Quantile Regression for Extraordinarily Large Data

Shih-Kang Chao
Department of Statistics
Purdue University

November, 2016
A joint work with Stanislav Volgushev and Guang Cheng
With the advance of technology, it is increasingly common that data set is so large that it cannot be stored in a single machine

- Social media (views, likes, comments, images...)
- Meteorological and environmental surveillance
- Transactions in e-commerce
- Others...

Figure: A server room in Council Bluffs, Iowa. Photo: Google/Connie Zhou.
To take advantage of the opportunities in massive data, we need to deal with storing (disc) and computational (memory, processor) bottlenecks.

Divide and conquer paradigm: Randomly divide $N$ samples into $S$ groups. $n = N/S$
• Aggregate solutions from subproblems to get a solution for original problem
• Can be implemented by computational platforms such as Hadoop (White, 2012)
• Data that are stored in distributed locations can be analyzed in similar manner
Does D&C fit for statistical analysis?

Sometimes it does, but sometimes it doesn’t...

In the following, we give two simple examples.
Example 1: sample mean

\[ \frac{1}{4} \sum_{s=1}^{4} X_s = \frac{1}{4n} \sum_{s=1}^{4} \sum_{i=1}^{n} X_{is} = \frac{1}{N} \sum_{i=1}^{N} X_i = \overline{X}_N. \]

It fits!
Example 2: sample median

\[ X_{(0.5)}^s = \text{the middle value of ordered } n \text{ samples in } s \text{ group}; \]

\[ X_{(0.5)} = \text{overall median} \]

\[ \frac{1}{4} \sum_{s=1}^{4} X_{(0.5)}^s = X_{(0.5)} \]
Example 2: sample median

Simulation 1: $X_i \sim N(0, 1)$; Simulation 2: $X_i \sim \text{Exponential}(1)$. $N = 2^{15}$. True median v.s. simulated $S^{-1} \sum_{s=1}^{S} X_s^{(0.5)}$
Example 2: sample median

Simulation 1: $X_i \sim N(0, 1)$; Simulation 2: $X_i \sim \text{Exponential}(1)$. $N = 2^{15}$. True median v.s. simulated $S^{-1} \sum_{s=1}^{S} X_s^{(0.5)}$
Major issues

- When does the D&C algorithm work?
  - Especially for skewed and heavy tail distribution
- Statistical inference
  - Asymptotic distribution
- Inference for the ”whole” distribution?
  - Take advantage of massive size to discover subtle patterns hidden in the ”whole” distribution
Outline

1. Quantile regression
2. Two-step algorithm: D&C and quantile projection
3. Oracle rules: linear model and nonparametric model
4. Simulation
Response $Y$, predictors $X$. For $\tau \in (0, 1)$, conditional quantile curve $Q(\cdot; \tau)$ of $Y \in \mathbb{R}$ conditional on $X$ is defined through

$$P(Y \leq Q(X; \tau) | X = x) = \tau \quad \forall x.$$ 

$Q(x; \tau)$ at $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$ under different models.
Quantile regression v.s. mean regression

Mean regression:
\[ Y_i = m(X_i) + \varepsilon_i, \mathbb{E}[\varepsilon|X = x] = 0 \]
- \( m \): Regression function, object of interest.
- \( \varepsilon_i \): 'errors'.

Quantile regression:
\[ P(Y \leq Q(x; \tau)|X = x) = \tau \]
- No strict distinction between 'signal' and 'noise'.
- Object of interest: properties of conditional distribution of \( Y|X = x \).
- Contains much richer information than just conditional mean.
Quantile curves v.s. conditional distribution

\[ Q(x_0; \tau) = F^{-1}(\tau|x_0), \]

where \( F_{Y|X}(y|x) \) is the conditional distribution function of \( Y \) given \( X \).
Quantile Regression as Optimization

\{(X_i, Y_i)\}_{i=1}^{N} \text{ independent and identical samples in } \mathbb{R}^{d+1}

Koenker and Bassett (1978): if \( Q(x; \tau) = \beta(\tau) \top x \), estimate by

\[ \hat{\beta}_{or}(\tau) := \text{arg min}_{\theta} \sum_{i=1}^{N} \rho_{\tau}(Y_i - \theta \top X_i) \] (1.1)

where \( \rho_{\tau}(u) := \tau u^{+} + (1 - \tau) u^{-} \) ’check function’.

- Optimization problem (1.2) is convex (but non-smooth), which can be solved by linear programming
- \( \hat{Q}(x_0; \tau) := x_0 \top \hat{\beta}_{or}(\tau) \) for any \( x_0 \)
- More generally, we can consider a series approximation model
A unified framework:

\[ Q(x; \tau) \approx Z(x)\top \beta(\tau) \]

\[ m := \text{dim}(Z) \text{ (it is possible that } m \to \infty) \]. Solve

\[ \hat{\beta}_{or}(\tau) := \arg \min_{\mathbf{b}} \sum_{i=1}^{N} \rho_{\tau}\{Y_i - \mathbf{b}^\top Z(X_i)\} \] (1.2)

- Examples of \( Z(x) \): linear model with fixed/increasing dimension, B-splines, polynomials, trigonometric polynomials
- \( \hat{Q}(x; \tau) := Z(x)\top \hat{\beta}_{or}(\tau) \)
- Need to control the ”bias” \( Q(x; \tau) - Z(x)\top \beta(\tau) \)
\( \hat{\beta}_{or}(\tau) \) is computationally infeasible when \( N \) is so large that cannot be handled with a single machine

To infer the "whole" conditional distribution for fixed \( x_0 \), by

\[
F_{Y|X}(y|x_0) = \int_0^1 1\{Q(x_0; \tau) < y\} d\tau
\]

where \( 1(A) = 1 \) if \( A \) is true. To approximate the integral, we need a to compute \( Q(x_0; \tau) \) for many \( \tau \)
D&C algorithm at fixed $\tau$

\[ \hat{\beta}^s(\tau) := \arg \min_{b \in \mathbb{R}^m} \sum_{i=1}^{n} \rho_\tau \{ Y_{is} - b^\top Z(X_{is}) \} \quad (2.1) \]

\[ \overline{\beta}(\tau) := \frac{1}{S} \sum_{s=1}^{S} \hat{\beta}^s(\tau). \quad (2.2) \]

However, this is only for a fixed $\tau$!
Quantile projection algorithm

To avoid repetitively applying D&C, take $\mathbf{B} := (B_1, \ldots, B_q)^\top$

B-spline basis defined on equidistant knots $\{t_1, \ldots, t_G\} \subset \mathcal{T}$ with
degree $r_\tau \in \mathbb{N}$,

$$\hat{\beta}(\tau) := \hat{\Xi}^\top \mathbf{B}(\tau).$$  \hspace{1cm} (2.3)

Computation of $\hat{\Xi}$:

1. Define a grid of quantile levels $\{\tau_1, \ldots, \tau_K\}$ on $[\tau_L, \tau_U]$, $K > q$. For each $\tau_k$, compute $\bar{\beta}(\tau_k)$
2. Compute for $j = 1, \ldots, m$

$$\hat{\alpha}_j := \arg \min_{\alpha \in \mathbb{R}^q} \sum_{k=1}^K (\bar{\beta}_j(\tau_k) - \alpha^\top \mathbf{B}(\tau_k))^2.$$ \hspace{1cm} (2.4)

3. Set the matrix $\hat{\Xi} := [\hat{\alpha}_1 \ \hat{\alpha}_2 \ldots \hat{\alpha}_m]$. 
Define $\hat{Q}(x; \tau) := Z(x)^T \hat{\beta}(\tau) = Z(x)^T \Xi^T B(\tau)$.

Compute

$$\hat{F}_Y|_X(y|x_0) := \tau_L + \int_{\tau_L}^{\tau_U} 1\{\hat{Q}(x_0; \tau) < y\} d\tau. \tag{2.5}$$

where $0 < \tau_L < \tau_U < 1$. 
Both $S$ and $K$ can grow with $N$, and they decide the computational limit as well as the statistical performance.

- Find $S$ and $K$ such that $\beta(\tau)$ and $\hat{\beta}(\tau)$ are ”close” to $\hat{\beta}_{or}(\tau)$ in some statistical sense.
- This results in ”sharp” upper bound of $S$ and lower bound of $K$ which impose computational limit.

The two-step procedure requires only one pass through the entire data.

Quantile projection requires only $\{\beta(\tau_1), ... , \beta(\tau_K)\}$ of size $m \times K$, without the need to access the raw data set.
Oracle rules

For $\mathcal{T} = [\tau_L, \tau_U] \subset (0, 1)$, under regularity conditions, C., Volgushev and Cheng (2016):

1. \[ a_N u^\top (\hat{\beta}_{or}(\tau) - \beta(\tau)) \rightsquigarrow \mathcal{N} \text{ at any fixed } \tau \in \mathcal{T} \quad (3.1) \]
2. \[ a_N u^\top (\hat{\beta}_{or}(\tau) - \beta(\tau)) \rightsquigarrow \mathcal{G}(\tau) \text{ as a process in } \tau \in \mathcal{T} \quad (3.2) \]

where $\mathcal{N}$ is a centered normal distribution and $\mathcal{G}$ is a centered Gaussian process.

- The **oracle rule** holds for $\overline{\beta}(\tau)$ and $\hat{\beta}(\tau)$ if $\overline{\beta}(\tau)$ satisfies (3.1) and $\hat{\beta}(\tau)$ satisfies (3.2).
- Inference follows from the oracle rule.
Two leading models

**Linear model:** $m = \dim(Z(x))$ is fixed, and $Q(x; \tau) = Z(x)\top \beta(\tau);$

**Univariate spline nonparametric model:** $m = \dim(Z(x)) \to \infty$ with $N. c_N(\gamma_N) := \left| Q(x; \tau) - Z(x)\top \gamma_N(\tau) \right| \neq 0,$

$$
\gamma_N(\tau) := \arg \min_{\gamma \in \mathbb{R}^m} \mathbb{E}[(Z\top \gamma - Q(X; \tau))^2 f(Q(X; \tau) | X)]. \quad (3.3)
$$

- $\gamma_N$ can be defined in many other ways. We do not go into detail here.
- Conditions imposed throughout the talk:
- $J_m(\tau) := \mathbb{E}[ZZ\top f_Y|X(Q(X; \tau)|X)]$
**Linear model with fixed dimension**

\[ \mathcal{P}_1(\xi_m, M, \bar{f}, \bar{f}', f_{\min}): \text{all distributions satisfying (A1)-(A3)} \]

with some constants \(0 < \xi_m, M, \bar{f}, \bar{f}' < \infty\) and \(f_{\min} > 0\)

---

**Theorem 3.1 ( Oracle Rule for } \bar{\beta}(\tau))

- Suppose that \(S = o(N^{1/2}(\log N)^{-2})\). \(\forall \tau \in \mathcal{T} \) and \(u \in \mathbb{R}^m\),

\[
\frac{\sqrt{N}u^\top (\bar{\beta}(\tau) - \beta(\tau))}{(u^\top J_m(\tau)^{-1}E[Z(X)Z(X)^\top]J_m(\tau)^{-1}u)^{1/2}} \Rightarrow N(0, \tau(1-\tau)),
\]

- If \(S \gtrsim N^{1/2}\), then the weak convergence result above fails for some \(P \in \mathcal{P}_1(1, d, \bar{f}, \bar{f}', f_{\min})\).

\(S = o(N^{1/2}(\log N)^{-2})\) is sharp: only miss by a factor of \((\log N)^{-2}\)
\( \Lambda_{c}^{\eta_{\tau}}(\mathcal{T}) \): Hölder class with smoothness \( \eta_{\tau} > 0 \).

\( \ell^\infty(\mathcal{T}) \): set of all uniformly bounded real functions on \( \mathcal{T} \)

**Theorem 3.2 (Oracle Rule for \( \hat{\beta}(\cdot) \))**

Suppose that \( \tau \mapsto Q(x_0; \tau) \in \Lambda_{c}^{\eta_{\tau}}(\mathcal{T}) \) for a \( x_0 \in \mathcal{X} \).

- If \( S = o(N^{1/2}(\log N)^{-2}) \), and \( K \gg G \gg N^{1/(2\eta_{\tau})} \) with \( r_{\tau} \geq \eta_{\tau} \) then the projection estimator \( \hat{\beta}(\tau) \) defined in (2.3) satisfies

\[
\sqrt{N}(Z(x_0)^{\top}\hat{\beta}(\cdot) - Q(x_0; \cdot)) \rightsquigarrow G(\cdot) \text{ in } \ell^\infty(\mathcal{T}), \quad (3.4)
\]

where \( G \) is a centered Gaussian process.

- If \( G \sim N^{1/(2\eta_{\tau})} \) the weak convergence in (3.4) fails for some \( P \in \mathcal{P}_1(1, d, \bar{f}, \bar{f}^{\top}, f_{\min}) \) with \( \tau \mapsto Q(x_0; \tau) \in \Lambda_{c}^{\eta_{\tau}}(\mathcal{T}) \).

\( K \gg N^{1/(2\eta_{\tau})} \) is sharp!
We require an additional condition:

\[(L)\text{ For each } x \in \mathcal{X}, \text{ the vector } Z(x) \text{ has zeroes in all but at most } r \text{ consecutive entries, where } r \text{ is fixed. Furthermore, }\]

\[
\sup_{x \in \mathcal{X}} \tilde{\mathcal{E}}(Z(x), \gamma_N) = O(1).
\]

which guarantees that the matrix \(J_m(\tau)\) to be a block matrix

\[
\mathcal{P}_L(Z, M, \bar{f}, \bar{f}', f_{\min}, R) := \{ \text{all sequences of distributions of } (X, Y) \text{ on } \mathbb{R}^{d+1} \text{ satisfying } (A1)-(A3) \text{ with } M, \bar{f}, \bar{f}' < \infty, f_{\min} > 0, \text{ and } (L) \text{ for some } r < R; \]

\[
m^2(\log N)^6 = o(N), c_N(\gamma_N)^2 = o(N^{-1/2}) \}. \tag{3.5}
\]
Splines nonparametric model

**Theorem 3.3 (Oracle rule for $\overline{\beta}(\tau)$)**

Let $\{(X_i, Y_i)\}_{i=1}^N$ be distributed $P_N \in \mathcal{P}_L(Z, M, \bar{f}, \bar{f}', f_{\text{min}}, R)$, $m \gg N^\zeta$ for some $\zeta > 0$.

- $S = o((Nm^{-1}(\log N)^{-4})^{1/2})$ then $\forall \tau \in \mathcal{T}, x_0 \in \mathcal{X}$,

$$
\frac{\sqrt{N}Z(x_0)^\top (\overline{\beta}(\tau) - \gamma_N(\tau))}{(Z(x_0)^\top J_m(\tau)^{-1}E[ZZ^\top]J_m(\tau)^{-1}Z(x_0))^{1/2}} \xrightarrow{\text{w}} \mathcal{N}(0, \tau(1 - \tau))
$$

- If $S \gtrsim (N/m)^{1/2}$ the weak convergence above fails for some $P_N \in \mathcal{P}_L(Z, M, \bar{f}, \bar{f}', f_{\text{min}}, R)$

Sharpness of $S = o((N/m)^{1/2}(\log N)^{-2})$: nonparametric rate is slower, again miss by a factor of $(\log N)^{-2}$
Theorem 3.4 (Oracle rule for \( \hat{\beta}(\tau) \))

Let the assumptions of Theorem 3.3 hold and assume that 
\( \tau \mapsto Q(x_0; \tau) \in \Lambda_c^{\eta_\tau}(\mathcal{T}) \), \( r_\tau \geq \eta_\tau \),

- Suppose that \( c_N(\gamma_N) = o(N^{-1/2}\|Z(x_0)\|) \) and 
  \( K \gg G \gg N^{1/(2\eta_\tau)}\|Z(x_0)\|^{-1/\eta_\tau} \), and the limit
  \( H(\tau_1, \tau_2) := \lim_{N \to \infty} K_N(\tau_1, \tau_2) > 0 \quad \forall \tau_1, \tau_2 \in \mathcal{T} \).

\[
\frac{\sqrt{N}}{\|Z(x_0)\|} (Z(x_0) \trans \hat{\beta}(\cdot) - Q(x_0; \cdot)) \rightsquigarrow G(\cdot) \quad \text{in } \ell^\infty(\mathcal{T}),
\]

where \( G \) is a centered Gaussian process with covariance structure \( \mathbb{E}[G(\tau)G(\tau')] = H(\tau, \tau') \).

- If \( G \preceq N^{1/(2\eta_\tau)}\|Z(x_0)\|^{-1/\eta_\tau} \) the weak convergence fails for some \( \tau \mapsto Q(x_0; \tau) \in \Lambda_c^{\eta_\tau}(\mathcal{T}) \) for all \( S \).
Distribution function

\[ \hat{F}_{Y|X}^{or}(\cdot|x_0) := \tau_L + \int_{\tau_L}^{\tau_U} 1\{Z(x)^\top \hat{\beta}_{or}(\tau) < y\} d\tau \]

The oracle rule holds for both models.

**Corollary 3.5**

Under the same conditions as Theorem 3.2 (linear model) or Theorem 3.4 (nonparametric spline model), we have for any \( x_0 \in \mathcal{X} \),

\[ \sqrt{N}(\hat{F}_{Y|X}(\cdot|x_0) - F_{Y|X}(\cdot|x_0)) \sim - f_{Y|X}(\cdot|x_0) \mathbb{G}(F_{Y|X}(\cdot|x_0)), \]

in \( \ell^\infty((Q(x_0; \tau_L), Q(x_0; \tau_U))) \), where \( \mathbb{G} \) is a centered Gaussian process defined in respective theorem. The same holds for \( \hat{F}_{Y|X}^{or}(\cdot|x_0) \).
Phase transitions

Figure: Regions \((S, K)\) for the oracle rule of linear model and spline nonparametric model. "?" region is the discrepancy between the sufficient and necessary conditions.
Setting: linear model

- Model: \( Y_i = 0.21 + \beta_{m-1}^\top X_i + \varepsilon_i, \ m = 4, 16, 32 \)
- \( X_i \) follows a multivariate uniform distribution \( \mathcal{U}([0, 1]^{m-1}) \) with covariance matrix \( \Sigma_X := \mathbb{E}[X_iX_i^\top] \) with 
  \( \Sigma_{jk} = 0.1^20.7|j-k| \) for \( j, k = 1, \ldots, m - 1 \)
- \( \varepsilon \sim \mathcal{N} \) or \( \varepsilon \sim \text{EXP} \) (skewed)
- \( x_0 = (1, (m - 1)^{-1/2}1_{m-1}^\top) ^\top \)
- From Theorem 3.1, the coverage probability of the \( \alpha = 95\% \) confidence interval is 
  \[
P\{x_0^\top \beta(\tau) \in [x_0^\top \beta(\tau) \pm N^{-1/2} f_{\varepsilon,\tau}^{-1} \sqrt{\tau(1 - \tau)x_0^\top \Sigma_{X}^{-1} x_0 \Phi^{-1}(1 - \alpha/2)}] \}
\]
- \( S^* \): the point of \( S \) where the coverage starts to drop
$\beta(\tau), \varepsilon \sim \mathcal{N}(0, 0.1^2)$

Coverage Probability

Normal$(0, 0.1^2)$, $\tau = 0.1$

Coverage Probability

Normal$(0, 0.1^2)$, $\tau = 0.5$

Coverage Probability

Normal$(0, 0.1^2)$, $\tau = 0.9$
$\overline{\beta}(\tau), \varepsilon \sim \text{Exp}(0.8)$
Summary for simulation of $\bar{\beta}(\tau)$

- $m$ increases, $S^*$ decreases
- $N$ increases, $S^*$ gets close to $N^{1/2}$
- $\varepsilon \sim \mathcal{N}$, coverage is symmetric in $\tau$, $\varepsilon \sim \text{Exp}$, coverage is asymmetric in $\tau$
- $t$ distribution behaves similarly to normal distribution
Quantile regression

Two-step algorithm

Oracle rules

Simulation

References

Quantile projection setting

- \( \mathbf{B} \): cubic B-spline with \( q = \text{dim}(\mathbf{B}) \) defined on \( G = 4 + q \) equidistant knots on \([\tau_L, \tau_U]\). We require \( K > q \) so that \( \hat{\beta}(\tau) \) is computable (see (2.4))

- \( N = 2^{14} \)

- \( y_0 = Q(x_0; \tau) \) so that \( F_{Y|X}(y_0|x_0) = \tau \)

- From Theorem 3.2, the coverage probability with size \( \alpha = 0.95 \) is

\[
P\{ \tau \in \left[ \hat{F}_{Y|X}(Q(x_0; \tau)|x_0) \pm N^{-1/2} \sqrt{\tau(1 - \tau)x_0^\top \Sigma_X^{-1} x_0 \Phi^{-1}(1 - \alpha/2)} \right] \}
\]
$F_{Y \mid X}(y \mid x), \varepsilon \sim \mathcal{N}(0, 0.1^2), m = 4$ for $\hat{\beta}(\tau)$
$F_{Y|X}(y|x), \varepsilon \sim \mathcal{N}(0, 0.1^2), \ m = 32$ for $\hat{\beta}(\tau)$
$F_{Y|X}(y|x), \varepsilon \sim \text{Exp}(0.8), m = 4$ for $\hat{\beta}(\tau)$
$F_{Y|X}(y|x), \varepsilon \sim \text{Exp}(0.8), \ m = 32$ for $\hat{\beta}(\tau)$
Summary for simulation of $F_{Y|X}(y|x)$

- Increase in $m$ lowers both $S^*$ and $q^*$ (the critical point in $q$ for the oracle rule).
- Either $S > S^*$ or $q < q^*$ ($q = G - 4$) leads to the collapse of the oracle rule.
- Increase in $q$ and $K$ improves the coverage probability.
- For $\varepsilon \sim \mathcal{N}$, coverage is no longer symmetric in $\tau$. 

Thank you for your attention

Assumption (A): data \((X_i, Y_i)_{i=1,...,N}\) form triangular array and are row-wise i.i.d. with

(A1) Assume that \(\|Z_i\| \leq \xi_m < \infty\), where \(\xi_m = O(N^b)\) is allowed to increase to infinity, and that

\[
1/M \leq \lambda_{\text{min}}(\mathbb{E}[ZZ^T]) \leq \lambda_{\text{max}}(\mathbb{E}[ZZ^T]) \leq M
\]

holds uniformly in \(n\) for some fixed constant \(M\).

(A2) The conditional distribution \(F_{Y|X}(y|x)\) is twice differentiable w.r.t. \(y\). Denote the corresponding derivatives by \(f_{Y|X}(y|x)\) and \(f'_{Y|X}(y|x)\). Assume that

\[
\bar{f} := \sup_{y,x} |f_{Y|X}(y|x)| < \infty, \quad \bar{f}' := \sup_{y,x} |f'_{Y|X}(y|x)| < \infty
\]

uniformly in \(n\).

(A3) Assume that

\[
0 < f_{\text{min}} \leq \inf_{\tau \in T} \inf_x f_{Y|X}(Q(x; \tau)|x)
\]

uniformly in \(n\).
\[ \begin{align*}
\mathbb{E}[G(\tau)G(\tau')] &= \mathbb{E}[Z(x_0) \top J_m(\tau)^{-1} \mathbb{E}[ZZ\top] J_m(\tau')^{-1}Z(x_0)(\tau \wedge \tau' - \tau \tau')]. \\
\mathcal{K}(\tau_1, \tau_2) &= \|Z(x_0)\|^2 Z(x_0) \top J_m^{1}(\tau_1) \mathbb{E}[ZZ\top] J_m^{1}(\tau_2)Z(x_0)(\tau_1 \wedge \tau_2 - \tau_1 \tau_2).
\end{align*} \]
\[ \beta_3 = (0.21, -0.89, 0.38)^\top; \]
\[ \beta_{15} = (\beta_3^\top, 0.63, 0.11, 1.01, -1.79, -1.39, 0.52, -1.62, 1.26, -0.72, 0.43, -0.41, -0.02)^\top; \]
\[ \beta_{31} = (\beta_{15}^\top, 0.21, \beta_{15}^\top)^\top. \]

(5.2)

\[ \beta(\tau) = (0.21 + 0.1 \times \Phi_{\sigma=0.1}^{-1}(\tau), \beta_{m-1}^\top)^\top \]

\( f_{\varepsilon,\tau} > 0 \) is the height of the density of \( \varepsilon_i \) evaluated at \( \varepsilon_i \)'s \( \tau \) quantile.