Statistical Papers (2023) 64:1943–1968
https://doi.org/10.1007/s00362-022-01367-z

REGULAR ARTICLE

Optimal subsampling for functional quantile regression

Qian Yan[1] · **Hanyu Li**[1] · **Chengmei Niu**[1]

Received: 6 May 2022 / Accepted: 27 September 2022 / Published online: 19 October 2022

Abstract
Subsampling is an efficient method to deal with massive data. In this paper, we investigate the optimal subsampling for linear quantile regression when the covariates are functions. The asymptotic distribution of the subsampling estimator is first derived. Then, we obtain the optimal subsampling probabilities based on the A-optimality criterion. Furthermore, the modified subsampling probabilities without estimating the densities of the response variables given the covariates are also proposed, which are easier to implement in practise. Numerical experiments on synthetic and real data show that the proposed methods always outperform the one with uniform sampling and can approximate the results based on full data well with less computational efforts.

Keywords Functional quantile regression · A-optimality · Asymptotic distribution · Optimal subsampling · Massive data

Mathematics Subject Classification 62K05 · 62G08 · 62R10

1 Introduction

Technological advances have made data easier to collect, store, and process, allowing multiple points in the temporal or spatial domain to be observed and recorded. These observations can be viewed as smooth functions with respect to time or space, which is called functional data in statistics. Functional data analysis (FDA) is particularly important given the widespread availability of functional data. Traditional FDA methods, however, are no longer available due to limited computer resources as a result of

✉ Hanyu Li
lihy.hy@gmail.com ; hyli@cqu.edu.cn

Qian Yan
qianyan@cqu.edu.cn

Chengmei Niu
chengmeiniu@cqu.edu.cn

[1] College of Mathematics and Statistics, Chongqing University, Chongqing 401331, People's Republic of China

these massive data. In order to overcome this problem, random subsampling methods are alternative approaches that have shown good performance in extracting meaningful information from large-scale datasets and making statistical methods scalable to massive data.

To the best of our knowledge, there are two main types of random subsampling methods in statistical models: Randomized Numerical Linear Algebra (RandNLA) subsampling approaches and optimal subsampling approaches. Popular RandNLA subsampling approaches include uniform sampling, leverage score sampling and shrinkage leverage score sampling; see e.g., (Drineas et al. 2006; Mahoney 2011; Drineas et al. 2012). Currently, some researchers have studied statistical properties of these RandNLA subsampling estimators for regression models. For example, Ma et al. (2015) presented the bias and variance of subsampling estimator for least squares regression, and Wang et al. (2018) and Homrighausen and McDonald (2019) extended them to ridge regression. Raskutti and Mahoney (2016) and Dobriban and Liu (2019) investigated error bounds for the statistical efficiency of the estimator based on subsampling least squares regression. Ma et al. (2020) comprehensively analyzed the asymptotic properties of the RandNLA subsampling estimator for linear regression under certain regularity assumptions.

On the other hand, several scholars have developed optimal subsampling methods for parametric regression problems. For example, Wang et al. (2018) proposed an inverse weighted subsampling method for logistic regression based on the A- or L-optimality criterion. Subsequently, a more efficient estimation method and Poisson subsampling were considered by Wang (2019) to correct the bias of the subsampling estimator given in Wang et al. (2018) and to improve the computational efficiency. Later, Yao and Wang (2019), Yu et al. (2020) and Ai et al. (2021) extended the subsampling method to softmax regression, quasi-likelihood and generalized linear models, respectively. Very recently, Wang and Ma (2021), Ai et al. (2021), Fan et al. (2021), and Shao et al. (2022) employed the optimal subsampling method in ordinary quantile regression, and Shao and Wang (2021) and Yuan et al. (2022) developed the subsampling for composite quantile regression.

All of the aforementioned studies of subsampling methods focus on statistical models with scalar variables, and now only little work has been done in the area of subsampling for functional regression. As far as we know, these studies are mainly concerned with functional mean regression, which is an extension of the multiple mean regression model in the functional data setting. Specifically, He and Yan (2022) proposed a functional principal subspace sampling probability for functional linear regression with scalar response, which eliminates the impact of eigenvalue inside the functional principal subspace and properly weights the residuals. Liu et al. (2021) extended the optimal subsampling method to functional linear regression and functional generalized linear model with scalar response. As we know, the mean regression is more sensitive to outliers and less able to handle heavy-tailed errors. Also, the assumption of homoskedasticity of errors in mean regression is usually invalid in massive data. The quantile regression proposed by Koenker and Bassett (1978) can tackle these issues and hence has attracted a lot of attention from scholars as a robust alternative to mean regression. Specifically, the quantile regression gives much more complete information about the conditional response distribution than the traditional

mean regression, and exhibits robustness to outliers and data located in the tail of the conditional response distribution. Additionally, the quantile regression naturally incorporates heteroscedasticity. For functional quantile regression with scalar response, there are also many works; see e.g., (Cardot et al. 2004, 2005; Chen and Müller 2012; Kato 2012; Sang and Cao 2020). More specifically, Cardot et al. (2004, 2005) studied penalized spline estimator and its convergence rate. Chen and Müller (2012) and Kato (2012) obtained the estimation of slope function based on functional principal component analysis basis. Sang and Cao (2020) studied penalized spline estimator for functional single index quantile regression. However, these methods cannot be directly applied to massive data, and, to the best of our knowledge, there is almost no work on random subsampling for functional quantile regression, in contrast to quantile regression with scalar variables, where there is a lot of work as previously mentioned.

Based on the above motivation, we investigate the optimal subsampling for quantile regression in massive data when the covariates are functions. We first derive the asymptotic distribution of the general subsampling estimator and then obtain the optimal subsampling probabilities by minimizing the asymptotic integrated mean squared error (IMSE) under the A-optimality criterion. In addition, we also provide a feasible modified version of the optimal subsampling probabilities to ensure the feasibility of the subsampling method. It is worth pointing out that our work differs substantially from the one by Liu et al. (2021). On the one hand, the loss function of the mean regression considered in Liu et al. (2021) is differentiable, but the functional quantile regression is a non-differentiable problem, which makes the theoretical deduction more challenging. On the other hand, compared with the informative sampling conditional on full data in Liu et al. (2021), our subsampling methods are unconditional and non-informative, which can make the subsampling estimator more stable (Ai et al. 2021).

The rest of this paper is organized as follows. Section 2 briefly introduces the scalar-on-function linear quantile regression problem and presents asymptotic behaviors of the penalized spline estimator. In Section 3, we derive the asymptotic distribution of the subsampling estimator and the optimal subsampling probabilities based on the A-optimality criterion. The modified version of these probabilities is also considered in this section. Section 4 illustrates our methodology through both numerical simulations and real data sets. Section 5 concludes this paper with some discussions. All proofs are delivered to the Appendix.

2 Model and Estimation

2.1 Functional quantile regression

Suppose that $\{x_i(t), y_i\}_{i=1}^n$ are n independent observations of $(\boldsymbol{X}(t), \boldsymbol{Y})$, where the covariates $x_i(t)$ are square integrable functions defined on $[0, 1]$, i.e., the elements of the space $L^2[0, 1]$, and are assumed to be non-random, and y_i are scalar responses. A scalar-on-function linear quantile regression model is defined as follows

$$y_i = \int_0^1 x_i(t)\beta(t)\mathrm{d}t + \epsilon_i \quad with \quad \mathrm{P}(\epsilon_i < 0 \mid x_i(t)) = \tau, \tag{1}$$

where $\beta(t)$ is an unknown slope function satisfying $\beta(t) \in L^2[0, 1]$, ϵ_i are independent random errors with probability density function $f_{\epsilon|\boldsymbol{X}(t)}(\epsilon_i, x_i(t))$, and the quantile level $\tau \in (0, 1)$. Thus, the τ-th conditional quantile of y_i given $x_i(t)$ is

$$Q_\tau(y_i \mid x_i(t)) = \int_0^1 x_i(t)\beta(t)\mathrm{d}t.$$

2.2 Full data estimation of $\boldsymbol{\beta(t)}$

To estimate the slope function $\beta(t)$, we consider the B-spline basis functions defined on equispaced knots. Specifically, let K equispaced interior knots divide the interval $[0, 1]$ into $K + 1$ sub-intervals, i.e., $[t_j, t_{j+1}]$, $j = 0, \ldots, K$. In these intervals, we can find $K + p + 1$ normalized B-spline basis functions $\{B_k(t), 1 \le k \le K + p + 1\}$, as denoted by $\boldsymbol{B}(t) = (B_1(t), B_2(t), \ldots, B_{K+p+1}(t))^T$. They are the piecewise polynomials of degree p on each sub-interval $[t_j, t_{j+1}]$ and $p - 1$ times continuously differentiable on $[0, 1]$. More properties of the B-spline function can be found in de Boor (2001). Thus, we can estimate $\beta(t)$ using a linear combination of the normalized B-spline basis functions (Stone 1985), which allows us to find a vector $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{K+p+1}$ such that

$$\hat{\beta}(t) = \sum_{k=1}^{K+p+1} \hat{\boldsymbol{\theta}}_k B_k(t) = \boldsymbol{B}^T(t)\hat{\boldsymbol{\theta}},$$

where $\hat{\boldsymbol{\theta}}$ is a solution of the minimization problem

$$L(\boldsymbol{\theta}; \lambda, K) = \sum_{i=1}^{n} \rho_\tau\Big(y_i - \int_0^1 x_i(t)\boldsymbol{B}^T(t)\boldsymbol{\theta}\mathrm{d}t\Big) + \frac{\lambda}{2}\int_0^1 \left\{\left(\boldsymbol{B}^{(q)}(t)\right)^T \boldsymbol{\theta}\right\}^2 \mathrm{d}t, \tag{2}$$

where $\rho_\tau(\epsilon) = \epsilon\{\tau - I(\epsilon < 0)\}$ is the quantile loss function with $I(\cdot)$ being the indicator function, $\lambda > 0$ is the smoothing parameter, and $\boldsymbol{B}^{(q)}(t)$ in the penalty term is the integrated squared q-th order derivative of all the B-splines functions for some integer $q \le p$. Furthermore, let $\boldsymbol{B}_i = \int_0^1 x_i(t)\boldsymbol{B}(t)\mathrm{d}t$ and $\boldsymbol{D}_q = \int_0^1 \boldsymbol{B}^{(q)}(t)\{\boldsymbol{B}^{(q)}(t)\}^T\mathrm{d}t$, the loss function (2) thus can be rewritten as

$$L(\boldsymbol{\theta}; \lambda, K) = \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{B}_i^T\boldsymbol{\theta}) + \frac{\lambda}{2}\boldsymbol{\theta}^T\boldsymbol{D}_q\boldsymbol{\theta}. \tag{3}$$

2.3 Asymptotic theory of $\hat{\boldsymbol{\beta}}(\boldsymbol{t})$

In this section, we show the asymptotic properties of $\hat{\beta}(t)$ based on full data. To get the desired results, here we assume that the following assumptions are satisfied.

Assumption 1 For the functional covariate $\boldsymbol{X}(t)$, assume there exist a constant C_1 such that $\|\boldsymbol{X}(t)\|_2 \leq C_1 < \infty$ a.s..

Assumption 2 Assume the unknown functional coefficient $\beta(t)$ is sufficiently smooth. That is, $\beta(t)$ has a d'-th derivative $\beta^{(d')}(t)$ such that

$$| \beta^{(d')}(t) - \beta^{(d')}(s) | \leq C_2 | t - s |^v, \quad t, s \in [0, 1],$$

where the constant $C_2 > 0$ and $v \in [0, 1]$. In what follows, we set $d = d' + v \geq p + 1$.

Assumption 3 Assume the density functions $f_{\epsilon|\boldsymbol{X}(t)}(\epsilon_i, x_i(t))$, $i = 1, 2, \ldots, n$, are continuous and uniformly bounded away from 0 and ∞ at $\epsilon_i = 0$. Furthermore, assume $\max_{i=1,2,\ldots,n} \mathrm{E}(\epsilon_i^4) < \infty$.

Assumption 4 Assume the smoothing parameter λ satisfies $\lambda = o(n^{1/2} K^{1/2-2q})$ with $q \leq p$.

Assumption 5 Assume the number of knots $K = o(n^{1/2})$ and $K/n^{1/(2d+1)} \to \infty$ as $n \to \infty$.

Remark 1 Assumptions 1 and 2 are quite usual in the functional setting; see e.g., (Cardot et al. 2005; Claeskens et al. 2009; Yoshida 2013). Assumption 3 is a regular condition also used in (Koenker 2005; Cardot et al. 2005) and can imply the uniqueness of the conditional quantile of order τ. Assumptions 4 and 5 are used to ensure the unbiasedness of the estimator (Liu et al. 2021).

To describe the asymptotic form of $\hat{\beta}(t)$, we also need the following preparations. Define $\boldsymbol{G} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{B}_i \boldsymbol{B}_i^T$, $\boldsymbol{G}_\tau = \frac{1}{n} \sum_{i=1}^n f_{\epsilon|\boldsymbol{X}(t)}(0, x_i(t)) \boldsymbol{B}_i \boldsymbol{B}_i^T$, and

$$\boldsymbol{H}_\tau = \boldsymbol{G}_\tau + \lambda/n \boldsymbol{D}_q. \tag{4}$$

Then, we have $\|\boldsymbol{G}\|_\infty = O(K^{-1})$ and $\|\boldsymbol{D}_q\|_\infty = O(K^{2q-1})$, where $\|\boldsymbol{A}\|_\infty = \max_{ij}\{| a_{ij} |\}$ for a matrix $\boldsymbol{A} = (a_{ij})$; see Lemma 1 in the Appendix. Related results can also be found in (Cardot et al. 2003; Claeskens et al. 2009; Liu et al. 2021) and the references therein. Meanwhile, combining Assumptions 3 and 4, we have $\|\boldsymbol{H}_\tau^{-1}\|_\infty = O(K)$. Furthermore, Assumption 2 implies that there exists a spline function $\beta_0(t) = \boldsymbol{B}^T(t)\boldsymbol{\theta}_0$, called spline approximation of $\beta(t)$, which as $K \to \infty$, satisfies

$$\sup_{t \in [0,1]} | \beta(t) + b_a(t) - \boldsymbol{B}^T(t)\boldsymbol{\theta}_0 | = o(K^{-d}),$$

where

$$b_a(t) = -\frac{\beta^d(t)}{K^d d!} \sum_{j=0}^{K} I(t_j \leq t < t_{j+1}) \mathrm{Br}_d \left(\frac{t - t_j}{K^{-1}} \right) = O(K^{-d})$$

is the spline approximation bias with $I(a < x < b)$ being the indicator function of an interval (a, b) and $\mathrm{Br}_d(t)$ being the d-th Bernoulli polynomial; see e.g., Zhou et al. (1998). Thus, the penalized spline quantile estimator can be decomposed as

$$\hat{\beta}(t) - \beta(t) = \hat{\beta}(t) - \beta_0(t) + \beta_0(t) - \beta(t) = \hat{\beta}(t) - \beta_0(t) + b_a(t) + o(K^{-d}).$$

Now, we present the asymptotic distribution of $\hat{\beta}(t)$ in the following theorem.

Theorem 1 *Under the Assumptions 1–3, for $t \in [0, 1]$, as $n \to \infty$, we have*

$$\left\{\boldsymbol{B}(t)^T \boldsymbol{V}_0 \boldsymbol{B}(t)\right\}^{-1/2} \sqrt{n/K} \left(\hat{\beta}(t) - \beta(t) - b_a(t) - b_\lambda(t)\right) \to N(0, 1),$$

where the shrinkage bias is define as

$$b_\lambda(t) = -\frac{\lambda}{n} \boldsymbol{B}^T(t) \boldsymbol{H}_\tau^{-1} \boldsymbol{D}_q \boldsymbol{\theta}_0 = O(\lambda K^{2q}/n),$$

and $\boldsymbol{V}_0$ is the asymptotic variance-covariance of $\sqrt{n/K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and is given as

$$\boldsymbol{V}_0 = \frac{\tau(1-\tau)}{K} \boldsymbol{H}_\tau^{-1} \boldsymbol{G} \boldsymbol{H}_\tau^{-1} = O(1).$$

Since Assumption 5 ensures that the order of K is n^v, where $v \geq 1/(2d+1)$, the spline approximation bias $b_a(t) = O(K^{-d})$ is negligible as $n \to \infty$. In addition, from Assumption 4, we can get $b_\lambda(t) = o(\sqrt{K/n})$. Thus the shrinkage bias is also negligible. By the above discussions, we have the following theorem.

Theorem 2 *Under the Assumptions 1–5, for $t \in [0, 1]$, as $n \to \infty$,*

$$\{\boldsymbol{B}(t)^T \boldsymbol{V}_0 \boldsymbol{B}(t)\}^{-1/2} \sqrt{n/K} (\hat{\beta}(t) - \beta(t)) \to N(0, 1),$$

where $\boldsymbol{V}_0$ is given in Theorem 1.

3 Optimal subsampling

3.1 Subsampling estimator and its asymptotic distribution

We first introduce a random subsampling approach, in which subsamples are taken at random with replacement based on some sampling distributions. Let R_i be the total number of times that the i-th data point is selected from the full data in a subsample and $\sum_{i=1}^n R_i = r$, which is carried out by using a random subsampling method with the probabilities π_i, $i = 1, \ldots, n$, such that $\sum_{i=1}^n \pi_i = 1$. Each R_i has a binomial distribution $\mathrm{Bin}(r, \pi_i)$ since we use subsampling with replacement. Because π_i may depend on the full data $\mathcal{F}_n = \{(x_i(t), y_i), i = 1, \ldots, n, t \in [0, 1]\}$, we need to add inverses of π_i's as weights to the objective function of the subsample to guarantee that

the loss function is unbiased. Thus, the subsampling estimator of the spline coefficient vector, says $\tilde{\boldsymbol{\theta}}$, is determined by minimizing

$$L^*(\boldsymbol{\theta}; \lambda, K) = \frac{1}{r}\sum_{i=1}^{n} \frac{R_i \rho_\tau(y_i - \boldsymbol{B}_i^T\boldsymbol{\theta})}{\pi_i} + \frac{\lambda}{2}\boldsymbol{\theta}^T \boldsymbol{D}_q \boldsymbol{\theta}. \tag{5}$$

Now, we investigate the asymptotic properties of $\tilde{\beta}(t) = \boldsymbol{B}^T(t)\tilde{\boldsymbol{\theta}}$ under Assumption 6 listed below, which restricts the weights in the loss function (5) and hence can be used to protect the loss function from inflating greatly by data points with extremely small subsampling probabilities. This assumption is also required in Ai et al. (2021) and Liu et al. (2021).

Assumption 6 Assume that $\max_{i=1,\ldots,n}(n\pi_i)^{-1} = O(r^{-1})$ and $r = o(K^2)$.

Theorem 3 *Under the Assumptions 1–6, letting* $\eta = \lim_{n\to\infty} r/n$, *for* $t \in [0, 1]$, *as* $r, n \to \infty$, *we have*

$$\left\{\boldsymbol{B}(t)^T \boldsymbol{V} \boldsymbol{B}(t)\right\}^{-1/2} \sqrt{r/K}\left(\tilde{\beta}(t) - \beta(t)\right) \to N(0, 1),$$

in distribution, where

$$\boldsymbol{V} = \frac{\tau(1-\tau)}{K}\boldsymbol{H}_\tau^{-1}(\boldsymbol{V}_\pi + \eta \boldsymbol{G})\boldsymbol{H}_\tau^{-1}, \quad \boldsymbol{V}_\pi = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\boldsymbol{B}_i \boldsymbol{B}_i^T}{\pi_i}. \tag{6}$$

3.2 Optimal subsampling probabilities

To better approximate $\beta(t)$, it is important to choose the proper subsampling probabilities. It would be meaningful if the asymptotic integrated mean squared error (IMSE) of $\tilde{\beta}(t)$ attains its minimum. By Theorem 3 and observing that $\tilde{\beta}(t)$ is asymptotic unbiased, we have the asymptotic IMSE of $\tilde{\beta}(t)$ as follows

$$\mathrm{IMSE}(\tilde{\beta}(t) - \beta(t)) = \frac{K}{r}\int_0^1 \boldsymbol{B}^T(t)\boldsymbol{V}\boldsymbol{B}(t)\mathrm{d}t. \tag{7}$$

Note that, in (7), $\boldsymbol{V}$ defined in (6) is the asymptotic variance-covariance matrix of $\sqrt{r/K}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and the integral inequality $\int_0^1 \boldsymbol{B}^T(t)\boldsymbol{V}\boldsymbol{B}(t)\mathrm{d}t \le \int_0^1 \boldsymbol{B}^T(t)\boldsymbol{V}'\boldsymbol{B}(t)\mathrm{d}t$ holds if and only if $\boldsymbol{V} \le \boldsymbol{V}'$ holds in the Löwner-ordering sense. Thus, we focus on minimizing the asymptotic variance-covariance matrix $\boldsymbol{V}$ and choose the subsampling probabilities such that $\mathrm{tr}(\boldsymbol{V})$ is minimized. This is called the A-optimality criterion in optimal experimental designs; see e.g., Atkinson et al. (2007). Using this criterion, we are able to derive the optimal subsampling probabilities in the following theorem.

Theorem 4 (A-optimality) *If the subsampling probabilities* $\pi_i, i = 1, \ldots, n$, *are chosen as*

$$\pi_i^{FAopt} = \frac{\|\boldsymbol{H}_\tau^{-1}\boldsymbol{B}_i\|_2}{\sum_{i=1}^n \|\boldsymbol{H}_\tau^{-1}\boldsymbol{B}_i\|_2}, \tag{8}$$

then the total asymptotic MSE of $\sqrt{r/K}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, $\mathrm{tr}(\boldsymbol{V})$, *attains its minimum, and so does the asymptotic IMSE of* $\tilde{\beta}(t)$.

However, from (4), we have that $\boldsymbol{H}_\tau$ in (8) depends on the density functions of ϵ_i $(i = 1, \ldots, n)$ at zero given the respective $x_i(t)$ and hence the implementation of this subsampling method requires reasonable estimation for all the density functions $f_{\epsilon|\boldsymbol{X}(t)}(0, x_i(t))$, which are often infeasible in practice without additional information. In addition, it also requires the chosen of smoothing parameter λ in $\boldsymbol{H}_\tau$ and the calculation of $\|\boldsymbol{H}_\tau^{-1}\boldsymbol{B}_i\|_2$, which costs $O(n(K+p+1)^2)$. These weaknesses make this optimal subsampling method not suitable for practical use. While, for the independent identically distributed (i.i.d.) errors case, the $\boldsymbol{G}_\tau$ in $\boldsymbol{H}_\tau$ can be simply replaced by $f_{\epsilon|\boldsymbol{X}(t)}(0, x(t))\boldsymbol{G}$ since $f_{\epsilon|\boldsymbol{X}(t)}(0, x_i(t)) = f_{\epsilon|\boldsymbol{X}(t)}(0, x(t))$ for all i.

As observed in (6), only $\boldsymbol{V}_\pi$ involves π_i in the asymptotic variance-covariance matrix $\boldsymbol{V}$ and $\boldsymbol{H}_\tau^{-1}\boldsymbol{V}_\pi\boldsymbol{H}_\tau^{-1} \leq \boldsymbol{H}_\tau^{-1}\boldsymbol{V}_{\pi'}\boldsymbol{H}_\tau^{-1}$ if and only if $\boldsymbol{V}_\pi \leq \boldsymbol{V}_{\pi'}$ in the Löwner-ordering. Thus, we focus on $\boldsymbol{V}_\pi$ and choose to minimize its trace, which can be interpreted as minimizing the asymptotic MSE of $\sqrt{r/K}\boldsymbol{H}_\tau(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ due to its asymptotic unbiasedness. This is called L-optimality criterion in optimal experimental designs (Atkinson et al. 2007). Therefore, to circumvent density function estimation and save calculation cost, we consider the modified optimal criterion: minimizing $\mathrm{tr}(\boldsymbol{V}_\pi)$.

Theorem 5 (L-optimality) *If the subsampling probabilities* $\pi_i, i = 1, \ldots, n$, *are chosen as*

$$\pi_i^{FLopt} = \frac{\|\boldsymbol{B}_i\|_2}{\sum_{i=1}^n \|\boldsymbol{B}_i\|_2}, \tag{9}$$

then $\mathrm{tr}(\boldsymbol{V}_\pi)$ *attains its minimum.*

The functional L-optimal subsampling probabilities π_i^{FLopt} (9) do not depend on the densities of ϵ_i given the respective $x_i(t)$, and thus are much easier to implement compared with the functional A-optimal subsampling probabilities π_i^{FAopt} in (8). In addition, π_i^{FLopt} requires $O(n(K+p+1))$ flops to compute, which is much cheaper than π_i^{FAopt} as K increases.

Furthermore, it is worth noting that the subsampling probabilities π_i^{FLopt} $(i = 1, \ldots, n)$ do not contain responses and do not depend on the covariates directly. In fact, the structural information of the covariates is described by the expression $\|\boldsymbol{B}_i\|_2 = \|\int_0^1 x_i(t)\boldsymbol{B}(t)\mathrm{d}t\|_2$, which is similar to the statistical leverage score in linear model. As a result, the subsampling probabilities result in the non-informative sampling. This allows us to try different models based on the subsamples. It is in

contrast to the subsampling probabilities used in functional linear regression, which result in the informative sampling (Liu et al. 2021).

3.3 Tuning parameters selection

There are four tuning parameters in estimation of $\beta(t)$: the number of knots K, the degree p for spline functions, the smoothing parameter λ and the order of derivation q for the estimator. However, the number of knots K is not a crucial parameter because smoothing is controlled by the roughness penalty parameter λ; see e.g., (Ruppert 2002; Cardot et al. 2003). In addition, the degree of spline functions p and the order of derivatives q are also known to be less important. This is because, in practice, we usually smooth with B-splines of degree 3 and a second-order penalty. Once other parameters are fixed, a natural way to determine the parameter λ is to minimize a leave-one-out cross-validation criterion. We preferably employ the generalized approximate cross-validation (GACV) criterion introduced by Yuan (2006) in smoothing splines problems, which is defined by

$$\mathrm{GACV}(\lambda) = \frac{\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{B}_i^T \hat{\boldsymbol{\theta}})}{n - \mathrm{df}_\lambda},$$

where df_λ denotes the effective degrees of freedom of the fit. In the present paper, we implement $\hat{\boldsymbol{\theta}} = (\boldsymbol{B}^T \boldsymbol{W} \boldsymbol{B} + \lambda \boldsymbol{D})^{-1} \boldsymbol{B}^T \boldsymbol{W} \boldsymbol{y}$ and $\mathrm{df}_\lambda = \mathrm{tr}\left(\boldsymbol{B}(\boldsymbol{B}^T \boldsymbol{W} \boldsymbol{B} + \lambda \boldsymbol{D})^{-1} \boldsymbol{B}^T \boldsymbol{W}\right)$ in the penalized iteratively reweighted least squares (PIRLS) method which is useful to solve the functional quantile regression problem; see e.g., (Cardot et al. 2005; Reiss and Huang 2012). In the above expressions, $\boldsymbol{W}$ is a diagonal matrix whose diagonal elements are weights,

$$w_i^{(k)} = \frac{\tau - I[y_i - \boldsymbol{B}_i^T \hat{\boldsymbol{\theta}}^{(k)}]}{2[y_i - \boldsymbol{B}_i^T \hat{\boldsymbol{\theta}}^{(k)}]}, \quad i = 1, 2, \ldots, n,$$

which are iterated until convergence; see Appendix A of Reiss and Huang (2012). However, using full data to select the optimal λ is computationally expensive, so we select the smoothing parameter λ by GACV under the optimal subsample data.

4 Numerical Experiments

In this section, we aim to study the finite sample performance of the proposed methods by using synthetic and real data.

4.1 Simulation

We generate the functional covariates in a similar way to that adopted in Liu et al. (2021). More specifically, the functional covariates are identically and independently generated as:

$$x_i(t) = \sum a_{ij} \boldsymbol{B}_j(t), \quad i = 1, 2, \ldots, n,$$

where $\boldsymbol{B}_j(t)$ are cubic B-spline basis functions that are sampled at 100 equally spaced points between 0 and 1. We consider the following three different distributions for the basis coefficient $\boldsymbol{A} = (a_{ij})$:

1. **mvNormal**. Multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{ij} = 0.5^{|i-j|}$;
2. **mvT3**. Multivariate t distribution with 3 degrees of freedom, $t_3(\mathbf{0}, \boldsymbol{\Sigma})$;
3. **mvT2**. Multivariate t distribution with 2 degrees of freedom, $t_2(\mathbf{0}, \boldsymbol{\Sigma})$.

The responses are generated as follows:

$$y_i = \int_0^1 x_i(t)\beta(t)\mathrm{d}t + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where the slope function $\beta(t) = 2t^2 + 0.25t + 1$ and the random errors, ϵ_i's, are generated in three cases:

1. **Normal**. The standard normal distribution;
2. **T1**. t_1 distribution ;
3. **Hetero**. The standard normal distribution times $\int_0^1 \mid x_i(t)(t+1) \mid \mathrm{d}t$.

The first two designs consider symmetric i.i.d. random errors while the last one considers conditional heteroscedastic errors.

We first take $n = 10^5$ for training, $m = 1000$ for testing and $\tau = 0.5, 0.75$ to investigate the influence of different quantile levels on performance of the proposed subsampling methods. From Assumption 5, we let the number of knots $K = \lceil n^{1/4} \rceil$. We shall compare the functional A-optimal subsampling (FAopt) and L-optimal subsampling (FLopt) methods with the uniform subsampling (Unif) method. For fair comparison, we use the same basis functions and the same smoothing parameter in the three methods with the same full data. For each τ, we will compute the root integrated mean squared error (IMSE) from 1000 repetitions:

$$\mathrm{IMSE} = \frac{1}{1000} \sum_{k=1}^{1000} \sqrt{\int_0^1 \left\{ \tilde{\beta}^{(k)}(t) - \beta(t) \right\}^2 \mathrm{d}t},$$

where $\tilde{\beta}^{(k)}(t)$ is the estimator from the k-th run. All the experiments are implemented in R programming language on a PC with an Intel I5 processor and 16GB memory.

Figures 1 and 2 display the simulation results corresponding to various subsampling sizes of 600, 800, 1000, 1200, 1400 and 1600 under different quantile levels[1]. It is clear

[1] In Figures 1, 2, and 3, the three columns correspond to the three distributions of the basis coefficients (mvNormal, mvT3, mvT2), respectively, and the three rows correspond to the three distributions of random

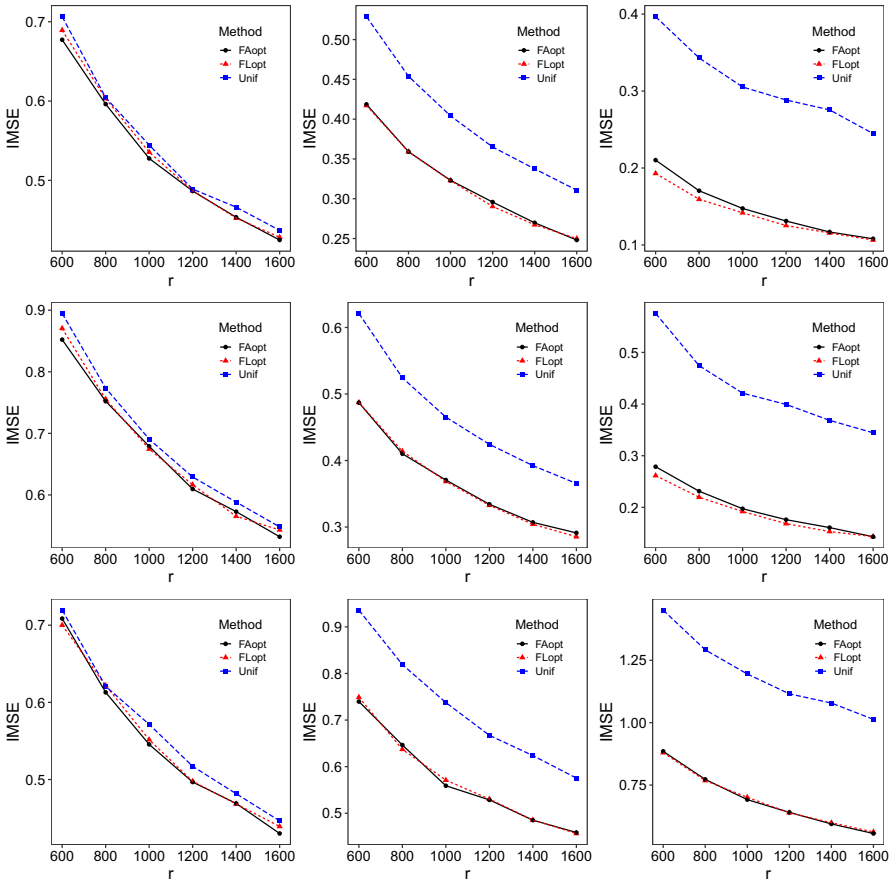

Fig. 1 IMSE for different subsampling size r with different distributions when $\tau = 0.5$ and $n = 10^5$

to see that the FAopt and FLopt subsampling methods always have smaller IMSEs than the Unif subsampling method for all cases, which is in agreement with the theoretical results that they aim to minimize the asymptotic IMSEs of the subsampling estimator. Moreover, the advantages of the FAopt and FLopt subsampling methods become more significant as the tail of the basis coefficient distribution becomes heavier. Besides, we also see that the FAopt and FLopt methods tend to perform similarly, even though the FLopt method does not theoretically minimize the MSE of the subsample spline coefficient $\tilde{\boldsymbol{\theta}}$.

To further assess the relative performance of the proposed methods in comparison with the full data estimator, the prediction efficiency (PE) is adopted on the test data of simulation, which is defined as follows:

Footnote 1 continued
errors (Normal, T1, Hetero), respectively. For example, the figure in the first column and the first row is for the mvNormal-Normal datasets.

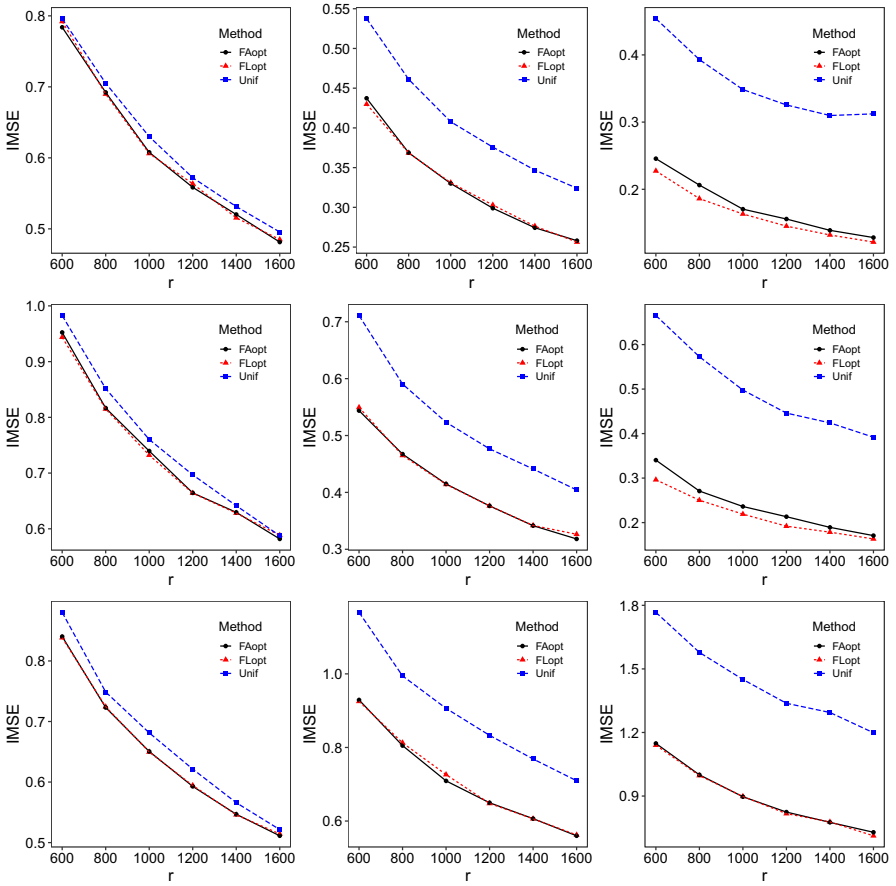

Fig. 2 IMSE for different subsampling size r with different distributions when $\tau = 0.75$ and $n = 10^5$

$$PE = \frac{\sum_i \left[\int_0^1 x_i(t)\beta(t)\mathrm{d}t - \int_0^1 x_i(t)\tilde{\beta}(t)\mathrm{d}t\right]^2}{\sum_i \left[\int_0^1 x_i(t)\beta(t)\mathrm{d}t - \int_0^1 x_i(t)\hat{\beta}(t)\mathrm{d}t\right]^2}, \quad i \in \text{testset}.$$

We plot the logarithm of prediction efficiency for the FAopt, FLopt and Unif methods when $\tau = 0.75$ in Figure 3, from which we can see that the FAopt and FLopt methods significantly outperform the Unif method, and the FLopt method has comparable or slightly smaller prediction efficiency than the FAopt method. Results for the case $\tau = 0.5$ are similar and thus are omitted.

To evaluate the computational efficiency of the subsampling methods, we record the computing time of the three subsampling methods. We use the function **Sys.time()** to count start and end times of the corresponding code only for the estimated part of $\tilde{\boldsymbol{\theta}}$. Since all the cases have similar performance, we only show the results of mvNormal–Normal datasets here. The results on different r for the FAopt, FLopt

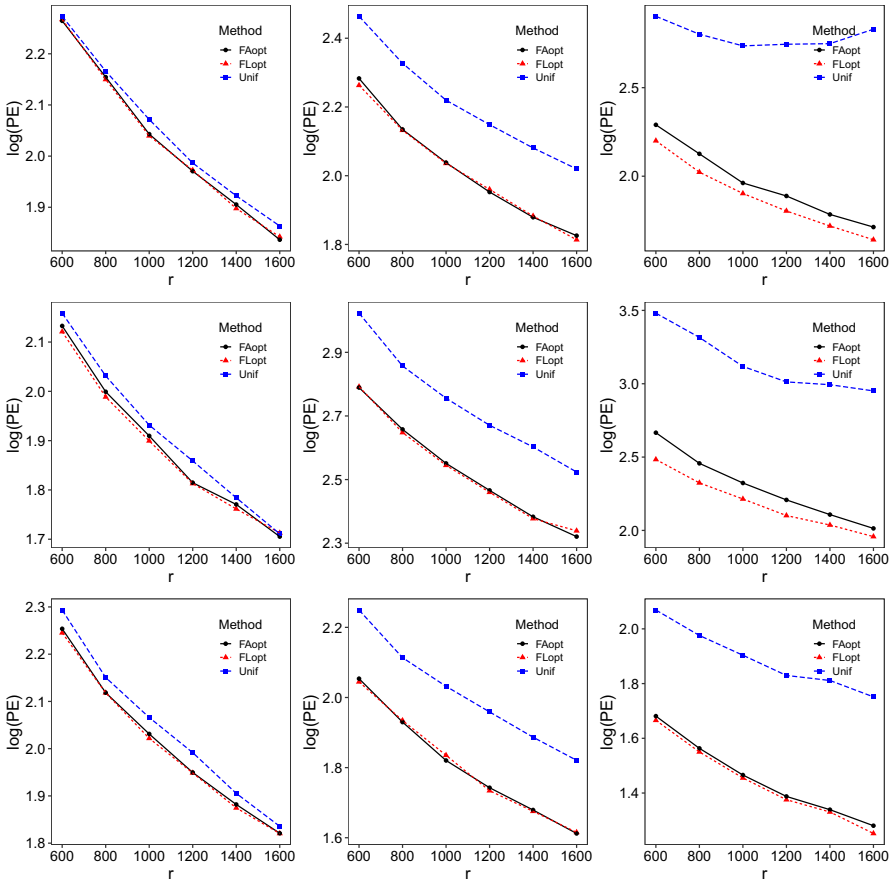

Fig. 3 Log prediction efficiency for different subsampling size r with different distributions when $\tau = 0.75$ and $n = 10^5$ for 1000 repetitions

and Unif subsampling methods with $\tau = 0.75$ and $n = 10^5$ are given in Table 1. It is not surprising to find that the Unif method takes the least time because it does not need to calculate the additional optimal subsampling probabilities. As we expected, the FLopt method is faster than the FAopt method, which agrees with the theoretical analysis. The computing time for using full data is also given in the last row of Table 1, which is the longest one and confirms that our proposed methods can reduce the computational burden.

To further demonstrate the performance of our proposed methods in large datasets, we set the full data size to $n = 10^4$, 10^5, 10^6 and 5×10^6, respectively. In addition, we let $r = 1000$, $\lambda = 0.001$ and enlarge the number of knots for spline function to $K = 50$. Table 2 presents the CPU seconds for repeating different subsampling methods 500 times. The results indicate that our proposed methods can improve the computational efficiency compared with the full data, and their advantage is more

Table 1 CPU seconds for different subsampling size r with $\tau = 0.75$ and $n = 10^5$ for 1000 repetitions

Method	r					
	600	800	1000	1200	1400	1600
FLopt	0.155	0.166	0.180	0.201	0.215	0.227
FAopt	0.472	0.462	0.469	0.496	0.515	0.533
Unif	0.115	0.133	0.142	0.161	0.178	0.193
Full data CPU seconds: 4.086						

Table 2 CPU seconds for different full data size n with $r = 1000$ when $\tau = 0.75$, $K = 50$ and $\lambda = 0.001$ for 500 repetitions

Method	n			
	10^4	10^5	10^6	5×10^6
FLopt	0.307	0.431	0.656	2.666
FAopt	0.355	0.790	5.415	33.625
Unif	0.304	0.378	0.383	0.575
Full	2.47	24.543	238.940	1668.454

significant as the full data size increases. For our two methods, we recommend the FLopt method for practical use.

4.2 Beijing multi-site air-quality data

Carbon monoxide (CO) is formed by incomplete combustion of fossil fuels and is ubiquitous in ambient air. The adverse health effects of very high CO concentrations, such as CO poisoning and cardiovascular deaths, are well documented; see e.g., (Liu et al. 2018; Kinoshita et al. 2020; Chen et al. 2021). Thus, air quality prediction is vital to management of human health, especially the respiratory system. There has been extensive research on prediction CO concentrations, see e.g., (Moazami et al. 2016; Shams et al. 2020).

Now, we analyze a data set available from https://archive-beta.ics.uci.edu/ml/datasets/beijing+multi+site+air+quality+data. This data set consists of hourly air pollutants data from 12 nationally controlled air-quality monitoring sites in Beijing from March 1, 2013 to February 28, 2017. Our primary interest here is to predict the maximum CO concentrations (mg/m^3) using the CO trajectory (24 hours) of the last day. After removing 4001 days' records with missing values, we have a dataset of 13531 days' complete records. It is randomly partitioned into a training set of $n = 10824$ observations and $m = 2707$ for testing. The raw observations are first transformed into functional data using 15 Fourier basis functions. This transformation can be implemented with the **Data2fd** function in the **fda** package, suggested in Sang and Cao (2020). A random subset of 100 curves of 24-hourly CO concentrations is presented in the left panel of Figure 4, where the time scale has been transformed to [0, 1]. The right panel of Figure 4 further supports the fact that the covariates are heavy-tailed. It depicts the histogram of the maximal values of intraday CO concentrations.

Since the true value $\beta(t)$ is unknown for real dataset, we use full data estimator instead. We calculate the empirical IMSE using eIMSE $= \frac{1}{1000}\sum_{k=1}^{1000}$

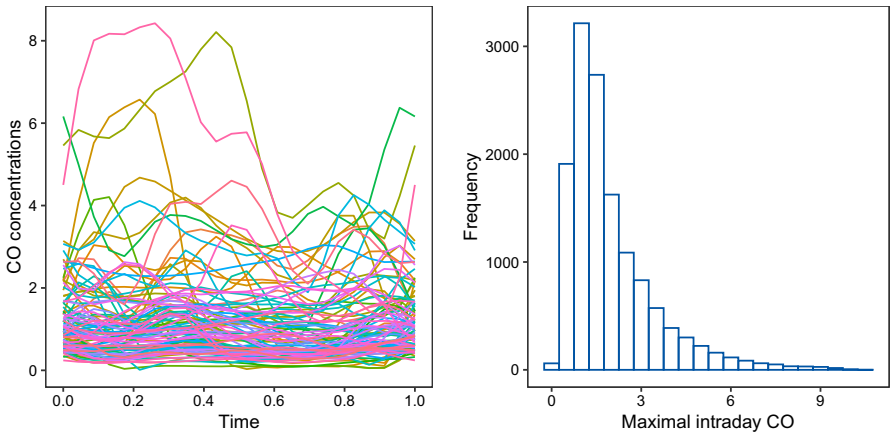

Fig. 4 Left subfigure: A random subset of 100 curves of 24-hourly CO concentrations. Right subfigure: Histogram of the maximal values of intraday CO concentrations

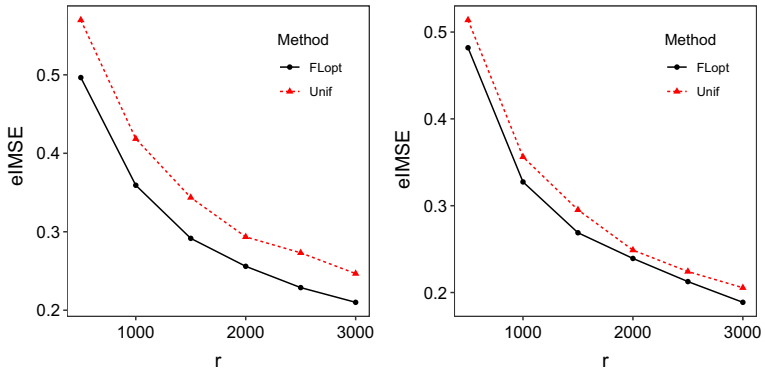

Fig. 5 eIMSE for different subsampling size r when $\tau = 0.5$ (left) and 0.75 (right)

$\sqrt{\int_0^1 \left\{\tilde{\beta}^{(k)}(t) - \hat{\beta}(t)\right\}^2 \mathrm{d}t}$, and compare the FLopt method with the Unif method. Figure 5 shows the eIMSE of subsampling estimator for different subsampling size $r = 500, 1000, 1500, 2000, 2500, 3000$ when $\tau = 0.5$ and 0.75. We can find that the FLopt method always has smaller eIMSE than the Unif method. All eIMSEs decrease as the subsampling size r gets large, showing the estimation consistency of the subsampling methods.

We further compare these two methods in terms of prediction accuracy. The relative efficiency (RE) is defined as follows:

$$RE = \frac{\sum_i \left[\int_0^1 x_i(t)\tilde{\beta}(t)\mathrm{d}t - \int_0^1 x_i(t)\hat{\beta}(t)\mathrm{d}t\right]^2}{\sum_i \left[\int_0^1 x_i(t)\hat{\beta}(t)\mathrm{d}t\right]^2}, \quad i \in \text{testset}. \tag{10}$$

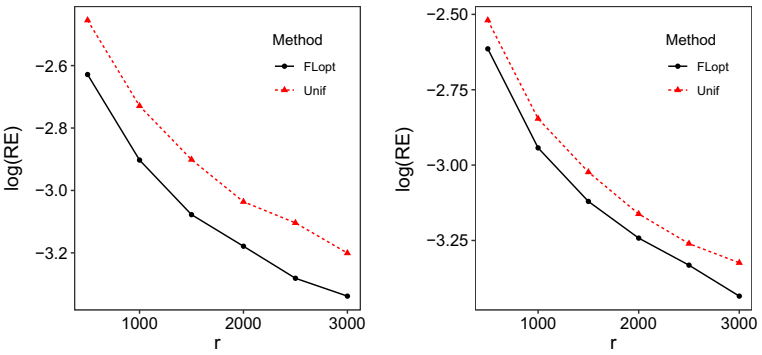

Fig. 6 Log relative efficiency for different subsampling size r when $\tau = 0.5$ (left) and 0.75 (right) for 1000 repetitions

Figure 6 displays the relative efficiency based on the subsampling methods with $\tau = 0.5$ and 0.75. In general, the relative efficiency of the subsampling estimator gradually decreases as the r increases, and the FLopt method is better than the Unif method. So it yields a better approximation to the results based on full data.

5 Conclusions

Existing optimal subsampling methods mainly focus on statistic models with scalar variables or functional mean regression. In this paper, we develop the optimal subsampling for quantile regression model when the covariates are functions. Not only is asymptotic normality estimated, but also the optimal and feasible optimal subsampling probabilities are derived according to the functional A- and L-optimality criteria, respectively. The latter results in the non-informative subsampling, which is more flexible and feasible to apply to other models compared with information sampling. Our numerical experiments show that, the FAopt and FLopt methods outperform the Unif subsampling method and are computationally feasible for massive data, and they yield good approximations to the results based on full data.

In this paper, we only consider the subsampling for the scalar-on-function quantile regression at the single quantile level. As done in (Shao and Wang 2021; Yuan et al. 2022), it is interesting to investigate multiple quantile levels. Observing that our FLopt sampling probabilities are irrelevant to quantile levels, this problem should be doable. In fact, a more interesting problem worth further investigations is how to apply optimal subsampling methods to the quantile regression process. In addition, other functional regression models are worth exploring, such as function-on-function regression and function-on-scalar regression.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No. 11671060) and the Natural Science Foundation Project of CQ CSTC (No. cstc2019jcyj-msxmX0267). The authors would like to thank the editor and the anonymous reviewers for their detailed comments and helpful suggestions.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A: proofs for theoretical results

To prove our theorems, we begin with the following several lemmas. Note that the subsampling model involves two kinds of random errors: sampling error and model error, so we need to consider these two types of randomness in the calculation.

Lemma 1 *Under Assumptions 1 and 5, for any vector* $\boldsymbol{\mu} \in \mathbb{R}^{K+p+1}$, *there are some positive constants* C_3, C_4, C_5 *and* C_6 *such that*

$$C_3 K^{-1} \leq \sigma_{min}(\boldsymbol{G}) \leq \sigma_{max}(\boldsymbol{G}) \leq C_4 K^{-1},$$
$$C_5 K^{2q-1} \|\boldsymbol{\mu}\|_2^2 \leq \boldsymbol{\mu}^T \boldsymbol{D}_q \boldsymbol{\mu} \leq C_6 K^{2q-1} \|\boldsymbol{\mu}\|_2^2,$$

where $\sigma_{min}(\cdot)$ *and* $\sigma_{max}(\cdot)$ *denote the smallest and largest eigenvalues of a matrix, respectively. In addition, we have* $\|\boldsymbol{G}\|_\infty = O(K^{-1})$ *and* $\|\boldsymbol{D}_q\|_\infty = O(K^{2q-1})$.

Proof These results can be derived directly from Lemma S2 and S3 in the supplementary file of Liu et al. (2021). □

Lemma 2 *Under Assumptions 1, and 3–5, there are two positive constants* C_7 *and* C_8 *such that*

$$C_7 K^{-1} \leq \sigma_{min}(\boldsymbol{H}_\tau) \leq \sigma_{max}(\boldsymbol{H}_\tau) \leq C_8 K^{-1},$$

and $\|\boldsymbol{H}_\tau\|_\infty = O(K^{-1})$.

Proof From Assumption 3, we have that there are two positive constants c_ϵ and C_ϵ such that $c_\epsilon \leq f_{\epsilon|\boldsymbol{X}(t)}(0, x(t)) \leq C_\epsilon$. On the other hand, by Lemma 1, we have $\|\boldsymbol{G}_\tau\|_\infty = O(K^{-1})$. Thus, the lemma can be directly proved by combining Lemma 1 with Assumptions 3 and 4. □

Lemma 3 *Let* $\psi_\tau(u) = \tau - I(u < 0)$ *and* $u_i = y_i - \boldsymbol{B}_i^T \boldsymbol{\theta}_0$. *Under the same assumptions as Theorem 3, for any non-zero* $\boldsymbol{\delta} \in \mathbb{R}^{K+p+1}$, *we have*

$$-\sqrt{\frac{K}{r}} \sum_{i=1}^{n} \frac{R_i}{n\pi_i} \boldsymbol{B}_i^T \boldsymbol{\delta} \psi_\tau(u_i) = -\sqrt{K} \boldsymbol{W}^T \boldsymbol{\delta} + o_P(1), \tag{A1}$$

where $\{\tau(1-\tau)(\boldsymbol{V}_\pi + \eta \boldsymbol{G})\}^{-1/2} \boldsymbol{W} \to N(\boldsymbol{0}, \boldsymbol{I})$ *in distribution.*

Proof Set

$$U_r = -\sqrt{\frac{K}{r}} \sum_{i=1}^{n} \frac{R_i}{n\pi_i} \boldsymbol{B}_i^T \boldsymbol{\delta} \psi_\tau(u_i).$$

To prove the asymptotic normality of U_r, it suffices to verify that U_r satisfies the Lindeberg-Feller conditions. Firstly, the conditional expectation and conditional variance are given by

$$\begin{aligned}
\mathrm{E}\{U_r \mid \mathcal{F}_n\} &= -\sqrt{\frac{K}{r}} \sum_{i=1}^{n} \mathrm{E}\left\{ \frac{R_i}{n\pi_i} \boldsymbol{B}_i^T \boldsymbol{\delta} \psi_\tau(u_i) \mid \mathcal{F}_n \right\} \\
&= -\frac{\sqrt{rK}}{n} \sum_{i=1}^{n} \boldsymbol{B}_i^T \boldsymbol{\delta} \psi_\tau(u_i), \\
\mathrm{Var}\{U_r \mid \mathcal{F}_n\} &= \frac{K}{r} \sum_{i=1}^{n} \mathrm{Var}\left\{ \frac{R_i}{n\pi_i} \boldsymbol{B}_i^T \boldsymbol{\delta} \psi_\tau(u_i) \mid \mathcal{F}_n \right\} \\
&= \frac{K}{n^2} \sum_{i=1}^{n} \frac{\pi_i(1-\pi_i)}{\pi_i^2} (\boldsymbol{B}_i^T \boldsymbol{\delta})^2 \psi_\tau^2(u_i).
\end{aligned}$$

From the fact that $\mathrm{P}(y_i < \int_0^1 x_i(t)\beta(t)\mathrm{d}t \mid x_i(t)) = \tau$, we have

$$\begin{aligned}
\mathrm{E}\{\psi_\tau(u_i) \mid x_i(t)\} &= \tau - \mathrm{E}\{I(u_i < 0) \mid x_i(t)\} \\
&= \tau - \mathrm{P}\left(y_i < \boldsymbol{B}_i^T \theta_0 \mid x_i(t) \right) \\
&= \tau - \mathrm{P}\left(y_i < \int_0^1 x_i(t)(\beta(t) + b_a(t)(1 + o_P(1)))\mathrm{d}t \mid x_i(t) \right) \\
&= -b_i f_{\epsilon|\boldsymbol{X}(t)}(0, x_i(t))(1 + o_P(1)) \\
&= o_P(1),
\end{aligned}$$

where $b_i = \int_0^1 x_i(t) b_a(t) \mathrm{d}t$, and the third equality is from the definition of $\boldsymbol{\theta}_0$ and the fourth equality is obtained by the Taylor expansion of the cumulative distribution function of the error ϵ_i at point $\epsilon_i = 0$. As a result, the unconditional expectation of U_r can be calculated as

$$\begin{aligned}
\mathrm{E}[U_r] &= -\frac{\sqrt{rK}}{n} \mathrm{E}\left\{ \sum_{i=1}^{n} \boldsymbol{B}_i^T \boldsymbol{\delta} \psi_\tau(u_i) \mid x_i(t) \right\} \\
&= \frac{\sqrt{rK}}{n} \sum_{i=1}^{n} \boldsymbol{B}_i^T \boldsymbol{\delta} b_i f_{\epsilon|\boldsymbol{X}(t)}(0, x_i(t))(1 + o_P(1)) \\
&= O(\sqrt{rK} K^{-(d+1)}). \qquad \text{(A2)}
\end{aligned}$$

More specifically, since $x_i(t)$ are square integrable functions, by the Cauchy-Schwarz inequality in integral form, there exist constant c such that

$$\begin{aligned}\boldsymbol{B}_i^2 &= \left(\int_0^1 x_i(t)\boldsymbol{B}(t)\mathrm{d}t\right)^2 \\ &\le \int_0^1 x_i^2(t)\mathrm{d}t \cdot \int_0^1 \boldsymbol{B}^2(t)\mathrm{d}t \le c\int_0^1 \boldsymbol{B}^2(t)\mathrm{d}t.\end{aligned}$$

Similarly, we have

$$b_i^2 \le c\int_0^1 b_a^2(t)\mathrm{d}t.$$

Thus, by the property of B-spline function, $\int_0^1 \boldsymbol{B}(t)\mathrm{d}t = O(K^{-1})$, and $b_a(t) = O(K^{-d})$, we can find that $\|\boldsymbol{B}_i\|_\infty = O(K^{-1})$ and $b_i = O(K^{-d})$ are satisfied. Putting them together, we obtain (A2).

On the other hand, according to law of total variance, the unconditional variance is given by

$$\begin{aligned}\mathrm{Var}\,[U_r] &= \mathrm{Var}\left\{-\sqrt{\frac{K}{r}}\sum_{i=1}^n \frac{R_i}{n\pi_i}\boldsymbol{B}_i^T\boldsymbol{\delta}\psi_\tau(u_i)\right\} \\ &= \mathrm{E}\left\{\mathrm{Var}\left\{-\sqrt{\frac{K}{r}}\sum_{i=1}^n \frac{R_i}{n\pi_i}\boldsymbol{B}_i^T\boldsymbol{\delta}\psi_\tau(u_i)\mid\mathcal{F}_n\right\}\right\} \\ &\quad +\mathrm{Var}\left\{\mathrm{E}\left\{-\sqrt{\frac{K}{r}}\sum_{i=1}^n \frac{R_i}{n\pi_i}\boldsymbol{B}_i^T\boldsymbol{\delta}\psi_\tau(u_i)\mid\mathcal{F}_n\right\}\right\}. \qquad \text{(A3)}\end{aligned}$$

We first deal with the first term in (A3) as follows

$$\begin{aligned}&\mathrm{E}\left\{\mathrm{Var}\left\{-\sqrt{\frac{K}{r}}\sum_{i=1}^n \frac{R_i}{n\pi_i}\boldsymbol{B}_i^T\boldsymbol{\delta}\psi_\tau(u_i)\mid\mathcal{F}_n\right\}\right\} \\ &= \frac{K}{n^2}\mathrm{E}\left\{\sum_{i=1}^n \frac{\pi_i(1-\pi_i)}{\pi_i^2}(\boldsymbol{B}_i^T\boldsymbol{\delta})^2\psi_\tau^2(u_i)\mid x_i(t)\right\} \\ &= K\tau(1-\tau)\boldsymbol{\delta}^T\left\{\sum_{i=1}^n \frac{\boldsymbol{B}_i\boldsymbol{B}_i^T}{n^2\pi_i} - \sum_{i=1}^n \frac{\boldsymbol{B}_i\boldsymbol{B}_i^T}{n^2}\right\}\boldsymbol{\delta}(1+o_P(1)). \qquad \text{(A4)}\end{aligned}$$

Similarly, the second term in (A3) equals

$$\mathrm{Var}\left\{\mathrm{E}\left\{-\sqrt{\frac{K}{r}}\sum_{i=1}^n \frac{R_i}{n\pi_i}\boldsymbol{B}_i^T\boldsymbol{\delta}\psi_\tau(u_i)\mid\mathcal{F}_n\right\}\right\}$$

$$= \frac{rK}{n^2}\text{Var}\left\{\sum_{i=1}^{n} \boldsymbol{B}_i^T \boldsymbol{\delta}\psi_\tau(u_i) \mid x_i(t)\right\}$$
$$= rK\tau(1-\tau)\boldsymbol{\delta}^T\left(\sum_{i=1}^{n}\frac{\boldsymbol{B}_i\boldsymbol{B}_i^T}{n^2}\right)\boldsymbol{\delta}(1+o_P(1)). \quad \text{(A5)}$$

Thus, substituting (A4) and (A5) into (A3), we have

$$\text{Var}[U_r] = K\tau(1-\tau)\boldsymbol{\delta}^T\left\{\sum_{i=1}^{n}\frac{\boldsymbol{B}_i\boldsymbol{B}_i^T}{n^2\pi_i} + \frac{r-1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{B}_i\boldsymbol{B}_i^T}{n}\right\}\boldsymbol{\delta}(1+o_P(1))$$
$$= K\tau(1-\tau)\boldsymbol{\delta}^T\left(\boldsymbol{V}_\pi + \eta\boldsymbol{G}\right)\boldsymbol{\delta}(1+o_P(1)). \quad \text{(A6)}$$

Denote $\xi_i = -\sqrt{\frac{K}{r}}\frac{R_i}{n\pi_i}\boldsymbol{B}_i^T\boldsymbol{\delta}\psi_\tau(u_i)$. We now check the Lindeberg-Feller conditions. For every $\epsilon > 0$,

$$\sum_{i=1}^{n}\text{E}\left\{\|\xi_i\|^2 I(\|\xi_i\| > \epsilon)\right\} \le \frac{1}{\epsilon}\sum_{i=1}^{n}\text{E}\left\{\|\xi_i\|^3\right\}$$
$$\le \left(\frac{K}{r}\right)^{3/2}\frac{1}{\epsilon}\sum_{i=1}^{n}\text{E}\left\{\frac{R_i^3\|\boldsymbol{B}_i^T\boldsymbol{\delta}\|^3\|\psi_\tau(u_i)\|^3}{n^3\pi_i^3}\right\}$$
$$= \left(\frac{K}{r}\right)^{3/2}\frac{1}{\epsilon}\sum_{i=1}^{n}\frac{\text{E}\left[R_i^3\right] \mid \boldsymbol{B}_i^T\boldsymbol{\delta}\mid^3 \text{E}\left\{\|\psi_\tau(u_i)\|^3 \mid x_i(t)\right\}}{n^3\pi_i^3}$$
$$= o_P(1), \quad \text{(A7)}$$

where

$$\text{E}\left[R_i^3\right] = r(r-1)(r-2)\pi_i^3 + 3r(r-1)\pi_i^2 + r\pi_i,$$

and the last equality holds by combining Assumption 6 and the fact that $\mid\psi_\tau(u_i)\mid \le 1$. Thus, by Lindeberg-Feller central limit theorem, it can be concluded that as $n \to \infty$, $r \to \infty$,

$$\frac{U_r - \text{E}[U_r]}{\sqrt{\text{Var}[U_r]}} \to N(0,1)$$

in distribution, which implies that the equation (A1) holds because $\text{E}[U_r] = O(\sqrt{rK}K^{-(d+1)}) = o_P(1)$. This completes the proof. □

Lemma 4 *Let* $v_i = \sqrt{K/r}\boldsymbol{B}_i^T$. *Under the same assumptions as Theorem 3,*

$$\sum_{i=1}^{n}\frac{R_i\int_0^{v_i}\{I(u_i \le s) - I(u_i \le 0)\}\text{d}s}{n\pi_i} = \frac{K}{2}\boldsymbol{\delta}^T\boldsymbol{G}_\tau\boldsymbol{\delta} + o_P(1).$$

Proof Let

$$M_r = \sum_{i=1}^{n} \frac{R_i \int_0^{v_i} \{I(u_i \le s) - I(u_i \le 0)\} \mathrm{d}s}{n\pi_i}.$$

Since

$$\begin{aligned}
&\mathrm{E}\left\{ \frac{R_i \int_0^{v_i} \{I(u_i \le s) - I(u_i \le 0)\} \, \mathrm{d}s}{n\pi_i} \right\} \\
&= \mathrm{E}\left\{ \mathrm{E}\left\{ \frac{R_i \int_0^{v_i} \{I(u_i \le s) - I(u_i \le 0)\} \, \mathrm{d}s}{n\pi_i} \mid \mathcal{F}_n \right\} \right\} \\
&= \frac{r}{n} \mathrm{E}\left\{ \int_0^{v_i} \{I(u_i \le s) - I(u_i \le 0)\} \, \mathrm{d}s \mid x_i(t) \right\} \\
&= \frac{r}{n} \int_0^{v_i} \left\{ \mathrm{P}\left(y_i < \boldsymbol{B}_i^T \boldsymbol{\theta}_0 + s \mid x_i(t) \right) - \mathrm{P}\left(y_i < \boldsymbol{B}_i^T \boldsymbol{\theta}_0 \mid x_i(t) \right) \right\} \mathrm{d}s \\
&= \frac{\sqrt{rK}}{n} \int_0^{\boldsymbol{B}_i^T \boldsymbol{\delta}} \left\{ \mathrm{P}\left(y_i < \boldsymbol{B}_i^T \boldsymbol{\theta}_0 + l\sqrt{\frac{K}{r}} \mid x_i(t) \right) - \mathrm{P}\left(y_i < \boldsymbol{B}_i^T \boldsymbol{\theta}_0 \mid x_i(t) \right) \right\} \mathrm{d}l \\
&= \frac{K}{n} \int_0^{\boldsymbol{B}_i^T \boldsymbol{\delta}} f_{\epsilon|\boldsymbol{X}(t)}(\boldsymbol{B}_i^T \boldsymbol{\theta}_0, x_i(t)) l \mathrm{d}l \cdot (1 + o_P(1)) \\
&= \frac{K}{2n} f_{\epsilon|\boldsymbol{X}(t)}(\boldsymbol{B}_i^T \boldsymbol{\theta}_0, x_i(t)) (\boldsymbol{B}_i^T \boldsymbol{\delta})^2 (1 + o_P(1)),
\end{aligned}$$

we can obtain the total expectation of M_r as follows

$$\begin{aligned}
\mathrm{E}[M_r] &= \frac{K}{2n} \sum_{i=1}^{n} f_{\epsilon|\boldsymbol{X}(t)}(\boldsymbol{B}_i^T \boldsymbol{\theta}_0, x_i(t)) (\boldsymbol{B}_i^T \boldsymbol{\delta})^2 (1 + o_P(1)) \\
&= \frac{K}{2} \boldsymbol{\delta}^T \left(\frac{1}{n} \sum_{i=1}^{n} f_{\epsilon|\boldsymbol{X}(t)}(0 + o(1), x_i(t)) \boldsymbol{B}_i \boldsymbol{B}_i^T \right) \boldsymbol{\delta} (1 + o_P(1)) \\
&= \frac{K}{2} \boldsymbol{\delta}^T \boldsymbol{G}_\tau \boldsymbol{\delta} (1 + o_P(1)). \qquad \text{(A8)}
\end{aligned}$$

Now, we show the total variance of M_r satisfying $\mathrm{Var}[M_r] = o_P(1)$. Note that the variance of M_r can be evaluated as

$$\begin{aligned}
\mathrm{Var}[M_r] &\le \sum_{i=1}^{n} \mathrm{E}\left\{ \frac{R_i \int_0^{v_i} \{I(u_i \le s) - I(u_i \le 0)\} \mathrm{d}s}{n\pi_i} \right\}^2 \\
&\le \sqrt{\frac{K}{r}} \left\{ \max_{i=1,2,\ldots,n} \frac{\| \boldsymbol{B}_i^T \boldsymbol{\delta} \|}{n\pi_i} \right\} \cdot \mathrm{E}[M_r]
\end{aligned}$$

$$\leq \sqrt{\frac{K}{r}} \left\{ \max_{i=1,2,\ldots,n} \frac{1}{n\pi_i} \right\} \cdot \left\{ \max_{i=1,2,\ldots,n} | \boldsymbol{B}_i^T \boldsymbol{\delta} | \right\} \cdot \mathrm{E}[M_r], \tag{A9}$$

where the second inequality is from the fact that

$$\begin{aligned} \int_0^{v_i} \{I(u_i \leq s) - I(u_i \leq 0)\} \mathrm{d}s &\leq \left| \int_0^{v_i} |\{I(u_i \leq s) - I(u_i \leq 0)\}| \, \mathrm{d}s \right| \\ &\leq \sqrt{\frac{K}{r}} \left| \boldsymbol{B}_i^T \boldsymbol{\delta} \right|, \quad i = 1, 2, \ldots, n. \end{aligned}$$

Thus, from (A8), (A9) and Assumption 6, and noting $\mathrm{E}[M_r] = O(1)$, we have $\mathrm{Var}[M_r] = o_P(\sqrt{K/r^3}) = o_P(1)$. As a result, Lemma 4 holds by Chebyshev's inequality. □

In the following, we present the proofs of Theorems 1, 2, 3, 4, and 5 in turn.

Proof of Theorem 1 and 2 Theorem 1 can be proved similar to Theorem 1 of Yoshida (2013), and Theorem 2 can be obtained directly from Theorem 1 by considering Assumptions 4 and 5. Here we omit the details. □ □

Proof of Theorem 3 Let

$$\begin{aligned} Z_r(\boldsymbol{\delta}) = &\sum_{i=1}^{n} \frac{R_i(\rho_\tau(u_i - v_i) - \rho_\tau(u_i))}{\pi_i} \\ &+ \frac{r\lambda}{2}(\boldsymbol{\theta}_0 + \sqrt{\frac{K}{r}}\boldsymbol{\delta})^T \boldsymbol{D}_q(\boldsymbol{\theta}_0 + \sqrt{\frac{K}{r}}\boldsymbol{\delta}) - \frac{r\lambda}{2}\boldsymbol{\theta}_0^T \boldsymbol{D}_q \boldsymbol{\theta}_0, \end{aligned}$$

where $u_i = y_i - \boldsymbol{B}_i^T \boldsymbol{\theta}_0$ and $v_i = \sqrt{r/K}\boldsymbol{B}_i^T \boldsymbol{\delta}$. It is easy to see that this function is convex and minimized at $\sqrt{r/K}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

On the other hand, using Knight's identity,

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v \{I(u \leq s) - I(u \leq 0)\} \mathrm{d}s, \tag{A10}$$

where $\psi_\tau(u) = \tau - I(u < 0)$, we have

$$Z_r(\boldsymbol{\delta}) = Z_{1r}(\boldsymbol{\delta}) + Z_{2r}(\boldsymbol{\delta}) + Z_{3r}(\boldsymbol{\delta}) + Z_{4r}(\boldsymbol{\delta}), \tag{A11}$$

where

$$\begin{aligned} Z_{1r}(\boldsymbol{\delta}) &= -\sqrt{\frac{K}{r}} \sum_{i=1}^{n} \frac{R_i}{\pi_i} \boldsymbol{B}_i^T \boldsymbol{\delta} \psi_\tau(u_i), \\ Z_{2r}(\boldsymbol{\delta}) &= \sum_{i=1}^{n} \frac{R_i \int_0^{v_i} \{I(u_i \leq s) - I(u_i \leq 0)\} \, \mathrm{d}s}{\pi_i}, \end{aligned}$$

$$Z_{3r}(\boldsymbol{\delta}) = \frac{K\lambda}{2}\boldsymbol{\delta}^T \boldsymbol{D}_q \boldsymbol{\delta},$$
$$Z_{4r}(\boldsymbol{\delta}) = \sqrt{rK}\lambda \boldsymbol{\theta}_0^T \boldsymbol{D}_q \boldsymbol{\delta}.$$

From Lemma 3, $Z_{1r}(\boldsymbol{\delta})$ in (A11) satisfies

$$\frac{Z_{1r}(\boldsymbol{\delta})}{n} = -\sqrt{K}\boldsymbol{W}^T\boldsymbol{\delta} + o_P(1), \tag{A12}$$

where $\{\tau(1-\tau)(\boldsymbol{V}_\pi + \eta\boldsymbol{G})\}^{-1/2}\boldsymbol{W} \to N(\boldsymbol{0}, \boldsymbol{I})$ in distribution. Furthermore, Lemma 4 and $Z_{3r}(\boldsymbol{\delta})$ in (A11) yield

$$\frac{Z_{2r}(\boldsymbol{\delta})}{n} + \frac{Z_{3r}(\boldsymbol{\delta})}{n} = \frac{K}{2}\boldsymbol{\delta}^T\left(\boldsymbol{G}_\tau + \frac{\lambda}{n}\boldsymbol{D}_q\right)\boldsymbol{\delta} + o_P(1) = \frac{K}{2}\boldsymbol{\delta}^T\boldsymbol{H}_\tau\boldsymbol{\delta} + o_P(1). \tag{A13}$$

Therefore, from (A11), (A12) and (A13), we can obtain

$$\frac{Z_r(\boldsymbol{\delta})}{n} = -\sqrt{K}\boldsymbol{W}^T\boldsymbol{\delta} + \frac{K}{2}\boldsymbol{\delta}^T\boldsymbol{H}_\tau\boldsymbol{\delta} + \frac{\sqrt{rK}}{n}\lambda\boldsymbol{\theta}_0^T\boldsymbol{D}_q\boldsymbol{\delta} + o_P(1).$$

Since $Z_r(\boldsymbol{\delta})/n$ is convex with respect to $\boldsymbol{\delta}$ and has unique minimizer, from the corollary in page 2 of Hjort and Pollard (2011), its minimizer, $\sqrt{r/K}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, satisfies that

$$\sqrt{\frac{r}{K}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \boldsymbol{H}_\tau^{-1}\left(\frac{1}{\sqrt{K}}\boldsymbol{W} - \sqrt{\frac{r}{K}}\cdot\frac{\lambda}{n}\boldsymbol{D}_q\boldsymbol{\theta}_0\right) + o_P(1).$$

Because the random vector is only $\boldsymbol{W}$ in asymptotic form of $\tilde{\boldsymbol{\theta}}$ and $\tilde{\beta}(t) - \beta_0(t) = \boldsymbol{B}^T(t)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, the expectation of $\tilde{\beta}(t) - \beta_0(t)$ can be written as

$$\mathrm{E}\{\tilde{\beta}(t) - \beta_0(t)\} = b_\lambda(t)(1 + o_P(1)),$$

where $b_\lambda(t) = -\frac{\lambda}{n}\boldsymbol{B}^T(t)\boldsymbol{H}_\tau^{-1}\boldsymbol{D}_q\boldsymbol{\theta}_0$. Together with $\tilde{\beta}(t) - \beta(t) = \tilde{\beta}(t) - \beta_0(t) + \beta_0(t) - \beta(t)$, we have the asymptotic bias of $\tilde{\beta}(t)$ as

$$\mathrm{E}\{\tilde{\beta}(t) - \beta(t)\} = b_a(t)(1 + o_P(1)) + b_\lambda(t)(1 + o_P(1)).$$

Thus, we have

$$\begin{aligned}&\{\boldsymbol{B}(t)^T\boldsymbol{V}\boldsymbol{B}(t)\}^{-1/2}\sqrt{r/K}(\tilde{\beta}(t) - \beta(t) - b_a(t) - b_\lambda(t))\\ =&\{\boldsymbol{B}(t)^T\boldsymbol{V}\boldsymbol{B}(t)\}^{-1/2}\boldsymbol{B}^T(t)\boldsymbol{H}_\tau^{-1}\frac{1}{\sqrt{K}}\boldsymbol{W} + o_P(1).\end{aligned}$$

Combining the fact that

$$\{\boldsymbol{B}(t)^T\boldsymbol{V}\boldsymbol{B}(t)\}^{-1/2}\boldsymbol{B}^T(t)\boldsymbol{V}\boldsymbol{B}(t)\{\boldsymbol{B}(t)^T\boldsymbol{V}\boldsymbol{B}(t)\}^{-1/2} = 1,$$

by the definition of $\boldsymbol{W}$ and Slutsky's Theorem, we can obtain for $t \in [0, 1]$, as $r, n \to \infty$,

$$\{\boldsymbol{B}(t)^T \boldsymbol{V} \boldsymbol{B}(t)\}^{-1/2} \sqrt{r/K} (\tilde{\beta}(t) - \beta(t) - b_a(t) - b_\lambda(t)) \to N(0, 1).$$

Further, from the discussions before Theorem 2, we know that $b_\lambda(t)$ and $b_a(t) = o_P(1)$ are negligible. Thus, we have

$$\{\boldsymbol{B}(t)^T \boldsymbol{V} \boldsymbol{B}(t)\}^{-1/2} \sqrt{r/K} (\tilde{\beta}(t) - \beta(t)) \to N(0, 1).$$

So Theorem 3 is proved. □

Proof of Theorem 4 Note that

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{V}) &= \frac{\tau(1-\tau)}{K} \mathrm{tr}\left[\boldsymbol{H}_\tau^{-1} \left(\sum_{i=1}^n \frac{\boldsymbol{B}_i \boldsymbol{B}_i^T}{n^2 \pi_i} + \eta \sum_{i=1}^n \frac{\boldsymbol{B}_i \boldsymbol{B}_i^T}{n} \right) \boldsymbol{H}_\tau^{-1} \right] \\
&= \frac{\tau(1-\tau)}{Kn^2} \sum_{i=1}^n \mathrm{tr}\left[\frac{\boldsymbol{H}_\tau^{-1} \boldsymbol{B}_i \boldsymbol{B}_i^T \boldsymbol{H}_\tau^{-1}}{\pi_i} \right] \\
&\quad + \frac{\tau(1-\tau)\eta}{Kn} \sum_{i=1}^n \mathrm{tr}\left[H_\tau^{-1} \boldsymbol{B}_i \boldsymbol{B}_i^T \boldsymbol{H}_\tau^{-1} \right] \\
&= \frac{\tau(1-\tau)}{Kn^2} \sum_{i=1}^n \frac{\| H_\tau^{-1} \boldsymbol{B}_i \|_2^2}{\pi_i} + \frac{\tau(1-\tau)\eta}{Kn} \sum_{i=1}^n \| \boldsymbol{H}_\tau^{-1} \boldsymbol{B}_i \|_2^2 \\
&= \frac{\tau(1-\tau)}{Kn^2} \left(\sum_{i=1}^n \pi_i \right) \left(\sum_{i=1}^n \frac{\| H_\tau^{-1} \boldsymbol{B}_i \|_2^2}{\pi_i} \right) + \frac{\tau(1-\tau)\eta}{Kn} \sum_{i=1}^n \| H_\tau^{-1} \boldsymbol{B}_i \|_2^2 \\
&\geq \frac{\tau(1-\tau)}{Kn^2} \left(\sum_{i=1}^n \| H_\tau^{-1} \boldsymbol{B}_i \|_2 \right)^2 + \frac{\tau(1-\tau)\eta}{Kn} \sum_{i=1}^n \| H_\tau^{-1} \boldsymbol{B}_i \|_2^2,
\end{aligned}$$

where the last inequality is from the Cauchy-Schwarz inequality and the equality in it holds if and only if $\pi_i \propto \| \boldsymbol{H}_\tau^{-1} \boldsymbol{B}_i \|_2$. So the proof is completed by considering $\sum_{i=1}^n \pi_i = 1$. □

Proof of Theorem 5 Note that

$$\begin{aligned}
\mathrm{tr}\,[\boldsymbol{V}_\pi] &= \mathrm{tr}\left(\sum_{i=1}^n \frac{\boldsymbol{B}_i \boldsymbol{B}_i^T}{n^2 \pi_i} \right) = \frac{1}{n^2} \sum_{i=1}^n \mathrm{tr}\left(\frac{\boldsymbol{B}_i \boldsymbol{B}_i^T}{\pi_i} \right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \frac{\| \boldsymbol{B}_i \|_2^2}{\pi_i} = \frac{1}{n^2} \left(\sum_{i=1}^n \pi_i \right) \left(\sum_{i=1}^n \frac{\| \boldsymbol{B}_i \|_2^2}{\pi_i} \right) \\
&\geq \frac{1}{n^2} \left(\sum_{i=1}^n \| \boldsymbol{B}_i \|_2 \right)^2,
\end{aligned}$$

where the last inequality is from the Cauchy-Schwarz inequality and the equality in it holds if and only if $\pi_i \propto \|\boldsymbol{B}_i\|_2$. So the proof is completed by considering $\sum_{i=1}^{n} \pi_i = 1$. □

References

Ai M, Wang F, Yu J, Zhang H (2021) Optimal subsampling for large-scale quantile regression. J Complex 62:10512

Ai M, Yu J, Zhang H, Wang H (2021) Optimal subsampling algorithms for big data regression. Stat Sinica 31(2):749–772

Atkinson A, Donev AN, Tobias RD (2007) Optimum experimental designs, with SAS. Oxford University Press, New York

Cardot H, Ferraty F, Sarda P (2003) Spline estimators for the functional linear model. Stat Sin 13:571–591

Cardot H, Crambes C, Sarda P (2005) Quantile regression when the covariates are functions. J Nonparameter Stat 17(7):841–856

Cardot H, Crambes C, Sarda P (2004) Conditional quantiles with functional covariates: an application to ozone pollution forecasting. In: Compstat 2004 Proceedings, pp 769–776

Chen K, Müller H (2012) Conditional quantile analysis when covariates are functions, with application to growth data. J R Stat Soc B 74(2):67–89

Chen K, Breitner S, Wolf K et al (2021) Ambient carbon monoxide and daily mortality: a global time-series study in 337 cities. Lancet Planet Health 5(4):e191–e199

Claeskens G, Krivobokova T, Opsomer JD (2009) Asymptotic properties of penalized spline estimators. Biometrika 96(3):529–544

de Boor C (2001) A practical guide to splines. Springer, Berlin

Dobriban E, Liu S (2019) Asymptotics for sketching in least squares regression. In: Advances in Neural Information Processing Systems 32, pp 3675–3685

Drineas P, Magdon-Ismail M, Mahoney MW, Woodruff DP (2012) Fast approximation of matrix coherence and statistical leverage. J Mach Learn Res 13(1):3441–3472

Drineas P, Mahoney MW, Muthukrishnan S (2006) Sampling algorithms for l_2 regression and applications. In: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm, pp 1127–1136

Fan Y, Liu Y, Zhu L (2021) Optimal subsampling for linear quantile regression models. Can J Stat 49(4):1039–1057

He S, Yan X (2022) Functional principal subspace sampling for large scale functional data analysis. Electron J Stat 16(1):2621–2682

Hjort NL, Pollard D (2011) Asymptotics for minimisers of convex processes. arXiv preprint arXiv:1107.3806

Homrighausen D, McDonald DJ (2019) Compressed and penalized linear regression. J Comput Graph Stat 29:309–322

Kato K (2012) Estimation in functional linear quantile regression. Ann Stat 40(6):3108–3136

Kinoshita H, Türkan H, Vucinic S et al (2020) Carbon monoxide poisoning. Toxicol Rep 7:169–173

Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46(1):33–50

Liu C, Yin P, Chen R et al (2018) Ambient carbon monoxide and cardio-vascular mortality: a nationwide time-series analysis in 272 cities in China. Lancet Planet Health 2(1):e12–e18

Liu H, You J, Cao J (2021) Functional L-optimality subsampling for massive data. arXiv preprint arXiv:2104.03446

Ma P, Mahoney MW, Yu B (2015) A statistical perspective on algorithmic leveraging. J Mach Learn Res 16(27):861–911

Mahoney MW (2011) Randomized algorithms for matrices and data. Found Trends Mach Learn 3:123–224

Ma P, Zhang X, Xing X, Ma J, Mahoney MW (2020) Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, pp 1026–1035

Moazami S, Noori R, Amiri BJ et al (2016) Reliable prediction of carbon monoxide using developed support vector machine. Atmos Pollut Res 7(3):412–418

Raskutti G, Mahoney MW (2016) A statistical perspective on randomized sketching for ordinary least-squares. J Mach Learn Res 17(213):1–31
Reiss P, Huang L (2012) Smoothness selection for penalized quantile regression splines. Int J Biostat. https://doi.org/10.1515/1557-4679.1381
Ruppert D (2002) Selecting the number of knots for penalized splines. J Comput Graph Stat 11(4):735–757
Sang P, Cao J (2020) Functional single-index quantile regression models. Stat Comput 30(4):771–781
Shams R, Jahani A, Moeinaddini M, Khorasani N (2020) Air carbon monoxide forecasting using an artificial neural network in comparison with multiple regression. Model Earth Syst Environ 6:1467–1475
Shao Y, Wang L (2021) Optimal subsampling for composite quantile regression model in massive data. Stat Pap 63:1139–1161
Shao L, Song S, Zhou Y (2022) Optimal subsampling for large-sample quantile regression with massive data. Can J Stat. https://doi.org/10.1002/cjs.11697
Stone CJ (1985) Additive regression and other nonparametric models. Ann Stat 13(2):689–705
Wang H (2019) More efficient estimation for logistic regression with optimal subsamples. J Mach Learn Res 20(132):1–59
Wang H, Ma Y (2021) Optimal subsampling for quantile regression in big data. Biometrika 108(1):99–112
Wang H, Zhu R, Ma P (2018) Optimal subsampling for large sample logistic regression. J Am Stat Assoc 113(522):829–844
Wang S, Gittens A, Mahoney MW (2018) Sketched ridge regression: optimization perspective, statistical perspective, and model averaging. J Mach Learn Res 18(218):1–50
Yao Y, Wang H (2019) Optimal subsampling for softmax regression. Stat Pap 60(2):585–599
Yoshida T (2013) Asymptotics for penalized spline estimators in quantile regression. Commun Stat Theory M. https://doi.org/10.1080/03610926.2013.765477
Yu J, Wang H, Ai M, Zhang H (2020) Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. J Am Stat Assoc 117(537):265–276
Yuan M (2006) GACV for quantile smoothing splines. Comput Stat Data Ann 50(3):813–829
Yuan X, Li Y, Dong X, Liu T (2022) Optimal subsampling for composite quantile regression in big data. Stat Pap 63:1649–1676
Zhou S, Shen X, Wolfe D (1998) Local asymptotics for regression splines and confidence regions. Ann Stat 26(25):1760–1782