# Distributed High-Dimensional Quantile RegressionEstimation Efficiency and Support Recovery

Caixing Wang, Ziliang Shen

Shanghai University of Finance and Economics, School of Statistics and Management

Joint work with Ziliang Shen

# Outline

1. Background

2. Method

3. Theory

4. Experiment

# Outline

1. Background
2. Method
3. Theory
4. Experiment

# Background

Why quantile regression?

- **Reason 1:** Quantile regression allows us to study the impact of predictors on different quantiles of the response distribution, and thus provides **a complete picture** of the relationship between responses and covariates.
- **Reason 2: Robust** to outliers in response observations.
- **Reason 3:** Estimation and inference are **distribution-free**, and **heterogeneity** is usually allowed in quantile regression models.

# Motivation

Although quantile regression can better handle data heterogeneity, computational challenges arise when both sample size and dimension are large due to the **non-smooth** check loss function. Consequently, it is natural to consider a distributed estimation procedure to address scalability concerns.

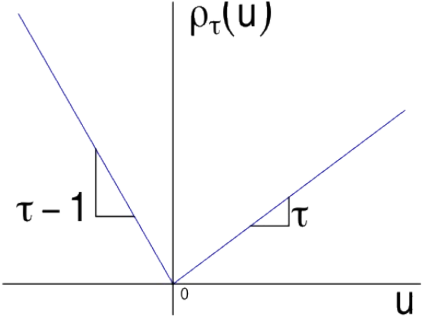

Figure: The check loss function.

# Contribution

- Methodology novelty. We transform the original quantile regression into the least-squares optimization. By applying a **double-smoothing** approach, we extend a previous Newton-type distributed approach without the restrictive independent assumption between the error term and covariates. An efficient algorithm is developed, which enjoys high computation and communication efficiency.
- Theoretical assessments. We prove that the proposed distributed estimator achieves a **near-oracle convergence rate and high support recovery accuracy** after a constant number of iterations.
- Numerical verification. Another contribution of this work is the comprehensive studies on the validity and effectiveness of the proposed algorithm in various synthetic and real-life examples, which further support the theoretical findings in this paper.

# Outline

1. Background
2. Method
3. Theory
4. Experiment

# The Linear Quantile Model

For a given quantile level $\tau \in (0,1)$, we consider to construct the conditional $\tau$-th quantile function $Q_\tau(Y|\boldsymbol{X})$ with a linear model $Q_\tau(Y|\boldsymbol{X}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*(\tau) = \sum_{j=0}^{p} x_j \beta_j^*(\tau)$, where $Y \in \mathbb{R}$ is the response and $\boldsymbol{X} = (x_0, x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^{p+1}$ is $p+1$-dimensional covariate vector with $x_0 \equiv 1$. Here, $\boldsymbol{\beta}^* = \boldsymbol{\beta}^*(\tau) = (\beta_0^*(\tau), \beta_1^*(\tau), \ldots, \beta_p^*(\tau))$ is the true coefficient that can be obtained by

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\ \mathcal{Q}(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\ \mathbb{E}\left[\rho_\tau(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\right],$$

where $\rho_\tau(u) = u\{\tau - I(u \le 0)\}$ is the standard check loss function and we denote the error term as $\varepsilon = Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}^*$.

# Newton-type Transformation on Quantile Regression

Given an initial estimator $\boldsymbol{\beta}_0$, the population form of the Newton-Raphson iteration is

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \boldsymbol{H}^{-1}(\boldsymbol{\beta}_0)\mathbb{E}\left[\partial \mathcal{Q}(\boldsymbol{\beta}_0)\right], \tag{1}$$

where $\partial \mathcal{Q}(\boldsymbol{\beta}) = \boldsymbol{X}\{I(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} \leq 0) - \tau\}$ is the subgradient of the check loss function with respect to $\boldsymbol{\beta}$, and

$$\boldsymbol{H}(\boldsymbol{\beta}) = \partial \mathbb{E}[\partial \mathcal{Q}(\boldsymbol{\beta})]/\partial \boldsymbol{\beta} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} f_{\varepsilon|\boldsymbol{X}}(\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*))]$$

denotes the population Hessian matrix of $\mathbb{E}[\mathcal{Q}(\boldsymbol{\beta})]$. Here, $f_{\varepsilon|\boldsymbol{X}}(\cdot)$ is the conditional density of $\varepsilon$ given $\boldsymbol{X}$.

When the initial estimator $\boldsymbol{\beta}_0$ is close to the true parameter $\boldsymbol{\beta}^*$, $\boldsymbol{H}(\boldsymbol{\beta}_0)$ will be close to $\boldsymbol{H}(\boldsymbol{\beta}^*) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} f_{\varepsilon|\boldsymbol{X}}(0)]$. Motivated by this, we further approximate $\boldsymbol{H}(\boldsymbol{\beta}^*)$ with $\boldsymbol{D}_h(\boldsymbol{\beta}_0)$ such that

$$\boldsymbol{H}(\boldsymbol{\beta}_0) \approx \boldsymbol{H}(\boldsymbol{\beta}^*) \approx \boldsymbol{D}_h(\boldsymbol{\beta}_0) = \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} K_h(e_0)),$$

where $e_0 = Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0$, and $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ denoting a symmetrix and non-negative kernel function, $h \to 0$ is the bandwidth.

# Newton-type Transformation on Quantile Regression

Denote a **pseudo covariate** as $\widetilde{\boldsymbol{X}}_h = \sqrt{K_h(e_0)}\boldsymbol{X}$, and replace $\boldsymbol{H}(\boldsymbol{\beta}_0)$ with $\boldsymbol{D}_h(\boldsymbol{\beta}_0)$ in (1) leads to the following iteration,

$$\boldsymbol{\beta}_1 = \boldsymbol{D}_h^{-1}(\boldsymbol{\beta}_0)\mathbb{E}\left\{\widetilde{\boldsymbol{X}}_h\left[\widetilde{\boldsymbol{X}}_h^{\mathrm{T}}\boldsymbol{\beta}_0 - \frac{1}{\sqrt{K_h(e_0)}}\left(I(e_0 \leq 0) - \tau\right)\right]\right\}.$$

If we further define a new **pseudo response** as

$$\widetilde{Y}_h = \widetilde{\boldsymbol{X}}_h^{\mathrm{T}}\boldsymbol{\beta}_0 - \frac{1}{\sqrt{K_h(e_0)}}\left(I(e_0 \leq 0) - \tau\right),$$

then $\boldsymbol{\beta}_1 = \boldsymbol{D}_h^{-1}(\boldsymbol{\beta}_0)\mathbb{E}(\widetilde{\boldsymbol{X}}_h\widetilde{Y}_h) = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\mathbb{E}(\widetilde{Y}_h - \widetilde{\boldsymbol{X}}_h^{\mathrm{T}}\boldsymbol{\beta})^2$ is the least squares regression coefficient of $\widetilde{Y}_h$ on $\widetilde{\boldsymbol{X}}_h$. To further encourage the sparsity of the coefficient vector, we consider the $\ell_1$-penalized least squares problem as

$$\boldsymbol{\beta}_{1,\ell_1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\ \frac{1}{2}\mathbb{E}\left(\widetilde{Y}_h - \widetilde{\boldsymbol{X}}_h^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + \lambda|\boldsymbol{\beta}|_1.$$

# Distributed Estimation with a Double-smoothing Shifted Loss Function

Let $\mathscr{Z}^N = \{(\boldsymbol{X}_i, Y_i)\}_{i=1}^N$ are randomly and evenly stored in $m$ machines $\mathcal{M}_1, \ldots, \mathcal{M}_m$ with the sample size $n$. Let $\widehat{\boldsymbol{D}}_{k,h} = \frac{1}{n}\sum_{i\in\mathcal{M}_k} \widetilde{\boldsymbol{X}}_{i,h}\widetilde{\boldsymbol{X}}_{i,h}^{\mathrm{T}}$ and $\widehat{\boldsymbol{D}}_h = \frac{1}{m}\sum_{k=1}^m \widehat{\boldsymbol{D}}_{k,h} = \frac{1}{N}\sum_{i=1}^N \widetilde{\boldsymbol{X}}_{i,h}\widetilde{\boldsymbol{X}}_{i,h}^{\mathrm{T}}$ as the $k$-th local and total sample covariance matrix.

The pseudo local and global loss functions are

$$\mathcal{L}_k(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i\in\mathcal{M}_k}(\widetilde{Y}_{i,h} - \widetilde{\boldsymbol{X}}_{i,h}^{\mathrm{T}}\boldsymbol{\beta})^2 \text{ and } L_N(\boldsymbol{\beta}) = \frac{1}{2N}\sum_{i=1}^N(\widetilde{Y}_{i,h} - \widetilde{\boldsymbol{X}}_{i,h}^{\mathrm{T}}\boldsymbol{\beta})^2.$$

According to the Taylor expansion of $\mathcal{L}_N(\boldsymbol{\beta})$ around $\widehat{\boldsymbol{\beta}}_0$, we have

$$\mathcal{L}_N(\boldsymbol{\beta}, \widehat{\boldsymbol{D}}_h(\boldsymbol{\beta}) = \mathcal{L}_N(\widehat{\boldsymbol{\beta}}_0) + \{\partial\mathcal{L}_N(\widehat{\boldsymbol{\beta}}_0)\}^{\mathrm{T}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_0) + \frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_0)^{\mathrm{T}}\widehat{\boldsymbol{D}}_h(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_0). \tag{2}$$

# Distributed Estimation with a Double-smoothing Shifted Loss Function

To save the communication cost, we replace the global Hessian $\widehat{\boldsymbol{D}}_h$ with the local Hessian $\widehat{\boldsymbol{D}}_{1,b}$. Here, $h$ and $b$ denote the *global bandwidth* and *local bandwidth*. Thus we can rewrite (2) as

$$\mathcal{L}_N(\boldsymbol{\beta}, \widehat{\boldsymbol{D}}_h) = \underbrace{\mathcal{L}_N(\boldsymbol{\beta}, \widehat{\boldsymbol{D}}_{1,b})}_{(i)\text{ Shifted loss}} + \underbrace{\mathcal{O}_{\mathbb{P}}\left\{\|\widehat{\boldsymbol{D}}_h - \widehat{\boldsymbol{D}}_{1,b}\|_{op} \cdot |\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_0|_2^2\right\}}_{(ii)\text{ Approximation error}}. \tag{3}$$

With proper local bandwidth $b$ andthe global bandwidth $h$, we can prove $\|\widehat{\boldsymbol{D}}_h - \widehat{\boldsymbol{D}}_{1,b}\|_{op} = o_{\mathbb{P}}(1)$. Remove the terms that are independent of $\boldsymbol{\beta}$ in (i) and the negligible approximation error (ii) in (3), and add the lasso penalty, the distributed estimation can be simplified to

$$\widehat{\boldsymbol{\beta}}_{1,h} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\ \frac{1}{2n}\sum_{i\in\mathcal{M}_1}(\widetilde{\boldsymbol{X}}_{i,b}^{\mathrm{T}}\boldsymbol{\beta})^2 - \boldsymbol{\beta}^{\mathrm{T}}\left\{\boldsymbol{z}_N + (\widehat{\boldsymbol{D}}_{1,b} - \widehat{\boldsymbol{D}}_h)\widehat{\boldsymbol{\beta}}_0\right\} + \lambda_N|\boldsymbol{\beta}|_1.$$

# Distributed Estimation with a Double-smoothing Shifted Loss Function

Given $\widehat{\boldsymbol{\beta}}_{1,h}$ as the estimator from the first iteration, we can similarly construct an iterative distributed estimation procedure,

$$\widehat{\boldsymbol{\beta}}_{t,h} = \mathop{\arg\min}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \frac{1}{2n} \sum_{i\in\mathcal{M}_1} \left( (\widetilde{\boldsymbol{X}}_{i,b}^{(t)})^{\mathrm{T}} \boldsymbol{\beta} \right)^2 - \boldsymbol{\beta}^{\mathrm{T}} \left\{ \boldsymbol{z}_N^{(t)} + \left( \widehat{\boldsymbol{D}}_{1,b}^{(t)} - \widehat{\boldsymbol{D}}_h^{(t)} \right) \widehat{\boldsymbol{\beta}}_{t-1,h} \right\} + \lambda_{N,t} |\boldsymbol{\beta}|_1. \tag{4}$$

In addition, we introduce the pooled estimator:

$$\widehat{\boldsymbol{\beta}}_{pool} = \mathop{\mathrm{argmin}}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \frac{1}{2N} \sum_{i=1}^{N} (\widetilde{Y}_{i,h} - \widetilde{\boldsymbol{X}}_{i,h}^{\mathrm{T}} \boldsymbol{\beta})^2 + \lambda |\boldsymbol{\beta}|_1. \tag{5}$$

# Outline

1. Background
2. Method
3. Theory
4. Experiment

# Statistical Guarantees: Convergence rate

**Theorem 1 (Convergence rate).** *Suppose that the initial estimator satisfies that* $|\widehat{\boldsymbol{\beta}}_{0,h} - \boldsymbol{\beta}^*|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{s \log N/n})$ *and let* $h \asymp (s \log N/N)^{1/3}$, $b \asymp (s \log n/n)^{1/3}$. *Define* $\kappa(a,b,g) = \max\left\{ s^a \left(\frac{\log n}{n}\right)^{\frac{2g+3}{6}}, s^b \left(\frac{\log N}{n}\right)^{\frac{g+1}{2}} \right\}$, *for* $1 \leq g \leq t$, *take*

$$\lambda_{N,g} = C\left(\sqrt{\frac{\log N}{N}} + \kappa(\frac{5g}{6}, g, g)\right),$$

*where* $C$ *is a sufficiently large constant. Then under Assumptions 1-6, we have*

$$\left|\widehat{\boldsymbol{\beta}}_{t,h} - \boldsymbol{\beta}^*\right|_2 = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{s \log N}{N}} + \kappa(\frac{5t+3}{6}, \frac{2t+1}{2}, t)\right). \qquad (6)$$

# Statistical Guarantees: Support recovery

**Theorem 2 (Support recovery).** *Let* $\widehat{S}_t = \{j : \widehat{\beta}^p_{t,h} \neq 0, j \in \mathbb{N}_+\}$ *be the support of* $\widehat{\boldsymbol{\beta}}_{t,h}$*, where* $t \geq 1$*. Under the same conditions of Theorem 1, we have* $\mathbb{P}(\widehat{S}_t \subseteq S) \to 1$*. Furthermore, if there exists a sufficiently large constant* $C > 0$ *such that*

$$\min_{j \in S} \left|\beta^*_j\right| \geq C \left\|\boldsymbol{I}^{-1}_{S \times S}\right\|_{\infty} \left( \sqrt{\frac{\log N}{N}} + \kappa(\frac{5t}{6}, t; t) \right).$$

*Then we have* $\mathbb{P}(\widehat{S}_t = S) \to 1$*.*

# Statistical Guarantees

**When $t \geq t_0$ for some $t_0$, the estimator can achieve global convergence rate $\mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{s \log N}{N}}\right)$ and beta-min condition**

$$\min_{j \in S} \left|\beta_j^*\right| \geq C \left\|\boldsymbol{I}_{S \times S}^{-1}\right\|_\infty \left(\sqrt{\frac{\log N}{N}}\right),$$

**here** $logN \asymp \log \max(p, N)$**).**

# Contrastive Method

We present four contrasting methods with our proposed DHSQR method:

(a) Averaged DC (*Avg-DC*) estimator which computes the $\ell_1$-penalized QR estimators on the local machine and then combines the local estimators by taking the average;

(b) The distributed high-dimensional sparse quantile regression estimator on a single machine with pooled data defined in (5), which is denoted by *Pooled DHSQR*;

(c) Distributed robust estimator with Lasso (*DREL*), see in Chen et al. 2020;

(d) Distributed penalty quantile regression estimator (*DPQR*) with convolution smoothing, see in Jiang and Yu 2021; Tan, Battey, and Zhou 2022.

# Theoretical Comparison

Furthermore, we wish to reiterate the theoretical innovations of our method, DHSQR. Unlike the DREL method, we relax the homoscedasticity assumption of the error term. In contrast to the Avg-DC and DPQR methods, we provide theoretical guarantees to support recovery.

Table: Comparison of theoretical properties of different methods.

| Method | Statistical convergence | Support recovery | Heterogeneity |
|---|---|---|---|
| DHSQR | ✓ | ✓ | ✓ |
| DREL | ✓ | ✓ | × |
| DPQR | ✓ | × | ✓ |
| Avg-DC | ✓ | × | × |

# Outline

1. Background

2. Method

3. Theory

4. Experiment

# Simuation Setting

We generate synthetic data from the following linear models, corresponding to the homoscedastic error case (model 1) and the heteroscedastic error case (model 2):

- **Model 1 (homoscedastic error)**: $Y_i = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}^* + \varepsilon_i$
- **Model 2 (heteroscedastic error)**: $Y_i = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}^* + (1 + 0.4 x_{i1}) \varepsilon_i$,

where $\boldsymbol{X}_i = (1, x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is a $(p+1)$-dimensional vector and $(x_{i1}, \ldots, x_{ip})$ is drawn from a multivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma}_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le p$, the true parameter $\boldsymbol{\beta}^* = (1, 1, 2, 3, 4, 5, \mathbf{0}_{p-5})^{\mathrm{T}}$. We fix the dimension $p = 500$ and consider values of $\tau = 0.5$. We consider the following three noise distributions:

1. Normal distribution: the noise $\varepsilon_i \sim N(0, 1)$;
2. $t_3$ distribution: the noise $\varepsilon_i \sim t(3)$;
3. Cauchy distribution: the noise $\varepsilon_i \sim \mathit{Cauchy}(0, 1)$.

# Effect of the number of iterations Under Heavy-Tailed Noise

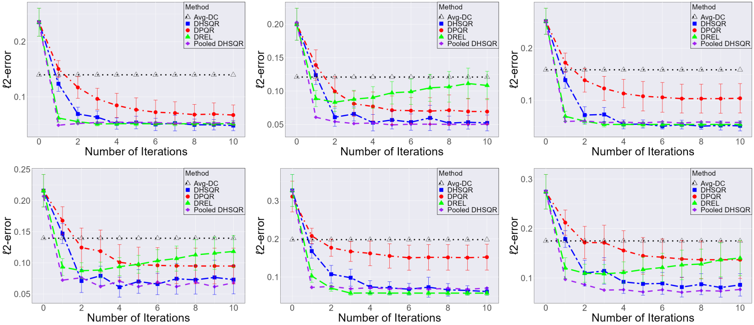

Figure: The $\ell_2$-error with an error bound between the true parameter and the estimated parameter versus the number of iterations with a fixed quantile level $\tau = 0.5$. In the left panel, from top to bottom represent noise distributions that are Normal, $t_3$, and Cauchy distribution for the homoscedastic error case, respectively. In the right panel, from top to bottom represent noise distributions as Normal, $t_3$, and Cauchy distribution for the heteroscedastic error case, respectively.

# Effect of Total Sample Size and Local Sample Size

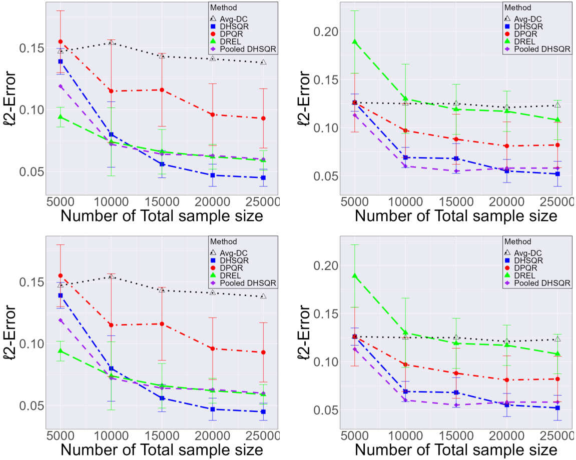

Figure: The $\ell_2$-error from the true parameter versus the number of total and local sample size with a fixed quantile level $\tau = 0.5$. In the top panel, from left to right show the effect of different total sample sizes $N$ for the homoscedastic and heteroscedastic error cases, respectively. In the bottom panel, from left to right show the effect of different local sample sizes $n$ for the homoscedastic and heteroscedastic error cases, respectively.

# Effect of Total Sample Size and Local Sample Size

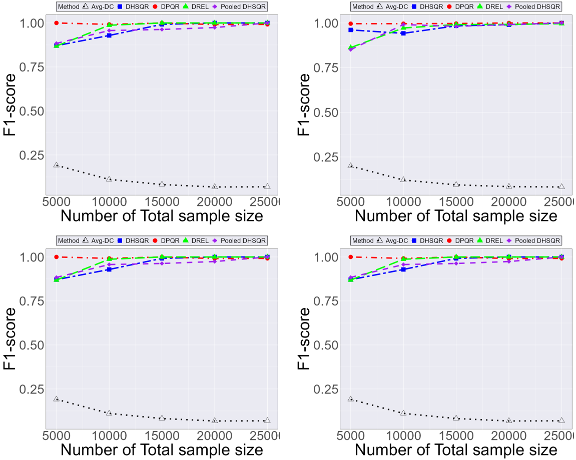

Figure: The $F_1$ score from the true parameter versus the number of total and local sample size with a fixed quantile level $\tau = 0.5$. In the top panel, from left to right show the effect of different total sample sizes $N$ for the homoscedastic and heteroscedastic error cases, respectively. In the bottom panel, from left to right show the effect of different local sample sizes $n$ for the homoscedastic and heteroscedastic error cases, respectively.

# Computation Time Comparison

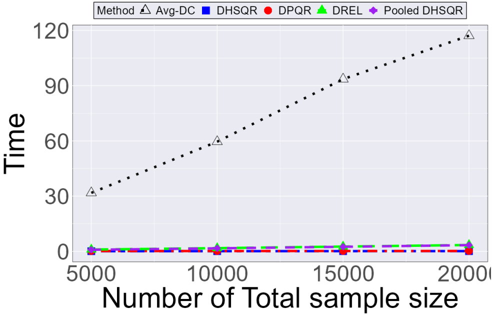

Figure: The Time computation of five estimators under different total sample sizes $N$.

In the figure, DHSQR outperforms others with the fastest single iteration time, followed by DPQP. Pooled DHSQR and DREL show comparable speeds, while Avg-DC is the slowest.

# References I

[1] Chen, X., Liu, W., Mao, X., and Yang, Z. Distributed high- dimensional regression under a quantile loss function. Journal of Machine Learning Research, 21(1):7432–7474, 2020.

[2] Jiang, R. and Yu, K. Smoothing quantile regression for a distributed system. Neurocomputing, 466:311–326, 2021.

[3] Tan, K. M., Battey, H., and Zhou, W.-X. Communication- constrained distributed quantile regression with optimal statistical guarantees. Journal of Machine Learning Re- search, 23:1–61, 2022.

Thank you for listening!