsymbol{\beta}_S\right)\right]}{E\left[\rho_\tau\left(y-\alpha\right)\right]}$$

where the numerator is the quantile loss function with respect to $\boldsymbol{x}_S$ and the denominator is the quantile loss function with respect to the intercept term α only. It is straightforward that $0 \leq R_\tau^1\left(\boldsymbol{\beta}_S\right) \leq 1$ and $R_\tau^1\left(\boldsymbol{\beta}_S\right)$ measures the goodness of fit in the quantile sense like R^2 in least square estimation. Equipped with this measurement, we show $R_\tau^1\left(\boldsymbol{\beta}_S\right)$ is weak submodular in the following lemma and then we can derive the performance bound for K-step quantile forward regression. Our approach follows Elenberg et al. (2018) and the basic idea is to show that restricted strong concavity and restricted smoothness of $R_\tau^1\left(\boldsymbol{\beta}_S\right)$.

Lemma 3.1. *(Restricted strong concavity and restricted smoothness of $R_\tau^1\left(\boldsymbol{\beta}_S\right)$) Under assumptions 2, 3, and 4, the population quantile goodness of fit $R_\tau^1\left(\boldsymbol{\beta}_S\right)=1-\frac{E\left[\rho_\tau\left(y-\boldsymbol{x}_S^T\boldsymbol{\beta}_S\right)\right]}{E\left[\rho_\tau\left(y-\alpha\right)\right]}$ satisfies restricted strong concavity and restricted smoothness from definition 2.2 for any $\boldsymbol{\beta}_S \in \mathbb{R}^p$ and index set S. i.e., let*

$$D=R_\tau^1\left(\boldsymbol{\beta}_2\right)-R_\tau^1\left(\boldsymbol{\beta}_1\right)-\left\langle\frac{E\left[\boldsymbol{x}^T\left(\tau-1\left(y-\boldsymbol{x}^T\boldsymbol{\beta}_1<0\right)\right)\right]}{E\left[\rho_\tau\left(y-\alpha\right)\right]}, \boldsymbol{\beta}_2-\boldsymbol{\beta}_1\right\rangle$$

, we have

$$-\frac{\lambda_{\min}\left(\boldsymbol{J}_S\right)}{2E\left[\rho_\tau\left(y-\alpha\right)\right]}\left\|\boldsymbol{\beta}_2-\boldsymbol{\beta}_1\right\|_2^2 \geq D \geq -\frac{\lambda_{\max}\left(\boldsymbol{J}_S\right)}{2E\left[\rho_\tau\left(y-\alpha\right)\right]}\left\|\boldsymbol{\beta}_2-\boldsymbol{\beta}_1\right\|_2^2$$

Remark 3.1. For least square regression, it is known that the negative quadratic loss function satisfies restricted strong concavity and restricted smoothness in Elenberg et al. (2018). Indeed, restricted strong concavity and restricted smoothness parameters are $\lambda_{\min}\left(\boldsymbol{x}\boldsymbol{x}^T\right)$ and $\lambda_{\max}\left(\boldsymbol{x}\boldsymbol{x}^T\right)$, the minimal and maximal eigenvalue of the Gram matrix. This is consistent with the restricted isometry property as in Das and Kempe (2011). In the quantile regression, we obtain a similar expression of restricted strong concavity and restricted smoothness parameters for $R_\tau^1\left(\boldsymbol{\beta}_S\right)$, i.e., $\frac{\lambda_{\min}\left(\boldsymbol{J}_S\right)}{E\left[\rho_\tau\left(y-\alpha\right)\right]}$ and $\frac{\lambda_{\max}\left(\boldsymbol{J}_S\right)}{E\left[\rho_\tau\left(y-\alpha\right)\right]}$. Indeed, $\boldsymbol{J}_S$ relates to the conditional density behavior of the data. Like classical quantile regression literature, if $\boldsymbol{J}_S$ behaves well, then good theoretical property follows.

It is worth mentioning another important point. Based on the same technique, we can also show the restricted concavity and restricted smoothness for the negative quantile loss function $-l\left(\boldsymbol{\beta}_S\right)$ with $\lambda_{\min}\left(\boldsymbol{J}_S\right)$ and $\lambda_{\max}\left(\boldsymbol{J}_S\right)$ as corresponding parameters. There are two reasons why we focus on $R_\tau^1\left(\boldsymbol{\beta}_S\right)$. First, $R_\tau^1\left(\boldsymbol{\beta}_S\right)$ is always a positive number between 0 and 1. Second, $R_\tau^1\left(\boldsymbol{\beta}_S\right)$ has an intuitive meaning as the goodness-of-fit to measure the amount of the response variable that is explained by the selected covariates.

Based on lemma 3.1, we have the following theorem for approximation guarantee bound of $R_\tau^1\left(\boldsymbol{\beta}_S\right)$ by directly applying the results in Elenberg et al. (2018) in the population level. This approximation guarantee bound means that it depends on the best approximation results under the

model misspecification with high dimensional covariates.

Theorem 3.1. *(Approximation guarantee bound for $R_\tau^1\left(\boldsymbol{\beta}_{S_K}\right)$) Under assumptions 2, 3, and 4, and algorithm 1 above, the submodularity ratio $\gamma_{S_K,K}$ is lower bounded by $\frac{\phi_{\min}(2K)(\boldsymbol{J})}{\phi_{\max}(2K)(\boldsymbol{J})}$. Moreover, we have*

$$R_\tau^1\left(\boldsymbol{\beta}_{S_K}\right) \geq \left(1-e^{-\gamma_{S_K,K}}\right) R_\tau^1\left(\boldsymbol{\beta}_{S^*}\right) \geq \left(1-e^{-\frac{\phi_{\min}(2K)(J)}{\phi_{\max}(2K)(J)}}\right) R_\tau^1\left(\boldsymbol{\beta}_{S^*}\right)$$

where we denote S^* as the optimal index set for covariates.

Proof. With lemma 3.1, we know that $R_\tau^1\left(\boldsymbol{\beta}_{S_K}\right)$ satisfies weak submodularity with restricted strong concavity parameter $\frac{\phi_{\min}(2K)(\boldsymbol{J})}{E[\rho_\tau(y-\alpha)]}$ and restricted smoothness parameter $\frac{\phi_{\min}(2K)(\boldsymbol{J})}{E[\rho_\tau(y-\alpha)]}$ under K-step quantile forward regression. Therefore, by theorem 1 in Elenberg et al. (2018),

$$\gamma_{S_K,K} \geq \frac{\phi_{\min}\left(2K\right)\left(\boldsymbol{J}\right)}{\phi_{\max}\left(2K\right)\left(\boldsymbol{J}\right)}$$

From theorem 3 in Elenberg et al. (2018), two inequalities hold naturally. □

Our theorem 3.1 is an analog of theorem 3 in Elenberg et al. (2018) in the context of quantile regression with respect to the goodness of fit R_τ^1. It is worth mentioning that we do not impose the true specification of the linear quantile model in this theorem. Regardless the model is correctly specified with a sparse true index set S^* or the model is misspecified, which means the number of covariates p diverges to infinity, we always have this constant factor approximation bound for R_τ^1. Since we have the population guarantee bounds for R_τ^1, it is straight forward to obtain the population prediction bounds for the quantile loss function $E\left[\rho_\tau\left(y-\boldsymbol{x}_{S_K}^T\boldsymbol{\beta}_{S_K}\right)\right]$.

Theorem 3.2. *(Approximation guarantee bound for K-step quantile forward regression) Under the conditions in theorem 3.1, we have*

$$E\left[\rho_\tau\left(y-\boldsymbol{x}_{S_K}^T\boldsymbol{\beta}_{S_K}\right)\right] \leq \left(1-e^{-\frac{\phi_{\min}(2K)(\boldsymbol{J})}{\phi_{\max}(2K)(\boldsymbol{J})}}\right) E\left[\rho_\tau\left(y-\boldsymbol{x}_{S^*}^T\boldsymbol{\beta}_{S^*}\right)\right] + e^{-\frac{\phi_{\min}(2K)(\boldsymbol{J})}{\phi_{\max}(2K)(\boldsymbol{J})}} E\left[\rho_\tau\left(y-\alpha\right)\right]$$

Proof. This result follows from theorem 3.1 and the definition of R_τ^1. □

From the above theorem, the population approximation bound for $E\left[\rho_\tau\left(y-\boldsymbol{x}_{S_K}^T\boldsymbol{\beta}_{S_K}\right)\right]$ is a linear combination of the population quantile loss with respect to the optimal sparse index set S^* and the intercept term α. In addition, we know $E\left[\rho_\tau\left(y-\alpha\right)\right]$ is always larger than $E\left[\rho_\tau\left(y-\boldsymbol{x}_{S^*}^T\boldsymbol{\beta}_{S^*}\right)\right]$. So there is a gap between the population quantile loss $E\left[\rho_\tau\left(y-\boldsymbol{x}_{S_K}^T\boldsymbol{\beta}_{S_K}\right)\right]$ and the optimal loss $E\left[\rho_\tau\left(y-\boldsymbol{x}_{S^*}^T\boldsymbol{\beta}_{S^*}\right)\right]$. However, from the monotonicity of $\phi_{\min}\left(2K\right)\left(\boldsymbol{J}\right)$ and $\phi_{\max}\left(2K\right)\left(\boldsymbol{J}\right)$, when K is increasing, $E\left[\rho_\tau\left(y-\boldsymbol{x}_{S_K}^T\boldsymbol{\beta}_{S_K}\right)\right]$ is approaching to $E\left[\rho_\tau\left(y-\boldsymbol{x}_{S^*}^T\boldsymbol{\beta}_{S^*}\right)\right]$ since the second term is diminishing.

For the sampling property, we have the following non-asymptotic probabilistic bound. It indicates that the sample quantile loss function is bounded by the linear combination of $E\left[\rho_\tau\left(y-\boldsymbol{x}_{S^*}^T\boldsymbol{\beta}_{S^*}\right)\right]$ and $E\left[\rho_\tau\left(y-\alpha\right)\right]$ under high probability if n is large.

Corollary 3.1. *Under the conditions in theorem 3.2, for any $\varepsilon > 0$, we have*

$$Pr\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i - \boldsymbol{x}_{i,S_K}^T\boldsymbol{\beta}_{S_K}\right) \leq \left(1 - e^{-\frac{\phi_{\min}(2K)(\boldsymbol{J})}{\phi_{\max}(2K)(\boldsymbol{J})}}\right) E\left[\rho_\tau\left(y - \boldsymbol{x}_{S^*}^T\boldsymbol{\beta}_{S^*}\right)\right] + e^{-\frac{\phi_{\min}(2K)(\boldsymbol{J})}{\phi_{\max}(2K)(\boldsymbol{J})}} E\left[\rho_\tau\left(y - \alpha\right)\right] + \varepsilon\right]$$
$$\geq 1 - e^{-\frac{n\varepsilon^2}{\tau\vee(1-\tau)(1+\frac{\varepsilon}{3})}}$$

As a summary, we derive the approximation guarantee bound and non-asymptotic bound for K-step quantile forward regression. It is clear that the probabilistic bound in corollary 3.1 converges to 1 as $n \to \infty$. The proof of this sampling bound is based on Bernstein-type inequality on bounded random variables. If we consider the case with $n \to \infty$ and $K \to \infty$, intuitively, we should obtain an excellent approximation of $E\left[\rho_\tau\left(y - \boldsymbol{x}_{S^*}^T\boldsymbol{\beta}_{S^*}\right)\right]$ by the sample $\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i - \boldsymbol{x}_{i,S_K}^T\boldsymbol{\beta}_{S_K}\right)$. More details on the asymptotic theory of K-step quantile forward regression will be discussed in section 4.

3.2 t-threshold quantile forward regression

In this subsection, we consider the performance bound of t-threshold quantile forward regression. Unlike simply choosing the number of step K, using threshold t as a tuning parameter is more flexible. The reason is that t can be a continuous positive number while K can only be an integer. However, the disadvantage of selecting t is that there are no traditional standards, e.g., information criterion, in practice.

Our t-threshold quantile forward regression algorithm is a quantile type analog of algorithm 1 in Kozbur (2020). To characterize the performance bound, we shall first prove the following lemmas. Lemma 3.2 is a quantile regression version of lemma 17 in Das and Kempe (2018), which is wildly used in the least square forward regression literature. Lemma 3.3 is a basic inequality for the loss function.

Lemma 3.2. *Under the assumptions 2, 3, 4 and algorithm 2 above, we have*

$$f\left(\hat{S}\cup S^*\right) - f\left(\hat{S}\right) \leq \frac{\phi_{\max}\left(K + s_0\right)(\boldsymbol{J})}{\phi_{\min}\left(K + s_0\right)(\boldsymbol{J})} s_0 t$$

Lemma 3.3. $f\left(\hat{S}\cup S^*\right) \geq f\left(S^*\right)$

As a remark, lemma 3.2 controls the difference of quantile loss functions between $\hat{S}\cup S^*$ and $\hat{S}$. Intuitively, the less of the cardinality of the true coefficients s_0 or the smaller the threshold t we choose, the better we can control this difference.

Based on previous lemmas, we can show the l_2-bound and quantile loss bound for prediction error. The proof procedure follows the first part of theorem 1 in Kozbur (2020). Moreover, we need to state an additional assumption on the density of the error term. The following assumption controls good behaviors of ε_0.

Assumption 5. *The unconditional probability density function of ε_0, f_{ε_0} is bounded below and above by $\underline{f}_{\varepsilon_0}$ and $\bar{f}_{\varepsilon_0}$ in the sense that*

$$0 < \underline{f}_{\varepsilon_0} \leq f_{\varepsilon_0} \leq \bar{f}_{\varepsilon_0} \leq \infty$$

Theorem 3.3. *Under the assumptions 2, 3, 4, 5 and algorithm 2 above, we have the following l_2 prediction bound*

$$E\left(\left(\boldsymbol{x}^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}^T\boldsymbol{\beta}_0\right)^2\right)^{\frac{1}{2}} \leq \sqrt{K+s_0}\left[\frac{\frac{2}{\underline{f}_{\varepsilon_0}}\left\|E\boldsymbol{x}\psi_\tau\left(y-\boldsymbol{x}^T\boldsymbol{\beta}_0\right)\right\|_\infty}{\varphi_{\min}\left(K+s_0\right)\left(E\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right)} + \sqrt{\frac{2}{\underline{f}_{\varepsilon_0}}\frac{\phi_{\max}\left(K+s_0\right)\left(\boldsymbol{J}\right)}{\phi_{\min}\left(K+s_0\right)\left(\boldsymbol{J}\right)}t}\right]$$

In addition, the quantile prediction loss has the same upper bound

$$\rho_\tau\left(\boldsymbol{x}^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}^T\boldsymbol{\beta}_0\right) \leq \sqrt{K+s_0}\left[\frac{\frac{2}{\underline{f}_{\varepsilon_0}}\left\|E\boldsymbol{x}\psi_\tau\left(y-\boldsymbol{x}^T\boldsymbol{\beta}_0\right)\right\|_\infty}{\varphi_{\min}\left(K+s_0\right)\left(E\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right)} + \sqrt{\frac{2}{\underline{f}_{\varepsilon_0}}\frac{\phi_{\max}\left(K+s_0\right)\left(\boldsymbol{J}\right)}{\phi_{\min}\left(K+s_0\right)\left(\boldsymbol{J}\right)}t}\right]$$

As a summary for this subsection, it is worth mentioning that these prediction bounds are in line with the simple forward selection in Kozbur (2020). In addition, these bounds do not depend on the data sampling assumption 1. Therefore, they are quite general for a different type of data structure. If we impose some high-level assumptions $K = O\left(s_0\right)$ and $t = O\left(\frac{\log p}{n}\right)$, we can achieve the same $O\left(\sqrt{\frac{s_0 \log p}{n}}\right)$ rate as Kozbur (2020) in the asymptotic world. Another extension of this result is the non-asymptotic bounds for testing-based quantile forward regression. We provide more details in the appendix.

4 Asymptotic analysis for K-step quantile forward regression

In this section, we investigate the asymptotic analysis for the K-step quantile forward regression at the sample level. We only analyze K-step quantile forward regression since our asymptotic framework requires both K and n to go to infinity with certain regularity conditions. The asymptotic analysis gives us the convergence rate for K-step quantile forward regression estimators and its quantile loss prediction error.

As far as we know, the connection between K-step and t-threshold quantile forward regressions is generally unknown. It is still an open question to find explainable reasoning for the asymptotic conditions of t-threshold quantile forward regression. However, the asymptotics with respect to K is intuitive. For the asymptotic consistency of $\hat{\boldsymbol{\beta}}_{\hat{S}}$, we do not impose conditions on the sparsity level s_0. Hence the model can be misspecified as before. Our proof procedure follows Lu and Su (2015) and Lee and Shin (2021). Later, we consider the asymptotic efficacy, which assumes the sparsity level s_0 is asymptotically negligible compared with the number of steps K. Following Shi and Huang (2021), this assumption gives us more elegant results on the asymptotic optimality of the K-step quantile forward regression.

4.1 K-step quantile forward regression

In addition to the assumptions above, we impose the following rate condition on K, p, and n.

Assumption 6. *(Rate conditions)*

(i) For any sequences K, p, and n, we have $\frac{1}{K} + \frac{K}{\log p} + \frac{\log p}{n} \to 0$ as $n \to \infty$ and $p \to \infty$.

(ii) For such sequences, we let $\frac{K}{\left(\frac{n}{\log p}\right)^{\frac{1}{2}}} \to 0$.

As a remark, assumption 6 (i) requires some regularity conditions on the number of steps K, the number of covariates p, and the sample size n. It contains three parts about the relative divergence rates among K, p, and n. For the first part, we impose the number of steps K goes to infinity. This condition is commonly used in forward selection literature; see assumption 1 in Shi and Huang (2021). An intuitive explanation is that the more we select, the better our approximation result. For the second part, we require the number of steps not to diverge too fast compared with the number of covariates. K can increase at most the logarithmic rate of p. This assumption is reasonable because if K increases too fast, then more irrelevant covariates will be selected. For the last part, we restrict the number of covariates so that it can not increase faster than the exponential rate of the sample size. This condition is wildly-used in the high-dimensional literature about variable selection, e.g., the l_1-penalization achieves $O\left(\sqrt{\frac{\log p}{n}}\right)$ rate. For assumption 6 (ii), it is a technical assumption that further restricts the divergence rate of K. It is prepared for the prediction consistency of quantile forward regression. Our assumption is more relaxed compared with a similar condition in assumption 1 of Shi and Huang (2021).

4.1.1 Asymptotic consistency of $\hat{\boldsymbol{\beta}}_{\hat{S}}$

Theorem 4.1. *If assumptions 1, 2, 3, 4, and 6 hold, then*

$$\max_{1\leq k\leq K} \max_{|\hat{S}|\leq k} \left\| \hat{\boldsymbol{\beta}}_{\hat{S}} - \boldsymbol{\beta}^0_{\hat{S}} \right\| = O_p\left(\sqrt{\frac{K\log p}{n}}\right)$$

As a remark, theorem 4.1 establishes the $O_p\left(\sqrt{\frac{K\log p}{n}}\right)$ consistency rate with $\hat{\boldsymbol{\beta}}_{\hat{S}}$ and $\boldsymbol{\beta}^0_{\hat{S}}$. This convergence rate is consistent with the rate in Lee and Shin (2021), which consider the complete subset prediction in quantile regression. This result is consistent with our intuition since K-step forward regression only chooses a subset of all possible combinations of covariates. Therefore, the uniform convergence rate should be similar to the complete subset approach. It is worth mentioning that we do not need to consider the uniform result over all possible subsets in forward quantile regression, so our result is more flexible. In addition, our convergence rate is comparable with Shi and Huang (2021), which use the least square forward regression in panel data. Their convergence rate is $O_p\left(\sqrt{\frac{K^3\log p}{n}}\right)$ in our notation. Our results differ in the power of K but share the same $\sqrt{\frac{\log p}{n}}$ factor that wildly appears in the high-dimensional literature. Indeed, the difference in the

power of K does not have too much influence since the rate of K is relatively slow compared with $\frac{n}{\log p}$.

Based on the consistency in parameters, the following lemma show the negligible difference in probability between $\hat{\rho}_{\tau,\hat{S}} = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i - \boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)$ and $\rho_{\tau,\hat{S}} = E\rho_\tau\left(y_i - \boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)$. This approximation is prepared for the asymptotic efficacy result later.

Corollary 4.1. *Denote* $\hat{\rho}_{\tau,\hat{S}} = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i - \boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)$ *and* $\rho_{\tau,\hat{S}} = E\rho_\tau\left(y_i - \boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)$. *Given the conditions in theorem 4.1, we have*

$$\max_{1\le k\le K}\max_{|\hat{S}|\le k}\left|\hat{\rho}_{\tau,\hat{S}} - \rho_{\tau,\hat{S}}\right| = O_p\left(\sqrt{\frac{K^2\log p}{n}}\right)$$

As we can notice, the probabilistic difference between $\hat{\rho}_{\tau,\hat{S}}$ and $\rho_{\tau,\hat{S}}$ converges at rate $O_p\left(\sqrt{\frac{K^2\log p}{n}}\right)$, which is slightly larger than the consistency rate of $\hat{\boldsymbol{\beta}}$. Under our rate assumption 6, we know $O_p\left(\sqrt{\frac{K^2\log p}{n}}\right) = o_p(1)$. Hence, this probabilistic difference is negligible. Based on this result, we can further consider the asymptotic efficacy when the true model is sparse with level s_0 in the next subsection.

4.1.2 Asymptotic efficacy

In this subsection, we investigate the asymptotic efficacy, which considers quantile loss consistency. The following theorem 4.2 is a quantile-type analog of theorem 2 in Shi and Huang (2021). Our asymptotic efficacy emphasizes the quantile forward regression can produce an excellent approximation of quantile prediction loss on the correctly specified model while the true sparsity level s_0 is dominated by the number of steps K asymptotically. Our proof procedure follows the proof of forward selection in Shi and Huang (2021), but we focus on the quantile forward regression.

The following lemma 4.1 shows every step in the quantile forward regression contributes to narrowing down the difference between two quantile losses within two index sets. The lemma 4.2 considers the overall difference between the selected index set and an arbitrary fixed set.

Lemma 4.1. *For any set* $U, V \subset [p]$ *such that* $U \supset V$ *and* $u = |U| > |V| = v$, *we have*

$$\max_{\{x\}\in U}\rho_\tau(V) - \rho_\tau(V\cup\{x\}) \ge \frac{\phi_{\min}(u)(\boldsymbol{J})}{\phi_{\max}(u)(\boldsymbol{J})}\frac{1}{u-v}\left(\rho_\tau(V) - \rho_\tau(U)\right)$$

We define the following sequences of index sets to show the next lemma,

$$U_K(\kappa) = \left\{(U_1,\ldots,U_K)\in[p]^K \middle| \begin{array}{c} U_{k-1}\subset U_k, |U_k\backslash U_{k-1}| = 1 \\ \rho_\tau(U_{k-1}) - \rho_\tau(U_k) \ge (1-\kappa)\max_{j\in[p]}\rho_\tau(U_{k-1}) - \rho_\tau(U_k\cup\{j\}) \end{array}\right\}$$

Lemma 4.2. *For any sequence of sets* $(U_1,\ldots,U_K)\in U_K(\kappa)$ *and any* $W\subset[p]$ *with* $|W| = w$, *we*

have

$$\rho_\tau\left(U_K\right)-\rho_\tau\left(W\right)\leq\rho_\tau\left(\emptyset\right)\left(1-\left(1-\kappa\right)\frac{\phi_{\min}\left(w+K\right)\left(\boldsymbol{J}\right)}{\phi_{\max}\left(w+K\right)\left(\boldsymbol{J}\right)}\frac{1}{w}\right)^K$$

Based on two previous lemmas, we can state the following theorem for asymptotic efficacy.

Theorem 4.2. *Under the conditions in theorem 4.1, for any S_0 with $|S_0|=s_0$ satisfying $\frac{s_0}{K}\to 0$ and $\frac{s_0^2K^2\log p}{n}\to 0$ as $p\to\infty$ and $n\to\infty$, we have*

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)-E\rho_\tau\left(y_i-\boldsymbol{x}_{i,S_0}^T\boldsymbol{\beta}_{S_0}\right)\leq\delta\right)\to 1$$

for any universal constant $\delta>0$.

As the theorem 4.2 shown above, the quantile forward regression can eventually control the difference between the empirical quantile loss and the optimal quantile loss as long as the number of step K dominates the number of optimal covariates s_0. Since we allow model misspecification, the optimal covariates can be true covariates or the best covariates to approximate the linear quantile model. In addition, we do not restrict s_0 to be finite or not in the theorem. If s_0 is fixed, the result is intuitive since when K diverges to ∞, all optimal covariates will be selected in the end. Even if s_0 diverges, as long as it is dominated by K and satisfies the rate condition $\frac{s_0^2K^2\log p}{n}\to 0$, our asymptotic results still hold for quantile forward regressions.

5 Monte Carlo simulations

In this section, we present Monte Carlo simulations to investigate the finite sample performance of quantile forward regression. We focus on K-step quantile forward regression and compare our results with l_1-penalized quantile regression. Our Monte Carlo experiments contain two parts. The first part is based on pure numerical simulations, while the second part, which is application-based, considers the data generating process from the GaR application in the next application section.

5.1 Numerical simulations

In this subsection, we consider two types of linear quantile regression models: high-dimensional spare model and high-dimensional dense model. We illustrate the performance comparison between K-step quantile forward selection, l_1-penalized quantile regression, and the oracle quantile regression estimators. The following gives a brief introduction to each estimator.

1. K-step QFR: K-threshold quantile forward selection in algorithm 1. In the simulation, the tuning parameter K is determined by a quantile-type modified BIC from Wang et al. (2009). More explicitly, the number of step K is chosen by

$$K_{BIC}=\arg\min_{K\in\mathbb{N}}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{S_K}\right)+\frac{K\log\log p\log n}{n}\right\}$$

2. l_1-QR: The l_1-penalized quantile regression estimator is defined as follows. The tuning parameter λ is chosen by the BIC in "rqPen" R package.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} \right) + \lambda \left| \boldsymbol{\beta} \right|_1$$

3. Oracle: If the model is sparse, S^* is assumed to be known and we use traditional quantile regression. As for the dense model, we do not have the oracle estimator.

The data generating process follows the following linear quantile regression setup. For each $i = 1, \ldots, n$, we let

$$\begin{aligned} y_i &= \boldsymbol{x}_i^T \boldsymbol{\beta}_0 + \varepsilon_i \\ x_{ij} &\sim N(0,1), corr(x_{ij}, x_{ik}) = \rho^{|j-k|} \\ \varepsilon_i &\sim N(0,\sigma) - \sigma \Phi^{-1}(\tau) \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, ε_i are independent across i, and Φ^{-1} is the inverse cumulative density function of standard normal distribution. Under this setup, it is obvious that linear conditional quantile condition of y, $Q_y(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_0$, is satisfied. We consider two sets of simulations for both sparse and dense models. We run $S = 1000$ replications for each DGP. We let $n = 50$, $p = 100$ or $n = 100$, $p = 200$ to mimic the high dimensionality issue. Different choices of noise ratio $\sigma = 0.1, 1$ and correlation levels between covariates $\rho = 0, 0.5$ are also imposed.

In the sparse model setup, we set the number of true covariates, with $s_0 = 4$, and let corresponding coefficients $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$. For other coefficients, we let $\beta_{i>4} = 0$. As for the dense model, we let $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ as before but set $\beta_{i>4} \sim Unif(-0.1, 0.1)$.

To compare the performances among K-step QFR and l_1-QR, we use the following measurements for the sparse model.

1. The average mean quantile prediction error (MQPE) defined as $\frac{1}{S} \sum_{s=1}^{S} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(\boldsymbol{x}_i^T \boldsymbol{\beta}_0 - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} \right) \right\}^{(s)}$.
2. The exact selection rate (ESR) in the sense that $r_e = 1$ if $\hat{S} = S^* = \{1, 2, 3, 4\}$; otherwise, $r_e = 0$. We calculate the average exact selection rate by $\frac{1}{S} \sum_{s=1}^{S} r_e^{(s)}$.
3. The selection rate (SR) in the sense that $r_{sr} = 1$ if $S^* = \{1, 2, 3, 4\} \subseteq \hat{S}$; otherwise, $r_{sr} = 0$. We calculate the average selection rate by $\frac{1}{S} \sum_{s=1}^{S} r_{sr}^{(s)}$.

For the dense model, we consider the following two criteria.

1. The mean quantile prediction error (MQPE) defined as $\frac{1}{S} \sum_{s=1}^{S} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(\boldsymbol{x}_i^T \boldsymbol{\beta}_0 - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} \right) \right\}^{(s)}$.
2. The average number of selected covariates (NS) as $\frac{1}{S} \sum_{s=1}^{S} \left\| \hat{\boldsymbol{\beta}} \right\|_0^{(s)}$.

The following tables 1, 2, and 3 show our comparison results for sparse and dense setup.

Table 1: Sparse model: MQPE, ESR and SR comparison among K-step QFR, l_1-QR, and Oracle model

		$\sigma = 0.1, \rho = 0$					
		$n = 50, p = 100$			$n = 100, p = 200$		
τ		MQPE	ESR	SR	MQPE	ESR	SR
0.05	K-step QFR	0.0477	0.265	0.895	0.0262	0.2	0.999
	l_1-QR	0.0861	0.015	1	0.0548	0.022	1
	Oracle	0.0255	1	1	0.0169	1	1
0.50	K-step QFR	0.0174	0.627	1	0.0113	0.806	1
	l_1-QR	0.0822	0.029	1	0.0468	0.039	1
	Oracle	0.0134	1	1	0.0096	1	1
0.95	K-step QFR	0.0490	0.259	0.866	0.0260	0.19	1
	l_1-QR	0.0852	0.012	1	0.0530	0.029	1
	Oracle	0.0254	1	1	0.0168	1	1
		$\sigma = 1, \rho = 0$					
		$n = 50, p = 100$			$n = 100, p = 200$		
τ		MQPE	ESR	SR	MQPE	ESR	SR
0.05	K-step QFR	0.4492	0.136	0.387	0.2672	0.192	0.952
	l_1-QR	0.4941	0	0.42	0.3940	0	0.91
	Oracle	0.2553	1	1	0.1690	1	1
0.50	K-step QFR	0.1964	0.541	0.93	0.1136	0.804	1
	l_1-QR	0.3268	0.005	0.962	0.2787	0.012	0.998
	Oracle	0.1339	1	1	0.0960	1	1
0.95	K-step QFR	0.4441	0.136	0.379	0.2663	0.189	0.956
	l_1-QR	0.4978	0.001	0.434	0.3939	0.001	0.895
	Oracle	0.2541	1	1	0.1679	1	1

Note: This table is based on 1000 replications for each cell. Every number in the MQPE, ESR, and SR columns report the average mean quantile prediction error, exact selection rate, and selection rate correspondingly following the definition above. The model in each case follows simulation model (1) with n, p, σ, ρ defined on the top of the table and one of the three methods (K-step QFR, l_1-QR, and Oracle) on the left. For each case, we examine $\tau \in \{0.05, 0.50, 0.90\}$.

Table 2: Sparse model: MQPE, ESR and SR comparison among K-step QFR, l_1-QR, and Oracle model

		$\sigma = 0.1, \rho = 0.5$					
		$n = 50, p = 100$			$n = 100, p = 200$		
τ		MQPE	ESR	SR	MQPE	ESR	SR
0.05	K-step QFR	0.0381	0.371	0.989	0.0258	0.215	1
	l_1-QR	0.0950	0.28	0.998	0.0565	0.233	1
	Oracle	0.0255	1	1	0.0169	1	1
0.50	K-step QFR	0.0170	0.652	1	0.0113	0.826	1
	l_1-QR	0.0940	0.327	1	0.0513	0.318	1
	Oracle	0.0134	1	1	0.0096	1	1
0.95	K-step QFR	0.0382	0.354	0.989	0.0262	0.197	1
	l_1-QR	0.0947	0.251	0.998	0.0558	0.265	1
	Oracle	0.0254	1	1	0.0168	1	1
		$\sigma = 1, \rho = 0.5$					
		$n = 50, p = 100$			$n = 100, p = 200$		
τ		MQPE	ESR	SR	MQPE	ESR	SR
0.05	K-step QFR	0.4198	0.111	0.278	0.2739	0.196	0.814
	l_1-QR	0.4133	0.027	0.679	0.3165	0.027	0.971
	Oracle	0.2553	1	1	0.1690	1	1
0.50	K-step QFR	0.2049	0.506	0.778	0.1143	0.816	0.989
	l_1-QR	0.2798	0.07	0.985	0.2240	0.136	1
	Oracle	0.1339	1	1	0.0960	1	1
0.95	K-step QFR	0.4208	0.112	0.29	0.2786	0.172	0.815
	l_1-QR	0.4108	0.043	0.718	0.3184	0.024	0.959
	Oracle	0.2541	1	1	0.1679	1	1

Note: This table is based on 1000 replications for each cell. Every number in the MQPE, ESR, and SR columns report the average mean quantile prediction error, exact selection rate, and selection rate correspondingly following the definition above. The model in each case follows simulation model (1) with n, p, σ, ρ defined on the top of the table and one of the three methods (K-step QFR, l_1-QR, and Oracle) on the left. For each case, we examine $\tau \in \{0.05, 0.50, 0.90\}$.

Table 3: Dense model: MQPE, ESR and SR comparison between K-step QFR and l_1-QR

		$\sigma = 1, \rho = 0$				$\sigma = 1, \rho = 0.5$			
		$n = 50, p = 100$		$n = 100, p = 200$		$n = 50, p = 100$		$n = 100, p = 200$	
τ		MQPE	NS	MQPE	NS	MQPE	NS	MQPE	NS
0.05	K-step QFR	0.4885	4.703	0.3994	6.526	0.4558	4.384	0.4065	6.401
	l_1-QR	0.5257	10.761	0.4676	18.183	0.4576	8.955	0.4191	14.967
0.50	K-step QFR	0.3332	4.319	0.3246	4.401	0.3420	4.095	0.3454	4.461
	l_1-QR	0.3711	24.622	0.3814	38.253	0.3407	17.041	0.3708	26.834
0.95	K-step QFR	0.4857	4.672	0.3980	6.498	0.4612	4.460	0.4085	6.377
	l_1-QR	0.5247	10.72	0.4732	18.488	0.4553	8.871	0.4200	15.156

Note: This table is based on 1000 replications for each cell. Every number in the MQPE and NS columns report the average mean quantile prediction error and number of selected variables correspondingly following the definition above. The model in each case follows simulation model (1) with n, p, σ, ρ defined on the top of the table and one of the two methods (K-step QFR and l_1-QR) on the left. For each case, we examine $\tau \in \{0.05, 0.50, 0.90\}$.

For the sparse model, the results are shown in table 1 and 2. The K-step QFR outperforms l_1-QR in terms of MQPE in most scenarios except the case $\sigma = 1, \rho = 0.5$ with small sample size $n = 50, p = 100$ and extreme $\tau = 0.05, 0.95$. Our explanation is that when sample size n is small, the number of effective samples for quantile regression at tail quantiles ($\tau = 0.05, 0.95$ in our simulation) is small; hence the estimation of quantile regression in every step of QFR becomes unstable. Intuitively, this issue should be alleviated when we increase the sample size n. We can see this pattern in which K-step QFR performs better with $n = 100, p = 200$. In addition, the exact selection ratio (ESR) of K-step QFR is better than l_1-QR across all cases. We find that l_1 penalization method can hardly choose the exact number of true covariates, and it always tends to include many irrelevant covariates. This characteristic is consistent with high-dimensional literature, and it impacts the prediction performance of l_1-QR. However, for K-step QFR, it always has some chance to select the true covariates exactly, especially when $\tau = 0.5$.

For the dense model results in table 3, it is clear that K-step QFR dominates l_1-QR in terms of MQPE performance across all cases in table 3. Note that the dense model is trying to imitate the model misspecification issue, so we do not have the Oracle method. In our dense DGP, we set the 4 most important independent variables. As we can see in the table, the average number of selected variables is between 4 to 7 for K-step QFR with modified BIC and l_1-penalized QR selects much more covariates. One possible explanation is that we directly choose our number of steps K in K-step QFR while l_1-QR selects the penalty parameter, which is λ. The penalty parameter indirectly controls the number of selected variables. Therefore, with the dense model setup, l_1-QR is more likely to be affected by those small but unimportant coefficients.

5.2 Application-based simulations

To complement our pure numerical simulation, we consider another experiment that is based on the real-world dataset. In the empirical dataset, the cross-sectional dependence across variables or

predictors usually exists and might have an impact on econometric methods. In macroeconomics, a lot of independent variables or predictors are well-structured and hence highly correlated, e.g., the producer price index for industrial commodities, intermediate products and finished consumer goods. This problem is even more severer when the dimension of covariates is large. Therefore, it is reasonable to investigate finite sample performance with simulations based on real data.

Our simulation is based on Fred-QD dataset from McCracken and Ng (2020) (see more description in the next empirical application section). We set $n = 194$ and $p = 231$, which are the same as the empirical application part. We focus on $\tau = 0.05$ and assume a linear model as follows

$$y_i = \alpha_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_0 + \varepsilon_i$$

where the variance-covariance matrix of $\boldsymbol{x}_i$ is calculated from the empirical dataset. We assume $Q_\tau(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_0$ and set $x_{ij} \sim N(0,1)$ and independent error term $\varepsilon_i \sim N(0,1)$. Similarly, we consider both sparse and dense models. For the sparse model, we randomly set 4 $\beta_j = 1$ for $j \in [p]$ in every simulation and set other β_js equal to 0. For the dense model, those un-selected β_js follows $Unif(-0.05, 0.05)$. The number of simulations S is 200 for both models. We evaluate the performance with MQPE and NS as the dense model setting in the following table 4.

Table 4: Application-based simulation: MQPE and NS comparison between K-step QFR and l_1-QR

	Sparse DGP		Dense DGP	
$\tau = 0.05$	MQPE	NS	MQPE	NS
K-step QFR	0.0803	6.6	0.0829	7.075
l_1-QR	0.1386	72.159	0.1508	79.05

Note: This table is based on 200 replications for each cell. Every number in the MQPE and NS columns report the average mean quantile prediction error and number of selected variables correspondingly following the definition above. For each case, we examine the case with $\tau = 0.05$.

Overall, the performance of K-step QFR is satisfactory compared with l_1-QR. In sum, K-step QFR has a smaller average prediction error, and the number of selected covariates is close to the true or important covariates, which is 4 in our setup. This result confirms that K-step QFR behaves well in the model selection and prediction accuracy even if there are high cross-sectional dependence structures among the covariates.

6 Empirical applications

We conduct two empirical applications in this section. The first part is growth-at-risk (GaR) forecasting. We consider predicting the tail risk of the GDP growth rate with a high dimensional Fred dataset from McCracken and Ng (2020). In the second part, we investigate the international economic growth convergence hypothesis from the quantile regression aspect with the famous cross-sectional dataset in Barro and Lee (1994).

6.1 Growth-at-Risk prediction example with Fred dataset

In this subsection, we investigate an empirical application of the prediction of growth-at-risk. GaR measures the extremely negative impact, i.e., the left tail risk, on the GDP growth rate, and it attracts more attention for empirical macroeconomic researchers nowadays. Adrian et al. (2019) first use simple quantile regression with the national financial condition index (NFCI) as the key variable to predict the GaR. Brownlees and Souza (2021) further compare the prediction performance of quantile regression with traditional volatility models, like GARCH, in this topic. Plagborg-Møller et al. (2020) use quantile regression with extracted financial and global factors to argue the predictability of GaR. Unlike previous research only considering one or a few variables or factors, our empirical example utilizes high dimensional data. With the assistance of abundant data, we can first consider choosing influential predictors and then making proper forecasting on GaR. Our data source is the Fred-QD dataset, which is provided by McCracken and Ng (2020). This dataset contains the quarterly GDP growth rate of the United States as the response variable and $p = 231$ time series for explanatory variables. We transform all variables into stationary series by the instructions in McCracken and Ng (2016). In addition, since macroeconomic variables are highly correlated if they belong to the same statistical category, a suitable variable selection approach seems important. To make use of as many predictors as possible, we set the truncated sample period from 1973Q1 to 2021Q2 with $n = 194 < p$, so this application suffers high dimensionality issue. We consider the linear predictive quantile regression at $\tau = 0.05$

$$Q_{\tau=0.05}\left(y_t|\mathcal{F}_{t-1}\right) = \alpha_0 + \boldsymbol{x}_{t-1}^T\boldsymbol{\beta}_0$$

In the application, we use a moving window with fixed 100 quarters as in-sample periods and compare both one-period and four-periods ahead prediction performance between K-step simple forward quantile regression and l_1-QR. For these methods, we use modified BIC to select the tuning parameter K and consider BIC to select the tuning parameter λ. We report the average prediction error and the average number of selected variables in the table 5 below.

Table 5: GaR application: MQPE and NS comparison between K-step QFR and l_1-QR

	1 period ahead		4 periods ahead	
$\tau = 0.05$	MQPE	NS	MQPE	NS
K-step QFR	0.144	3.828	0.193	3.8
l_1-QR	0.406	99	0.615	99

Note: The entries in this table are based on 94 fixed window prediction periods for 1-period ahead forecasting or 91 periods for 4-periods ahead forecasting. Every number in the MQPE and NS columns report the average mean quantile prediction error and number of selected variables correspondingly following the definition above. For each case, we examine the case with $\tau = 0.05$.

In sum, we confirm that K-step QFR shows stable prediction performance and select a relatively small number of important covariates. It outperforms l_1-QR in both criteria, and this result is consistent with our extensive simulations. We also show that l_1-QR is not very helpful in variable

dimension reduction while encountering real-world macroeconomic data.

6.2 Cross-sectional international growth example in Barro and Lee (1994)

We examine an international economic growth example from Barro and Lee (1994) in the subsection. Barro and Lee (1994) consider the effect of the initial level of per capita GDP on the GDP per capita growth rate. If this effect is negative empirically, the result should suggest a convergence of growth rates among the poor nations and rich nations. This is also named as convergence hypothesis. Barro and Sala-i Martin (1995) use a simple bivariate regression to investigate this effect, and their results reject the theoretical hypothesis. Since then, much literature has tried to include other cross-sectional predictors and examine whether this effect holds conditioning on other variables. Besides the literature focusing on the least square framework, Koenker and Machado (1999) investigate the quantile regression effect (the international economic growth risk) of this issue and provide explanations based on the quantile type goodness-of-fit. The quantile regression approach can provide more insights in terms of robustness.

Recently, since the number of predictors that the literature include is relatively large compared with the sample size, some researchers have started to use high-dimensional tools to examine this question, see Giannone et al. (2021). When the dimension of covariates is comparably high, a suitable variable section approach becomes a critical issue. For example, Belloni et al. (2011) use the Post-square-root-Lasso method and the Post-double-selection method to provide variable selection results for the international growth least square regression. In our case, we focus on the economic growth risk from the quantile perspective and apply our K-step quantile forward regression to conduct variable selection issues.

Our data source is from the "GrowthData" dataset in the "hdm" r-package provided by Chernozhukov et al. (2016). It is a cross-sectional international growth dataset that contains $n = 90$ observations and $p = 62$ covariates. We consider linear quantile regression model for $i = 1, \ldots, n$ as the following,

$$Q_\tau\left(y_i|d_i, x_i\right) = \delta_0 + \alpha_0 d_i + \sum_{j=1}^{p-1} \beta_{0j} x_j$$

where y_i is the average GDP growth rate between 1960-1985 for each country as the response variable, d_i is the logarithm of the nation's GDP in 1960 (initial GDP level), and $x_j, j = 1, \ldots, 61$ denotes other control covariates which measure different nation-specific characteristics in pre-1960 period, including national accounts, education, trade policy, political circumstance and others. α_0 and β_{0j} are corresponding parameters and δ_0 is the intercept term. Since we plan to check the quantile effect of the logarithm of initial GDP level d_i on the average growth rate y_i, d_i is always included in the model, and we use quantile forward regression algorithm on other covariates x_j. The application considers 19 evenly distributed choices of τ from 0.05 to 0.95. We show the results from K-step quantile forward regression for covariate selection. Since p is not large than n, we use BIC to select the tuning parameters K. We report the covariate selection results for different $\tau = 0.05, 0.25, 0.50, 0.75, 0.95$ in table 6 below. A complete result for all τ is provided in the

appendix.

Table 6: Variable selection result for the international growth example with K-step quantile forward regression (Real per capita GDP (log) is always included)

τ	Additional variables selected
0.05	Black market premium (log) Population proportion over 64 Terms of trade shock Percentage of "no schooling" in the male population Infant mortality rate Ratio of nominal government current expenditure on the defense to nominal GDP
0.25	Black market premium (log)
0.50	Black market premium (log)
0.75	Total gross enrollment ratio for primary education
0.95	Total gross enrollment ratio for primary education Ratio of domestic investment to real GDP Percentage of "no schooling" in the female population Infant mortality rate Exchange rate

Our variable selection result sheds some light on the choice of conditioning variables. For the downside economic growth risk, the logarithm of black market premium shows strong signals. In Belloni et al. (2011), they also select this covariate in the least square regression by the square-root Lasso. Therefore, the logarithm of black market premium not only has an impact on the mean level but also influences the lower quantile levels. This finding is also consistent with the result in Koenker and Machado (1999), suggesting market openness in a country plays an important role in long-term economic growth distribution. For the upper quantiles, we find that the total gross enrollment ratio for primary education seems important. This indicates education determines the potential level of high economic growth. Based on our additional selected variables, we can further examine the estimated quantile regression effect of the initial logarithm of per capita GDP (d_i) on the economic growth rate (y_i). The following table 7 and figure 1 shows squared root Lasso the estimated coefficients $\hat{\alpha}$ and its confidence bands across τ. In the figure 1, the solid red line is the estimated quantile regression effect $\hat{\alpha}$ as table 7. The dashed light red line and dashed dark red line are the upper and lower 95% confidence bands. The black dashed line is a horizontal line at -0.0112, indicating the coefficient estimated by post-square-root Lasso in Belloni et al. (2011). Our estimated $\hat{\alpha}$ is negative for all quantile levels, which is consistent with the convergence hypothesis that the poor countries grow faster and the rich countrie grows slower. Our results complement the findings of Belloni et al. (2011). Besides that, the quantile regression approach indicates more on the conditional distribution compared with the mean regression. Since our results suggest a decrease in $\hat{\alpha}$ for larger τ, this indicates a stronger convergence on the right tail of growth rate than the left tail. This evidence confirms the results in Koenker and Machado (1999).

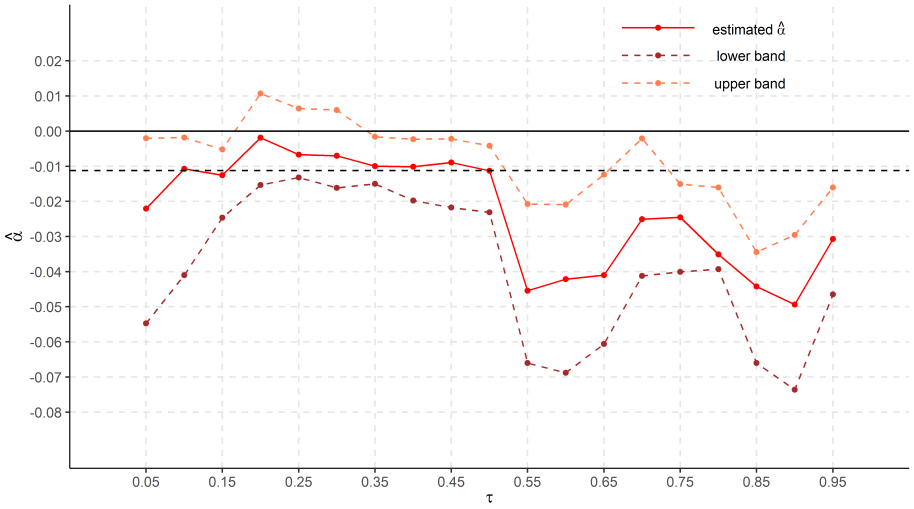

Figure 1: Estimated quantile regression effect $\hat{\alpha}$ and 95% confidence bands

Note: The figure illustrates the estimated quantile regression effect $\hat{\alpha}$ across all τ from 0.05 to 0.95. $\hat{\alpha}$ (red solid line) measures the effect of the initial level of GDP per capita on the per capital GDP growth rate. The two red dashed lines are the 95% confidence interval generated in the post selection. In most cases, theses bands exclude the null effect indicated by the horizontal 0 line. The black dashed line is the estimated effect of mean regression by the Post-square-root-Lasso in Belloni et al. (2011).

Table 7: The estimated coefficients $\hat{\alpha}$ given selected variables for different τ

τ	$\hat{\alpha}$	τ	$\hat{\alpha}$	τ	$\hat{\alpha}$	τ	$\hat{\alpha}$
0.05	-0.0221	0.30	-0.0070	0.55	-0.0454	0.80	-0.0351
0.10	-0.0107	0.35	-0.0099	0.60	-0.0422	0.85	-0.0442
0.15	-0.0125	0.40	-0.010	0.65	-0.0410	0.90	-0.0494
0.20	-0.0019	0.45	-0.0088	0.70	-0.0251	0.95	-0.0307
0.25	-0.0067	0.50	-0.0112	0.75	-0.0245		

Note: This table reports the estimated quantile effect $\hat{\alpha}$ cross all $\tau = 0.05, \ldots, 0.95$. The numbers corresponds to the red solid line in figure 1.

As a summary, we utilize quantile forward regression to choose significant covariates in the international economic growth application. Our finding supports and provides more information on the existing literature, which consider testing the convergence hypothesis in Barro and Lee (1994).

7 Conclusion

In this paper, we investigate the theoretical properties of quantile forward regressions, including K-step quantile forward regression and t-threshold quantile forward regression. We show that quantile forward regressions are useful in dealing with high-dimensionality issues in the linear quantile regression model. Under the non-asymptotic framework, we are able to show that the quantile forward

regressions satisfy similar performance bounds with the least square forward selection. From the asymptotic view, K-step quantile forward regression also satisfies desirable properties like the same $O\left(\sqrt{\frac{\log p}{n}}\right)$ convergence rate as penalization methods and prediction consistency. Numerical simulations illustrate that the quantile forward regressions have better performance in terms of prediction than l_1-penalized methods under either simple or complex data structures. These advantages indicate that quantile forward regression is a user-friendly and convenient tool for empirical researchers. Two empirical applications demonstrate its usage.

Finally, it is worth mentioning that there are many possible extensions for quantile forward regression. First, we only consider the asymptotic results for K-step quantile forward regression and omit the asymptotic analysis for t-threshold quantile forward regression. The connection between the two types of forward regression algorithms is generally unknown. It would be interesting if researchers could find a suitable asymptotic framework that links these two types together. Second, it is also interesting to ask whether there is variable selection consistency, like in Fan and Li (2001), Zou (2006), or Lahiri (2021), for forward selection algorithms. If so, it is important to discover the underlying conditions and derive the result on selection consistency. Third, our analysis is based on cross-sectional data. It is also curious to extend quantile forward regression into other data structures, like time-series data or panel data. There is some recent development on the least square estimator, see Ing (2020), but it is an open question for the general regression models. We leave those extensions for future research.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70, 91–117.

Abrevaya, J. (2002): "The effects of demographics and maternal behavior on the distribution of birth outcomes," in *Economic applications of quantile regression*, Springer, 247–257.

Adrian, T., N. Boyarchenko, and D. Giannone (2019): "Vulnerable growth," *American Economic Review*, 109, 1263–89.

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006): "Quantile regression under misspecification, with an application to the US wage structure," *Econometrica*, 74, 539–563.

Arellano, M. and S. Bonhomme (2017): "Quantile selection models with an application to understanding changes in wage inequality," *Econometrica*, 85, 1–28.

Athey, S. and G. W. Imbens (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685–725.

Barro, R. J. and J.-W. Lee (1994): "Sources of economic growth," in *Carnegie-Rochester conference series on public policy*, Elsevier, vol. 40, 1–46.

Barro, R. J. and X. I. Sala-i Martin (1995): *Economic growth*.

Belloni, A. and V. Chernozhukov (2011): "L1-penalized quantile regression in high-dimensional sparse models," *The Annals of Statistics*, 39, 82–130.

Belloni, A., V. Chernozhukov, and C. Hansen (2011): "Inference for high-dimensional sparse econometric models," *in Advances in Economics and Econometrics-World Congress of Econometric Society 2010*.

Brownlees, C. and A. B. Souza (2021): "Backtesting global growth-at-risk," *Journal of Monetary Economics*, 118, 312–330.

Bühlmann, P. and S. Van De Geer (2011): *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Chamberlain, G. (1994): "Quantile regression, censoring, and the structure of wages," in *Advances in econometrics: sixth world congress*, vol. 2, 171–209.

Chen, L.-Y. and S. Lee (2020): "Sparse Quantile Regression," *arXiv preprint arXiv:2006.11201*.

Chernozhukov, V. and C. Hansen (2004): "The effects of 401 (k) participation on the wealth distribution: an instrumental quantile regression analysis," *Review of Economics and statistics*, 86, 735–751.

Chernozhukov, V., C. Hansen, and M. Spindler (2016): "High-dimensional metrics in R," *arXiv preprint arXiv:1603.01700*.

Das, A. and D. Kempe (2011): "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," *arXiv preprint arXiv:1102.3975*.

——— (2018): "Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection," *The Journal of Machine Learning Research*, 19, 74–107.

Elenberg, E. R., R. Khanna, A. G. Dimakis, and S. Negahban (2018): "Restricted strong convexity implies weak submodularity," *The Annals of Statistics*, 46, 3539–3568.

Fan, J. and R. Li (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, 96, 1348–1360.

Fan, J., J. Lv, and L. Qi (2011): "Sparse high-dimensional models in economics," *Annu. Rev. Econ.*, 3, 291–317.

Fan, R., J. H. Lee, and Y. Shin (2021): "Predictive Quantile Regression with Mixed Roots and Increasing Dimensions," *arXiv preprint arXiv:2101.11568*.

Giannone, D., M. Lenza, and G. E. Primiceri (2021): "Economic predictions with big data: The illusion of sparsity," *Econometrica*, 89, 2409–2437.

Ing, C.-K. (2020): "Model selection for high-dimensional linear regression with dependent observations," *Annals of Statistics*, 48, 1959–1980.

Kato, K. (2011): "Group Lasso for high dimensional sparse quantile regression models," *arXiv preprint arXiv:1103.1458*.

Knight, K. (1998): "Limiting distributions for L1 regression estimators under general conditions," *Annals of statistics*, 755–770.

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

——— (2017): "Quantile regression: 40 years on," *Annual Review of Economics*, 9, 155–176.

Koenker, R. and G. Bassett (1978): "Regression quantiles," *Econometrica: journal of the Econometric Society*, 33–50.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017): "Handbook of quantile regression," .

Koenker, R. and J. A. Machado (1999): "Goodness of fit and related inference processes for quantile regression," *Journal of the american statistical association*, 94, 1296–1310.

Kozbur, D. (2020): "Analysis of Testing-Based Forward Model Selection," *Econometrica*, 88, 2147–2173.

Lahiri, S. N. (2021): "Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions," *The Annals of Statistics*, 49, 820–844.

Lee, J. H. and Y. Shin (2021): "Complete subset averaging for quantile regressions," *Econometric Theory*, 3.

Lu, X. and L. Su (2015): "Jackknife model averaging for quantile regressions," *Journal of Econometrics*, 188, 40–58.

Ma, S., R. Li, and C.-L. Tsai (2017): "Variable screening via quantile partial correlation," *Journal of the American Statistical Association*, 112, 650–663.

McCracken, M. and S. Ng (2020): "FRED-QD: A quarterly database for macroeconomic research," Tech. rep., National Bureau of Economic Research.

McCracken, M. W. and S. Ng (2016): "FRED-MD: A monthly database for macroeconomic research," *Journal of Business & Economic Statistics*, 34, 574–589.

Miller, A. (2002): *Subset selection in regression*, chapman and hall/CRC.

Plagborg-Møller, M., L. Reichlin, G. Ricco, and T. Hasenzagl (2020): "When is growth at risk?" *Brookings Papers on Economic Activity*, 2020, 167–229.

Sancetta, A. et al. (2016): "Greedy algorithms for prediction," *Bernoulli*, 22, 1227–1277.

Shi, Z. and J. Huang (2021): "Forward-selected panel data approach for program evaluation," *Journal of Econometrics*.

Shibata, R. (1981): "An optimal selection of regression variables," *Biometrika*, 68, 45–54.

——— (1982): "Amendments and corrections: An optimal selection of regression variables," *Biometrika*, 69, 492–492.

Wang, H. (2009): "Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, 104, 1512–1524.

Wang, H., B. Li, and C. Leng (2009): "Shrinkage tuning parameter selection with a diverging number of parameters," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 671–683.

Wang, L. (2013): "The L1 penalized LAD estimator for high dimensional linear regression," *Journal of Multivariate Analysis*, 120, 135–151.

Wu, Y. and Y. Liu (2009): "Variable selection in quantile regression," *Statistica Sinica*, 801–817.

Xiao, Z., H. Guo, and M. S. Lam (2015): "Quantile regression and value at risk," in *Handbook of Financial Econometrics and Statistics*, Springer, 1143–1167.

Zhang, T. (2009): "On the consistency of feature selection using greedy least squares regression." *Journal of Machine Learning Research*, 10.

Zheng, Q., C. Gallagher, and K. Kulasekera (2013): "Adaptive penalized quantile regression for high dimensional data," *Journal of Statistical Planning and Inference*, 143, 1029–1038.

Zheng, Q., L. Peng, and X. He (2015): "Globally adaptive quantile regression with ultra-high dimensional data," *Annals of statistics*, 43, 2225.

Zou, H. (2006): "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, 101, 1418–1429.

A Proofs

A.1 Proof of lemma 3.1

Proof. Denote $\boldsymbol{x} = \boldsymbol{x}_{\cdot}$. For $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \Omega \subset \mathbb{R}^p$, following the definition 3 in Elenberg et al. (2018), we consider

$$D = R_\tau^1(\boldsymbol{\beta}_2) - R_\tau^1(\boldsymbol{\beta}_1) - \left\langle \frac{E\left[\boldsymbol{x}^T\left(\tau - 1\left(y - \boldsymbol{x}^T\boldsymbol{\beta}_1 < 0\right)\right)\right]}{E\left[\rho_\tau\left(y-\alpha\right)\right]}, \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \right\rangle$$

Using famous Knight's identity from Knight (1998), we have

$$\begin{aligned}
D =& -\frac{1}{E\left[\rho_\tau\left(y-\alpha\right)\right]} E\left[\rho_\tau\left(y - \boldsymbol{x}^T\boldsymbol{\beta}_2\right) - \rho_\tau\left(y - \boldsymbol{x}^T\boldsymbol{\beta}_1\right)\right] - \left\langle E\left[\boldsymbol{x}^T\left(\tau - 1\left(y - \boldsymbol{x}^T\boldsymbol{\beta}_1 < 0\right)\right)\right], \beta_2 - \beta_1 \right\rangle \\
=& \frac{1}{E\left[\rho_\tau\left(y-\alpha\right)\right]} E\left[\boldsymbol{x}^{\boldsymbol{T}}\left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right)\psi_\tau\left(y - \boldsymbol{x}^T\boldsymbol{\beta}_1\right)\right] - \\
& \frac{1}{E\left[\rho_\tau\left(y-\alpha\right)\right]} E \int_0^{\boldsymbol{x}^T\left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right)} 1\left(y - \boldsymbol{x}^T\boldsymbol{\beta}_1 \le s\right) - 1\left(y - \boldsymbol{x}^T\boldsymbol{\beta}_1 \le 0\right) ds \\
& - \frac{1}{E\left[\rho_\tau\left(y-\alpha\right)\right]} \left\langle E\left[\boldsymbol{x}^T\left(\tau - 1\left(y - \boldsymbol{x}^T\boldsymbol{\beta}_1 < 0\right)\right)\right], \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \right\rangle \\
=& -\frac{1}{E\left[\rho_\tau\left(y-\alpha\right)\right]} E \int_0^{\boldsymbol{x}^T\left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right)} F_{u_1}\left(s|\boldsymbol{x}\right) - F_{u_1}\left(0|\boldsymbol{x}\right) ds \\
=& -\frac{1}{2E\left[\rho_\tau\left(y-\alpha\right)\right]} \left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right)^T E f_{u_1}\left(\xi|\boldsymbol{x}\right) \boldsymbol{x}\boldsymbol{x}^T \left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right) \\
=& -\frac{1}{2E\left[\rho_\tau\left(y-\alpha\right)\right]} \left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right)^T E f_y\left(\xi + \boldsymbol{x}^T\boldsymbol{\beta}_1|\boldsymbol{x}\right) \boldsymbol{x}\boldsymbol{x}^T \left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right) \\
=& -\frac{1}{2E\left[\rho_\tau\left(y-\alpha\right)\right]} \left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right)^T E f_y\left(\xi^{'}|\boldsymbol{x}\right) \boldsymbol{x}\boldsymbol{x}^T \left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right)
\end{aligned}$$

where $u_1 = y - \boldsymbol{x}^T\boldsymbol{\beta}_1$, and $\xi \in \left(0, \boldsymbol{x}^T\left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right)\right)$, $\xi^{'} \in \left(0, \boldsymbol{x}^T\boldsymbol{\beta}_2\right)$.

Under the assumption above and using Rayleigh quotient with $0 < \lambda_{\min}\left(\boldsymbol{J}\right) \le \boldsymbol{J} = E f_y\left(\xi^{'}|\boldsymbol{x}\right)\boldsymbol{x}\boldsymbol{x}^T \le \lambda_{\max}\left(\boldsymbol{J}\right) < \infty$ in the support of $(y, \boldsymbol{x})$, we have

$$-\frac{\lambda_{\min}\left(\boldsymbol{J}_S\right)}{2E\left[\rho_\tau\left(y-\alpha\right)\right]} \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right\|_2^2 \ge D \ge -\frac{\lambda_{\max}\left(\boldsymbol{J}_S\right)}{2E\left[\rho_\tau\left(y-\alpha\right)\right]} \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right\|_2^2$$

So $R_\tau^1(\cdot)$ satisfying restricted strong concavity with $\frac{\lambda_{\min}(\boldsymbol{J}_S)}{E[\rho_\tau(y-\alpha)]}$ and restricted smoothness with $\frac{\lambda_{\max}(\boldsymbol{J}_S)}{E[\rho_\tau(y-\alpha)]}$ on Ω. □

A.2 Proof of corollary 3.1

Proof. By theorem 3.2, we know that

$$
\begin{aligned}
&Pr\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,S_K}^T\boldsymbol{\beta}_{S_K}\right)\leq\left(1-e^{-\frac{\phi_{\min}(2K)(\boldsymbol{J})}{\phi_{\max}(2K)(\boldsymbol{J})}}\right)E\left[\rho_\tau\left(y-\boldsymbol{x}_{S^*}^T\boldsymbol{\beta}_{S^*}\right)\right]+e^{-\frac{\phi_{\min}(2K)(\boldsymbol{J})}{\phi_{\max}(2K)(\boldsymbol{J})}}E\left[\rho_\tau\left(y-a\right)\right]+\varepsilon\right]\\
\geq&Pr\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,S_K}^T\boldsymbol{\beta}_{S_K}\right)-E\left[\rho_\tau\left(y-\boldsymbol{x}_{S_K}^T\boldsymbol{\beta}_{S_K}\right)\right]\leq\varepsilon\right]\\
\geq&1-e^{-\frac{n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}E\left(\rho_\tau^2\left(y_i-\boldsymbol{x}_{i,S_K}^T\boldsymbol{\beta}_{S_K}\right)\right)+\frac{\tau\vee(1-\tau)\varepsilon}{3}}}\\
\geq&1-e^{-\frac{n\varepsilon^2}{\tau\vee(1-\tau)\left(1+\frac{\varepsilon}{3}\right)}}
\end{aligned}
$$

where the second last inequality is from Bernstein inequality for i.i.d. random variables. □

A.3 Proof of lemma 3.2

Proof. From the theorem 3.1 and $\left|\hat{S}\right|=K$, we know the weak submodularity ratio is bounded below by

$$\gamma_{\hat{S},s_0}\geq\frac{\phi_{\min}\left(K+s_0\right)(\boldsymbol{J})}{\phi_{\max}\left(K+s_0\right)(\boldsymbol{J})}$$

By the definition of weak submodularity ratio we have

$$\frac{\phi_{\min}\left(K+s_0\right)(\boldsymbol{J})}{\phi_{\max}\left(K+s_0\right)(\boldsymbol{J})}\leq\gamma_{\hat{S},s_0}\leq\frac{\sum_{j\in S^*\setminus\hat{S}}f\left(\hat{S}\cup\{j\}\right)-f\left(S\right)}{f\left(\hat{S}\cup S^*\right)-f\left(\hat{S}\right)}$$

Since $f\left(\hat{S}\cup\{j\}\right)-f\left(\hat{S}\right)<t$, we know

$$
\begin{aligned}
f\left(\hat{S}\cup S^*\right)-f\left(\hat{S}\right)&\leq\frac{\phi_{\max}\left(K+s_0\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_0\right)(\boldsymbol{J})}\sum_{j\in S^*\setminus\hat{S}}f\left(\hat{S}\cup\{j\}\right)-f\left(S\right)\\
&\leq\frac{\phi_{\max}\left(K+s_0\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_0\right)(\boldsymbol{J})}s_0t
\end{aligned}
$$

□

A.4 Proof of lemma 3.3

Proof. Since f is a non-decreasing monotone function by definition, the basic inequality holds naturally. □

A.5 Proof of theorem 3.3

Proof. By lemma 3.2 and 3.3, we have

$$f\left(S^{*}\right)-f\left(\hat{S}\right) \leq \frac{\phi_{\max}\left(K+s_{0}\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_{0}\right)(\boldsymbol{J})} s_{0} t$$

which is equivalent with

$$E\left(\rho_{\tau}\left(y-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}-\left(\boldsymbol{x}^{T} \hat{\boldsymbol{\beta}}-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)\right)-\rho_{\tau}\left(\boldsymbol{y}-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)\right) \leq \frac{\phi_{\max}\left(K+s_{0}\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_{0}\right)(\boldsymbol{J})} s_{0} t$$

Using Knight's identity, we obtain

$$E\left[-\boldsymbol{x}^{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right) \psi_{\tau}\left(y-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)+\int_{0}^{\boldsymbol{x}^{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right)} 1\left(\varepsilon_{0} \leq s\right)-1\left(\varepsilon_{0} \leq 0\right) d s\right] \leq \frac{\phi_{\max}\left(K+s_{0}\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_{0}\right)(\boldsymbol{J})} s_{0} t$$

where $\varepsilon=y-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}$. For some $\xi \in\left(0, \boldsymbol{x}^{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right)\right)$, it follows that

$$\begin{aligned}
E\left[\int_{0}^{\boldsymbol{x}^{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right)} 1\left(\varepsilon_{0} \leq s\right)-1\left(\varepsilon \leq 0\right) d s\right] & \leq\left|E \boldsymbol{x}^{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right) \psi_{\tau}\left(y-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)\right|+\frac{\phi_{\max}\left(K+s_{0}\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_{0}\right)(\boldsymbol{J})} s_{0} t \\
E\left[\int_{0}^{\boldsymbol{x}^{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right)} F_{\varepsilon_{0}}(s)-F_{\varepsilon_{0}}(0) d s\right] & \leq\left|E \boldsymbol{x}^{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right) \psi_{\tau}\left(y-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)\right|+\frac{\phi_{\max}\left(K+s_{0}\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_{0}\right)(\boldsymbol{J})} s_{0} t \\
E\left[\int_{0}^{\boldsymbol{x}^{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right)} s f_{\varepsilon_{0}}\left(\xi_{\varepsilon}\right) d s\right] & \leq\left|E \boldsymbol{x}^{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right) \psi_{\tau}\left(y-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)\right|+\frac{\phi_{\max}\left(K+s_{0}\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_{0}\right)(\boldsymbol{J})} s_{0} t \\
E\left(\left(\boldsymbol{x}^{T} \hat{\boldsymbol{\beta}}-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)^{2}\right) & \leq \frac{2}{f_{\varepsilon_{0}}\left(\xi_{\varepsilon}\right)}\left\|E \boldsymbol{x} \psi_{\tau}\left(y-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)\right\|_{\infty}\left|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right| \\
& \quad+\frac{2}{f_{\varepsilon_{0}}\left(\xi_{\varepsilon}\right)} \frac{\phi_{\max}\left(K+s_{0}\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_{0}\right)(\boldsymbol{J})} s_{0} t \\
E\left(\left(\boldsymbol{x}^{T} \hat{\boldsymbol{\beta}}-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)^{2}\right) & \leq \frac{2}{\underline{f}_{\varepsilon_{0}}}\left[\left\|E \boldsymbol{x} \psi_{\tau}\left(y-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)\right\|_{\infty}\left|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right|+\frac{\phi_{\max}\left(K+s_{0}\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_{0}\right)(\boldsymbol{J})} s_{0} t\right]
\end{aligned}$$

Since we know $\left|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right| \leq \sqrt{K+s_{0}}\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{0}\right\|_{2} \leq \sqrt{K+s_{0}} \frac{E\left(\left(\boldsymbol{x}^{T} \hat{\boldsymbol{\beta}}-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)^{2}\right)^{\frac{1}{2}}}{\varphi_{\min}\left(K+s_{0}\right)\left(E\left(\boldsymbol{x} \boldsymbol{x}^{T}\right)\right)}$ where $\varphi_{\min}(s)(\boldsymbol{G})=\min_{S \subseteq[p]:|S| \leq s} \lambda_{\min}\left(\boldsymbol{G}_{S}\right)$ and $\boldsymbol{G}_{S}$ is the principal sub-matrix of $\boldsymbol{G}$, we have

$$\begin{aligned}
E\left(\left(\boldsymbol{x}^{T} \hat{\boldsymbol{\beta}}-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)^{2}\right)^{\frac{1}{2}} \leq & \frac{2}{\underline{f}_{\varepsilon_{0}}}\left\|E \boldsymbol{x} \psi_{\tau}\left(y-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)\right\|_{\infty} \sqrt{K+s_{0}} \frac{1}{\varphi_{\min}\left(K+s_{0}\right)\left(E\left(\boldsymbol{x} \boldsymbol{x}^{T}\right)\right)} \\
& +\frac{2}{\underline{f}_{\varepsilon_{0}}} \frac{\phi_{\max}\left(K+s_{0}\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_{0}\right)(\boldsymbol{J})} s_{0} t \frac{1}{E\left(\left(\boldsymbol{x}^{T} \hat{\boldsymbol{\beta}}-\boldsymbol{x}^{T} \boldsymbol{\beta}_{0}\right)^{2}\right)^{\frac{1}{2}}}
\end{aligned}$$

If $E\left(\left(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}-\boldsymbol{x}^T\boldsymbol{\beta}_0\right)^2\right)^{\frac{1}{2}} > \sqrt{\frac{2}{\underline{\mathrm{f}}_{\varepsilon_0}}\frac{\phi_{\max(K+s_0)}(\boldsymbol{J})}{\phi_{\min(K+s_0)}(\boldsymbol{J})}s_0 t}$, then we have

$$E\left(\left(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}-\boldsymbol{x}^T\boldsymbol{\beta}_0\right)^2\right)^{\frac{1}{2}} \leq \frac{2}{\underline{\mathrm{f}}_{\varepsilon_0}}\left\|E\boldsymbol{x}\psi_\tau\left(y-\boldsymbol{x}^T\boldsymbol{\beta}_0\right)\right\|_\infty\sqrt{K+s_0}\frac{1}{\varphi_{\min}\left(K+s_0\right)\left(E\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right)} + \sqrt{\frac{2}{\underline{\mathrm{f}}_{\varepsilon_0}}\frac{\phi_{\max}\left(K+s_0\right)(\boldsymbol{J})}{\phi_{\min}\left(K+s_0\right)(\boldsymbol{J})}s_0 t}$$

Otherwise, the conclusion holds automatically.

The second argument follows from the definition of ρ_τ and Jensen inequality,

$$\rho_\tau\left(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}-\boldsymbol{x}^T\boldsymbol{\beta}_0\right) \leq E\left|\boldsymbol{x}^T\hat{\boldsymbol{\beta}}-\boldsymbol{x}^T\boldsymbol{\beta}_0\right| \leq E\left(\left(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}-\boldsymbol{x}^T\boldsymbol{\beta}_0\right)^2\right)^{\frac{1}{2}}$$

□

A.6 Proof of theorem 4.1

Proof. Consider $\delta_n = L\sqrt{\frac{K\log p}{n}}$ for some sufficient large constant L, we define $Q_{\hat{S}}\left(\boldsymbol{\beta}_{\hat{S}}\right) = E\left[\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}\right)\right]$. Moreover, we let

$$D\left(\delta_n\right) = \inf_{|\hat{S}|=K}\inf_{\left\|\boldsymbol{\beta}_{\hat{S}}-\boldsymbol{\beta}_{\hat{S}}^0\right\|>\delta_n} Q_{\hat{S}}\left(\boldsymbol{\beta}_{\hat{S}}\right)-Q_{\hat{S}}\left(\boldsymbol{\beta}_{\hat{S}}^0\right)$$

and define

$$\mathcal{S}_{\hat{S}}\left(\delta_n\right) = \left\{\boldsymbol{\beta}_{\hat{S}} : \left\|\boldsymbol{\beta}_{\hat{S}}-\boldsymbol{\beta}_{\hat{S}}^0\right\| > \delta_n, \left\|\boldsymbol{\beta}_{\hat{S}}-\boldsymbol{\beta}_{\hat{S}}^0\right\| = o\left(1\right)\right\}$$

Therefore, for any $\beta_{\hat{S}} \in \mathcal{S}_{\hat{S}}\left(\delta_n\right)$, we have

$$\begin{aligned}
Q_{\hat{S}}\left(\boldsymbol{\beta}_{\hat{S}}\right)-Q_{\hat{S}}\left(\boldsymbol{\beta}_{\hat{S}}^0\right) &= E\left[\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}\right)-\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)\right] \\
&= E\left[\rho_\tau\left(\varepsilon_i+u_{i,\hat{S}}-\boldsymbol{x}_{i,\hat{S}}^T\left(\boldsymbol{\beta}_{\hat{S}}-\boldsymbol{\beta}_{\hat{S}}^0\right)\right)-\rho_\tau\left(\varepsilon_i+u_{i,\hat{S}}\right)\right] \\
&= E\left[\int_0^{\boldsymbol{x}_{i,\hat{S}}^T\left(\boldsymbol{\beta}_{\hat{S}}-\boldsymbol{\beta}_{\hat{S}}^0\right)} F_{\varepsilon_i}\left(-u_{i,\hat{S}}+s|\boldsymbol{x}_i\right)-F_{\varepsilon_i}\left(-u_{i,\hat{S}}|\boldsymbol{x}_i\right)ds\right]+o\left(1\right) \\
&\approx \frac{1}{2}\left(\boldsymbol{\beta}_{\hat{S}}-\boldsymbol{\beta}_{\hat{S}}^0\right)^T A_{\hat{S}}\left(\boldsymbol{\beta}_{\hat{S}}-\boldsymbol{\beta}_{\hat{S}}^0\right) \\
&\geq \frac{1}{2}\underline{\mathrm{c}}_A\delta_n^2
\end{aligned}$$

where we define $u_{i,\hat{S}} = y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0-\varepsilon_i$ and use Knight's identity with $-\boldsymbol{x}_{i,\hat{S}}^T\left(\boldsymbol{\beta}_{\hat{S}}-\boldsymbol{\beta}_{\hat{S}}^0\right)\psi_\tau\left(\varepsilon_i+u_{i,\hat{S}}\right) = o_p\left(1\right)$.

Using Boole's inequality, we have

$$\begin{aligned}
&P\left(\max_{1\leq k\leq K}\max_{|\hat{S}|\leq k}\left\|\hat{\boldsymbol{\beta}}_{\hat{S}}-\boldsymbol{\beta}^0_{\hat{S}}\right\|>\delta_n\right)\\
\leq& Kp\left(p-1\right)\cdots\left(p-K\right)\max_{1\leq k\leq K}\max_{|\hat{S}|\leq k}P\left(\left\|\hat{\boldsymbol{\beta}}_{\hat{S}}-\boldsymbol{\beta}^0_{\hat{S}}\right\|>\delta_n\right)\\
\leq& Kp^K\max_{1\leq k\leq K}\max_{|\hat{S}|\leq k}P\left(Q_{\hat{S}}\left(\boldsymbol{\beta}_{\hat{S}}\right)-Q_{\hat{S}}\left(\boldsymbol{\beta}^0_{\hat{S}}\right)\geq D\left(\delta_n\right)\right)\\
\approx& Kp^K\max_{1\leq k\leq K}\max_{|\hat{S}|\leq k}P\left(W_{i,\hat{S}}\geq 2nD\left(\delta_n\right)\right)
\end{aligned}$$

where $\hat{W}_{i,\hat{S}}=n\left(\boldsymbol{\beta}_{\hat{S}}-\boldsymbol{\beta}^0_{\hat{S}}\right)^T A_{\hat{S}}\left(\boldsymbol{\beta}_{\hat{S}}-\boldsymbol{\beta}^0_{\hat{S}}\right)$.

Applying the same argument in the proof of theorem 3.2 of Lu and Su (2015) and theorem 2 of Lee and Shin (2021) with lemma 2.1 of Shibata (1981, 1982), we obtain

$$\begin{aligned}
&Kp^K\max_{1\leq k\leq K}\max_{|\hat{S}|\leq k}P\left(W_{i,\hat{S}}\geq 2nD\left(\delta_n\right)\right)\\
\leq&\lim_{n\to\infty}\sup Kp^kP\left(\chi^2\left(l_{\hat{S}}\right)\geq\frac{2nD\left(\delta_n\right)}{\bar{c}_A\bar{c}_B/\underline{c}_A^2}\right)\\
\leq&\lim_{n\to\infty}\sup Kp^KP\left(\chi^2\left(\bar{l}\right)\geq\frac{2nD\left(\delta_n\right)}{\bar{c}_A\bar{c}_B/\underline{c}_A^2}\right)\\
\leq&\lim_{n\to\infty}\sup Kp^KP\left(\chi^2\left(\bar{l}\right)\geq\bar{l}+\frac{2nD\left(\delta_n\right)}{\bar{c}_A\bar{c}_B/\underline{c}_A^2}-\bar{l}\right)\\
\leq&\lim_{n\to\infty}\sup Kp^Ke^{-\frac{n\delta_n^2\frac{\underline{c}_A^3}{\bar{c}_A\bar{c}_B}-\bar{l}}{2}\left(1-\frac{\log n\delta_n^2\frac{\underline{c}_A^3}{\bar{l}\bar{c}_A\bar{c}_B}}{n\delta_n^2\frac{\underline{c}_A^3}{\bar{l}\bar{c}_A\bar{c}_B}-1}\right)}
\end{aligned}$$

By assumption $\frac{1}{K}+\frac{K}{\log p}+\frac{\log p}{n}\to 0$, we note that $\frac{\log n\delta_n^2\frac{\underline{c}_A^3}{\bar{l}\bar{c}_A\bar{c}_B}}{n\delta_n^2\frac{\underline{c}_A^3}{\bar{l}\bar{c}_A\bar{c}_B}-1}=o\left(1\right)$. Moreover, we have

$$\begin{aligned}
Kp^Ke^{-\frac{n\delta_n^2\frac{\underline{c}_A^3}{\bar{c}_A\bar{c}_B}}{2}}&=Kp^Kp^{-\frac{1}{2}L^2K\frac{\underline{c}_A^3}{\bar{c}_A\bar{c}_B}}\\
&=o\left(1\right)
\end{aligned}$$

for sufficiently large L. Therefore, we obtain $P\left(\max_{1\leq k\leq K}\max_{|\hat{S}|\leq k}\left\|\hat{\boldsymbol{\beta}}_{\hat{S}}-\boldsymbol{\beta}^0_{\hat{S}}\right\|>\delta_n\right)=o\left(1\right)$, which the statement follows. □

A.7 Proof of corollary 4.1

Proof. By triangular inequality, we note that

$$\begin{aligned}
&\max_{1\le k\le K}\max_{|\hat{S}|\le k}\left|\hat{\rho}_{\tau,\hat{S}}-\rho_{\tau,\hat{S}}\right| \\
=&\max_{1\le k\le K}\max_{|\hat{S}|\le k}\left|\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)-E\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)\right| \\
=&\max_{1\le k\le K}\max_{|\hat{S}|\le k}\left|\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)-\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)\right| \\
&+\max_{1\le k\le K}\max_{|\hat{S}|\le k}\left|\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)-E\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)\right| \\
:=&V_1+V_2
\end{aligned}$$

For V_1,

$$\begin{aligned}
&\max_{1\le k\le K}\max_{|\hat{S}|\le k}\left|\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)-\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)\right| \\
\le&\max_{1\le k\le K}\max_{|\hat{S}|\le k}\frac{1}{n}\sum_{i=1}^{n}\left|\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)\left(\tau-1\left(y_i<\boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)\right)-\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)1\left(y_i<\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)\right| \\
\le&\frac{1}{n}\sum_{i=1}^{n}\max_{1\le k\le K}\max_{|\hat{S}|\le k}\left|\boldsymbol{x}_{i,\hat{S}}^T\left(\boldsymbol{\beta}_{\hat{S}}^0-\hat{\boldsymbol{\beta}}_{\hat{S}}\right)\right| \\
\le&\frac{1}{n}\sum_{i=1}^{n}\max_{1\le k\le K}\max_{|\hat{S}|\le k}\left\|\boldsymbol{x}_{i,\hat{S}}\right\|\left\|\boldsymbol{\beta}_{\hat{S}}^0-\hat{\boldsymbol{\beta}}_{\hat{S}}\right\| \\
\le&C\sqrt{K}\max_{1\le k\le K}\max_{|\hat{S}|\le k}\left\|\boldsymbol{\beta}_{\hat{S}}^0-\hat{\boldsymbol{\beta}}_{\hat{S}}\right\| \\
=&O_p\left(\sqrt{\frac{K^2\log p}{n}}\right)
\end{aligned}$$

For V_2, by Hoeffding's inequality with respect to $\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)$, we have

$$\begin{aligned}
P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)-E\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}^0\right)\right|\ge L\sqrt{\frac{K^2\log p}{n}}\right)&\le e^{-\frac{2L^2nK^2\log p}{\sum_{i=1}^{n}C_i}} \\
&\le e^{-\frac{L^2K^2\log p}{C}} \\
&=o\left(1\right)
\end{aligned}$$

Therefore, we have $V_2=O_p\left(\sqrt{\frac{K^2\log p}{n}}\right)$. So the lemma follows. □

A.8 Proof of lemma 4.1

Proof. We first consider the weak submodularity of the negative quantile loss function $-\rho_\tau$. Following the same procedure in lemma 3.1 and theorem 3.1, we can derive

$$\gamma_{U,k}\left(-\rho_\tau\right) \geq \frac{\phi_{\min}\left(|U|+k\right)\left(\boldsymbol{J}\right)}{\phi_{\max}\left(|U|+k\right)\left(\boldsymbol{J}\right)}$$

for generic set U and number k.

By the definition of weak submodularity ratio $\gamma_{U,k}\left(-\rho_\tau\right) = \min_{L\subseteq U, S:|S|\leq k, S\cap L=\emptyset} \frac{\sum_{x\in S}\rho_\tau(L)-\rho_\tau(L\cup\{x\})}{\rho_\tau(L)-\rho_\tau(L\cup\{S\})}$ and the result above, we note that

$$\frac{\sum_{x\in U\backslash V}\rho_\tau\left(V\right)-\rho_\tau\left(V\cup\{x\}\right)}{\rho_\tau\left(V\right)-\rho_\tau\left(V\cup\{U\backslash V\}\right)} \geq \gamma_{V,u-v} \geq \frac{\phi_{\min}\left(u\right)\left(\boldsymbol{J}\right)}{\phi_{\max}\left(u\right)\left(\boldsymbol{J}\right)}$$

Therefore, we obtain

$$\left(u-v\right)\frac{\max_{\{x\}\in U}\rho_\tau\left(V\right)-\rho_\tau\left(V\cup\{x\}\right)}{\rho_\tau\left(V\right)-\rho_\tau\left(U\right)} \geq \frac{\phi_{\min}\left(u\right)\left(\boldsymbol{J}\right)}{\phi_{\max}\left(u\right)\left(\boldsymbol{J}\right)}$$

Rearranging the inequality,

$$\max_{\{x\}\in U}\rho_\tau\left(V\right)-\rho_\tau\left(V\cup\{x\}\right) \geq \frac{\phi_{\min}\left(u\right)\left(\boldsymbol{J}\right)}{\phi_{\max}\left(u\right)\left(\boldsymbol{J}\right)}\frac{1}{u-v}\left(\rho_\tau\left(V\right)-\rho_\tau\left(U\right)\right)$$

□

A.9 Proof of lemma 4.2

Proof. Let $W, V \subset [p]$ and $W \neq V$. Define $U = W \cup V$ and $u - v \geq 1$. Since $u - v = |W \cup V| - v \leq w$, we have $\phi_{\min}\left(u\right)\left(\boldsymbol{J}\right) \geq \phi_{\min}\left(w+v\right)\left(\boldsymbol{J}\right)$ and $\phi_{\max}\left(u\right)\left(\boldsymbol{J}\right) \leq \phi_{\max}\left(w+v\right)\left(\boldsymbol{J}\right)$. Therefore, $\frac{\phi_{\min}(u)(\boldsymbol{J})}{\phi_{\max}(u)(\boldsymbol{J})} \geq \frac{\phi_{\min}(w+v)(\boldsymbol{J})}{\phi_{\max}(w+v)(\boldsymbol{J})}$ and we have

$$\begin{aligned}
&\frac{\phi_{\min}\left(u\right)\left(\boldsymbol{J}\right)}{\phi_{\max}\left(u\right)\left(\boldsymbol{J}\right)}\frac{1}{u-v}\left(\rho_\tau\left(V\right)-\rho_\tau\left(U\right)\right)\\
\geq&\frac{\phi_{\min}\left(w+v\right)\left(\boldsymbol{J}\right)}{\phi_{\max}\left(w+v\right)\left(\boldsymbol{J}\right)}\frac{1}{w}\left(\rho_\tau\left(V\right)-\rho_\tau\left(U\right)\right)\\
\geq&\frac{\phi_{\min}\left(w+v\right)\left(\boldsymbol{J}\right)}{\phi_{\max}\left(w+v\right)\left(\boldsymbol{J}\right)}\frac{1}{w}\left(\rho_\tau\left(V\right)-\rho_\tau\left(W\right)\right)
\end{aligned}$$

Multiplying both sides by $-(1-\kappa)$, adding $\rho_\tau(V)-\rho_\tau(W)$ and using lemma 5, we obtain

$$\begin{aligned}
&\left(1-(1-\kappa)\frac{\phi_{\min}(w+v)(\boldsymbol{J})}{\phi_{\max}(w+v)(\boldsymbol{J})}\frac{1}{w}\right)(\rho_\tau(V)-\rho_\tau(W))\\
\geq&\rho_\tau(V)-\rho_\tau(W)-(1-\kappa)\frac{\phi_{\min}(u)(\boldsymbol{J})}{\phi_{\max}(u)(\boldsymbol{J})}\frac{1}{u-v}(\rho_\tau(V)-\rho_\tau(U))\\
\geq&\rho_\tau(V)-\rho_\tau(W)-(1-\kappa)\max_{\{j\}\in U}\rho_\tau(V)-\rho_\tau(V\cup\{j\})\\
\geq&\rho_\tau(V)-\rho_\tau(W)-(1-\kappa)\max_{\{j\}\in[p]}\rho_\tau(V)-\rho_\tau(V\cup\{j\})
\end{aligned}$$

Suppose $\rho_\tau(U_K)\geq\rho_\tau(W)$. Otherwise, the statement holds automatically. We have

$$\begin{aligned}
\rho_\tau(U_K)-\rho_\tau(W)&=\rho_\tau(U_{K-1})-\rho_\tau(W)-(\rho_\tau(U_{K-1})-\rho_\tau(U_K))\\
&\leq\rho_\tau(U_{K-1})-\rho_\tau(W)-(1-\kappa)\max_{j\in[p]}\rho_\tau(U_{k-1})-\rho_\tau(U_k\cup\{j\})\\
&\leq\left(1-(1-\kappa)\frac{\phi_{\min}(w+K-1)(\boldsymbol{J})}{\phi_{\max}(w+K-1)(\boldsymbol{J})}\frac{1}{w}\right)(\rho_\tau(U_{K-1})-\rho_\tau(W))
\end{aligned}$$

Repeatedly, we can obtain

$$\begin{aligned}
&\rho_\tau(U_K)-\rho_\tau(W)\\
\leq&\left(1-(1-\kappa)\frac{\phi_{\min}(w+K-1)(\boldsymbol{J})}{\phi_{\max}(w+K-1)(\boldsymbol{J})}\frac{1}{w}\right)(\rho_\tau(U_{K-1})-\rho_\tau(W))\\
\leq&\left(1-(1-\kappa)\frac{\phi_{\min}(w+K-1)(\boldsymbol{J})}{\phi_{\max}(w+K-1)(\boldsymbol{J})}\frac{1}{w}\right)\left(1-(1-\kappa)\frac{\phi_{\min}(w+K-2)(\boldsymbol{J})}{\phi_{\max}(w+K-2)(\boldsymbol{J})}\frac{1}{w}\right)(\rho_\tau(U_{K-2})-\rho_\tau(W))\\
&\cdots\\
\leq&\Pi_{k=1}^{K-1}\left(1-(1-\kappa)\frac{\phi_{\min}(w+k)(\boldsymbol{J})}{\phi_{\max}(w+k)(\boldsymbol{J})}\frac{1}{w}\right)(\rho_\tau(U_1)-\rho_\tau(W))\\
\leq&\left(1-(1-\kappa)\frac{\phi_{\min}(w+K)(\boldsymbol{J})}{\phi_{\max}(w+K)(\boldsymbol{J})}\frac{1}{w}\right)^K(\rho_\tau(\emptyset)-\rho_\tau(W))\\
\leq&\rho_\tau(\emptyset)\left(1-(1-\kappa)\frac{\phi_{\min}(w+K)(\boldsymbol{J})}{\phi_{\max}(w+K)(\boldsymbol{J})}\frac{1}{w}\right)^K
\end{aligned}$$

where we use the assumption that $\phi_{\min}(x)(\boldsymbol{J})$ is monotonically decreasing in x and $\phi_{\max}(x)(\boldsymbol{J})$ is monotonically increasing in x. □

A.10 Proof of theorem 4.2

Proof. Note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)-E\rho_\tau\left(y_i-\boldsymbol{x}_{i,S_0}^T\boldsymbol{\beta}_{S_0}\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)-E\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}\right)+E\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}\right)-E\rho_\tau\left(y_i-\boldsymbol{x}_{i,S_0}^T\boldsymbol{\beta}_{S_0}\right)\\
:=&T_1+T_2
\end{aligned}$$

For T_1, using corollary 4.1, we know $\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\hat{\boldsymbol{\beta}}_{\hat{S}}\right)-E\rho_\tau\left(y_i-\boldsymbol{x}_{i,\hat{S}}^T\boldsymbol{\beta}_{\hat{S}}\right)=O_p\left(\sqrt{\frac{K^2\log p}{n}}\right)=o_p\left(1\right)$.

For T_2, we denote $\zeta_k=\max_{\mathcal{U}_k}\left|\hat{\rho}_\tau\left(U_k\right)-\rho_\tau\left(U_k\right)\right|$ where $\mathcal{U}_k=\left\{\mathcal{U}\subset[p],|\mathcal{U}|\leq r\right\}$ and define a collection of sets

$$A_k\left(\kappa\right)=\left\{V\subset[p]:|V|=k,\max_{j\in[p]}\rho_\tau\left(V\right)-\rho_\tau\left(V\cup\{j\}\right)>\frac{4\zeta_k}{\kappa}\right\}$$

We also denote $\tilde{j}=\arg\max_{j\in[p]}\hat{\rho}_\tau\left(V\right)-\hat{\rho}_\tau\left(V\cup\{j\}\right)$. Suppose $\hat{U}_k\in A_k\left(\kappa\right)$ for all $2\leq k\leq K$, then we have

$$\begin{aligned}
&\rho_\tau\left(\hat{U}_{k-1}\right)-\rho_\tau\left(\hat{U}_{k-1}\cup\left\{\tilde{j}\right\}\right)\\
\geq&\hat{\rho}_\tau\left(\hat{U}_{k-1}\right)-\hat{\rho}_\tau\left(\hat{U}_{k-1}\cup\left\{\tilde{j}\right\}\right)-\left|\hat{\rho}_\tau\left(\hat{U}_{k-1}\right)-\hat{\rho}_\tau\left(\hat{U}_{k-1}\cup\left\{\tilde{j}\right\}\right)-\left(\rho_\tau\left(\hat{U}_{k-1}\right)-\rho_\tau\left(\hat{U}_{k-1}\cup\left\{\tilde{j}\right\}\right)\right)\right|\\
\geq&\hat{\rho}_\tau\left(\hat{U}_{k-1}\right)-\hat{\rho}_\tau\left(\hat{U}_{k-1}\cup\left\{\tilde{j}\right\}\right)-2\max_{U_k}\left|\hat{\rho}_\tau\left(U_k\right)-\rho_\tau\left(U_k\right)\right|\\
=&\max_{j\in[p]}\hat{\rho}_\tau\left(\hat{U}_{k-1}\right)-\hat{\rho}_\tau\left(\hat{U}_{k-1}\cup\{j\}\right)-2\zeta_k\\
\geq&\max_{j\in[p]}\left\{\rho_\tau\left(\hat{U}_{k-1}\right)-\rho_\tau\left(\hat{U}_{k-1}\cup\{j\}\right)-\left|\hat{\rho}_\tau\left(\hat{U}_{k-1}\cup\{j\}\right)-\hat{\rho}_\tau\left(\hat{U}_{k-1}\right)-\left(\rho_\tau\left(\hat{U}_{k-1}\cup\{j\}\right)-\rho_\tau\left(\hat{U}_{k-1}\right)\right)\right|\right\}\\
&-2\zeta_k\\
\geq&\max_{j\in[p]}\rho_\tau\left(\hat{U}_{k-1}\right)-\rho_\tau\left(\hat{U}_{k-1}\cup\{j\}\right)-4\zeta_k\\
>&\left(1-\kappa\right)\max_{j\in[p]}\rho_\tau\left(\hat{U}_{k-1}\right)-\rho_\tau\left(\hat{U}_{k-1}\cup\{j\}\right)
\end{aligned}$$

where the last inequality follows by the definition of $A_k\left(\kappa\right)$. Therefore, we know $\left\{\hat{U}_1,\ldots,\hat{U}_K\right\}\in U_K\left(\kappa\right)$. Using lemma 6, we have

$$\rho_\tau\left(\hat{U}_K\right)-\rho_\tau\left(U_0\right)\leq\rho_\tau\left(\emptyset\right)\left(1-\left(1-\kappa\right)\frac{\phi_{\min}\left(s_0+K\right)\left(\boldsymbol{J}\right)}{\phi_{\max}\left(s_0+K\right)\left(\boldsymbol{J}\right)}\frac{1}{w}\right)^K\to0$$

when the event $\left\{\left\{\hat{U}_1,\ldots,\hat{U}_K\right\}\in U_K\left(\kappa\right)\right\}$ occurs with $\frac{s_0}{K}\to0$ as $K\to\infty$ and $p\to\infty$.

Suppose there exist $\hat{U}_k \notin A_k(\kappa)$ for some $2 \leq k \leq K$. We denote $\tilde{k} = \min_{1 \leq k \leq K} \left\{ \hat{U}_k \notin A_k(\kappa) \right\}$. By the definition of $A_k(\kappa)$, we have $\max_{j \in [p]} \rho_\tau \left(\hat{U}_{\tilde{k}} \right) - \rho_\tau \left(\hat{U}_{\tilde{k}} \cup \{j\} \right) \leq \frac{4\zeta_k}{\kappa}$.

Suppose $U^* \subset \hat{U}_{\tilde{k}}$, by monotonicity of ρ_τ, we have $\rho_\tau \left(\hat{U}_K \right) \leq \rho_\tau \left(\hat{U}_{\tilde{k}} \right) \leq \rho_\tau (U^*)$. Otherwise, using lemma 4, 5 and the assumptions on $\phi_{\min}(x)(\boldsymbol{J})$ and $\phi_{\max}(x)(\boldsymbol{J})$, we can obtain

$$
\begin{aligned}
\rho_\tau \left(\hat{U}_K \right) - \rho_\tau (U^*) &\leq \rho_\tau \left(\hat{U}_{\tilde{k}} \right) - \rho_\tau (U^*) \\
&\leq \rho_\tau \left(\hat{U}_{\tilde{k}} \right) - \rho_\tau \left(\hat{U}_{\tilde{k}} \cup U^* \right) \\
&\leq s_0 \frac{\phi_{\max} \left(s_0 + \tilde{k} \right)(\boldsymbol{J})}{\phi_{\min} \left(s_0 + \tilde{k} \right)(\boldsymbol{J})} \max_{j \in [p]} \rho_\tau \left(\hat{U}_{\tilde{k}} \right) - \rho_\tau \left(\hat{U}_{\tilde{k}} \cup \{j\} \right) \\
&\leq s_0 \frac{\phi_{\max}(s_0 + K)(\boldsymbol{J})}{\phi_{\min}(s_0 + K)(\boldsymbol{J})} \max_{j \in [p]} \rho_\tau \left(\hat{U}_{\tilde{k}} \right) - \rho_\tau \left(\hat{U}_{\tilde{k}} \cup \{j\} \right) \\
&\leq s_0 \frac{\phi_{\max}(s_0 + K)(\boldsymbol{J})}{\phi_{\min}(s_0 + K)(\boldsymbol{J})} \frac{4\zeta_{\tilde{k}}}{\kappa} \\
&= \frac{\phi_{\max}(s_0 + K)(\boldsymbol{J})}{\phi_{\min}(s_0 + K)(\boldsymbol{J})} \frac{4K\zeta_{\tilde{k}}}{\kappa} \frac{s_0}{K} \\
&= O_p \left(\sqrt{\frac{s_0^2 K^2 \log p}{n}} \right) \\
&= o_p(1)
\end{aligned}
$$

where last $o_p(1)$ is from the condition in theorem 4.2.

Combining the results above, we have $\frac{1}{n} \sum_{i=1}^n \rho_\tau \left(y_i - \boldsymbol{x}_{i,\hat{S}}^T \hat{\boldsymbol{\beta}}_{\hat{S}} \right) - E\rho_\tau \left(y_i - \boldsymbol{x}_{i,S_0}^T \boldsymbol{\beta}_{S_0} \right) = o_p(1)$. So theorem 4 holds. □

B Testing-based quantile forward selection

Based on the previous simple quantile forward selection algorithm, we incorporate the hypothesis testing procedure in the selection of covariate. In every step, with previous selected set S, the null hypothesis is defined as follows

$$
H_0 : f(S \cup \{j\}) - f(S) = 0
$$

Given the significance level α, the associated test $T_{jS\alpha} \in \{0, 1\}$ satisfies that if $T_{jS\alpha} = 1$, then H_0 is rejected and if $T_{jS\alpha} = 0$, we fail to reject H_0. The value of $T_{jS\alpha}$ is determined by test statistics W_{jS} for any specific test. In every step, if there are more than one $T_{jS\alpha} = 1$, the algorithm chooses the variable with the largest W_{jS}.

The following is the testing-based quantile forward regression algorithm.

Algorithm 3 Testing-based quantile forward regression

1. Set $S_0 = \emptyset$.
2. For $k = 1, 2, \ldots$
   (a) If $T_{jS_{k-1}\alpha} = 1$ for some $j \in [p] \setminus S_{k-1}$
      i. Select the variable index
      $$s = \arg \max_{j \in [p] \setminus S_{k-1}} \left\{ W_{jS_{k-1}} | T_{jS_{k-1}\alpha} = 1 \right\}$$
      ii. Set $S_k = S_{k-1} \cup \{s\}$.
   (b) Else, break the algorithm
3. Return $\hat{S} = S_K$ with step K, satisfying $T_{j\hat{S}\alpha} = 0$ for all $j \in [p] \setminus \hat{S}$, and the corresponding estimator $\boldsymbol{\beta}_{\hat{S}}$.

Since we are using traditional hypothesis testing procedure to choose variables we need in every step, some regular assumptions on the tests needs to be satisfied. These regular assumptions are similar with the condition 2 in Kozbur (2020).

Assumption 7. *The test T satisfies the following conditions. There exists an integer $K_{test} > s_0$ and some constants $\alpha, \delta_{test}, c_{test}, c'_{test}, c''_{test} > 0$ such that*

1. For all j, the test T has power
$$P\left(T_{jS\alpha} = 1, |S| \leq K_{test} \text{ such that } f\left(S \cup \{j\}\right) - f\left(S\right) > c_{test}\right) \geq 1 - \frac{\delta_{test}}{3}$$
   (a) For some j, the test T can control the size
   $$P\left(T_{jS\alpha} = 1, |S| \leq K_{test} \text{ such that } f\left(S \cup \{j\}\right) - f\left(S\right) \leq c'_{test}\right) \leq \alpha + \frac{\delta_{test}}{3}$$
   (b) For all j, k, the test T is continuous
   $$\begin{aligned} &P\{f\left(S \cup \{j\}\right) - f\left(S\right) \geq c''_{test}\left(f\left(S \cup \{k\}\right) - f\left(S\right)\right), \\ &\quad |S| \leq K_{test} \text{ such that } T_{jS\alpha} = 1 \text{ and } W_{jS} \geq W_{kS}\} \leq 1 - \frac{\delta_{test}}{3} \end{aligned}$$

Under those assumptions above, we shall show the following theorem to control the l_2-bound and quantile loss bound for prediction error.

Theorem B.1. *Under the assumptions and algorithm 3 above, we have the following l_2 prediction bound*

$$E\left(\left(X^T\hat{\beta} - X^T\beta_0\right)^2\right)^{\frac{1}{2}} \leq \sqrt{K + s_0}\left[\frac{\frac{2}{f_{\varepsilon_0}}\left\|EX\psi_\tau\left(Y - X^T\beta_0\right)\right\|_\infty}{\varphi_{\min}\left(K + s_0\right)\left(E\left(XX^T\right)\right)} + \sqrt{\frac{2}{f_{\varepsilon_0}}\frac{\phi_{\max}\left(K + s_0\right)\left(\boldsymbol{J}\right)}{\phi_{\min}\left(K + s_0\right)\left(\boldsymbol{J}\right)}c_{test}}\right]$$

In addition, the quantile prediction loss has the same upper bound

$$\rho_\tau\left(X^T\hat{\beta}-X^T\beta_0\right)\leq\sqrt{K+s_0}\left[\frac{\frac{2}{\underline{f}_{\varepsilon_0}}\left\|EX\psi_\tau\left(Y-X^T\beta_0\right)\right\|_\infty}{\varphi_{\min}\left(K+s_0\right)\left(E\left(XX^T\right)\right)}+\sqrt{\frac{2}{\underline{f}_{\varepsilon_0}}\frac{\phi_{\max}\left(K+s_0\right)\left(\boldsymbol{J}\right)}{\phi_{\min}\left(K+s_0\right)\left(\boldsymbol{J}\right)}c_{test}}\right]$$

hold with probability at least $1-\alpha-\delta_{test}$.

Proof. Suppose the test T satisfying all three conditions in assumption 5. We denote this event as E. So $P\left(E\right)\geq 1-\alpha-\delta_{test}$. On this event, similar to the proof of lemma 2 and theorem 2, we replace t by c_{test}. Then we can obtain

$$f\left(S^*\right)-f\left(\hat{S}\right)\leq\frac{\phi_{\max}\left(K+s_0\right)\left(\boldsymbol{J}\right)}{\phi_{\min}\left(K+s_0\right)\left(\boldsymbol{J}\right)}s_0c_{test}$$

and

$$E\left(\left(X^T\hat{\beta}-X^T\beta_0\right)^2\right)^{\frac{1}{2}}\leq\sqrt{K+s_0}\left[\frac{\frac{2}{\underline{f}_{\varepsilon_0}}\left\|EX\psi_\tau\left(Y-X^T\beta_0\right)\right\|_\infty}{\varphi_{\min}\left(K+s_0\right)\left(E\left(XX^T\right)\right)}+\sqrt{\frac{2}{\underline{f}_{\varepsilon_0}}\frac{\phi_{\max}\left(K+s_0\right)\left(\boldsymbol{J}\right)}{\phi_{\min}\left(K+s_0\right)\left(\boldsymbol{J}\right)}c_{test}}\right]$$

In addition, the quantile prediction loss bound follows naturally from the l_2 bound. □

C Additional results

C.1 Additional application results in the international growth rate example

The following table 8 contains the covariate selection result for all τ varying from 0.05 to 0.95. It shows a clear pattern that black market premium exists in all lower quantile levels while total gross enrollment ratio for primary education appears in all upper quantile levels.

Table 8: Variable selection result for the international growth example with K-step quantile forward regression (Real per capita GDP (log) is always included)

τ	Additional variables selected
0.05	Population proportion over 64 Black market premium (log) Terms of trade shock Percentage of "no schooling" in the male population Infant mortality rate Ratio of nominal government current expenditure on the defense to nominal GDP
0.10	Black market premium (log) Ratio of real domestic investment
0.15	Black market premium (log) Ratio of real domestic investment
0.20	Black market premium (log)
0.25	Black market premium (log)
0.30	Black market premium (log)
0.35	Black market premium (log)
0.40	Black market premium (log)
0.45	Black market premium (log)
0.50	Black market premium (log)
0.55	Black market premium (log) Life expectancy
0.60	Infant mortality rate Ratio of real government consumption expenditure net of spending on defense and on education to real GDP
0.65	Infant mortality rate
0.70	Total gross enrollment ratio for primary education
0.75	Total gross enrollment ratio for primary education
0.80	Total gross enrollment ratio for primary education
0.85	Total gross enrollment ratio for primary education Measure of tariff restriction
0.90	Total gross enrollment ratio for primary education Measure of tariff restriction Exchange rate
0.95	Infant mortality rate Ratio of domestic investment to real GDP Percentage of "no schooling" in the female population Total gross enrollment ratio for primary education Exchange rate